Quantitative Description of MicroRNA Target Site Occupancy in Mouse Embryonic Stem Cells and Derived Cells of Neuronal Lineage

Marek Kudla

Watson School of Biological Sciences

A thesis submitted to the faculty of the Watson School of Biological Sciences
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

July 2013

Contents

1 Abstract 8

2 Introduction 10

2.1 MicroRNAs as Crucial Elements of Ubiquitous Mechanism of Post-transcriptional Regulation of Gene Expression 10

2.2 Introduction to MicroRNA-mediated Inhibition Mechanism and Molecular Architecture of ARGONAUTE 14

2.3 Experimental Methods for Determination of MicroRNA Interaction Sites 19

2.4 Embryonic Stem Cells and Importance of MicroRNAs in the Differentiation Mechanism 23

3 Experimental Methods and Data Analysis 27

3.1 Transcriptome Library Preparation Using "Not So Random" Hexamer Priming Method 27

3.2 Cell Culture and Differentiation Protocol 28

3.3 Cloning MicroRNA Target Sites Using HITS-CLIP on ARGONAUTE2 29

3.3.1 UV cross-linking of cells 30

3.3.2 Preparing beads for IP 30

3.3.3 Lysis and DNase/RNase treatment 31

3.3.4 Immunoprecipitation, RNase digestion, Phosphatase reaction and ligation . . . 31

3.3.5 Protein electrophoresis, transfer to nitrocellulose, autoradiography 33

3.3.6 RNA extraction, 5' phosphorylation and linker ligation 33

3.3.7 Reverse Transcription, PCR, size selection and amplification 35

3.4 Processing and Mapping Sequencing Reads 36

3.5 Determining a Reference Set of Messenger RNA Fragments Strongly Bound by RISC . 38

3.6 Statistical Analysis of Count Data 38

3.7 GO Terms Analysis and Visualization of Networks 39

3.8 Brain Gene Co-Expression Network Association Analysis for Genes Containing Differentially Bound Peaks 39

4 Results 41

4.1 Establishing a Reliable HITS-CLIP Protocol 41

4.2 Progressive Data Filtering Shows Ubiquitous Binding of RISC with Preference for 3'UTRs and CDS . . . 44
4.3 MicroRNA Expression in ES Cells Is Dominated by the MiR-290-295 and MiR-302 clusters . . . 47
4.4 Excellent Correlation of the Reference Set of RISC Binding Sites Across Replicates Produces a Data Set Suitable for Differential Binding Analysis . . . 49
4.5 Reference Set of Binding Sites is Well Supported By Different Antibodies and Existing ES Cell CLIP Datasets . . . 49
4.6 MicroRNA Seed Families Reduce the Complexity of the MicroRNA Expression Profile 52
4.7 Search for Overrepresented Sequences in the Reference Set Target Site Data Identifies Motifs Corresponding to Reverse Complement of Seed Sites from Most Highly Expressed MicroRNAs . . . 56
4.8 Impact of MicroRNA Level Change on RISC Binding to Matching Target Sites . . . 58
4.9 Conservation and TargetScan microRNA Target Site Prediction Support For Reference Target Site Set . . . 61
4.10 Long Non-coding RNA are Among Targets of RISC . . . 63
4.11 Changes of Underlying Transcript Levels Are Not a Major Factor Influencing Target Site CLIP Reads . . . 65
4.12 Calling RISC Binding Sites Differing Between Conditions Using Negative Binomial Distribution Statistics and a Generalized Linear Model . . . 67
4.13 Functional Analysis of Gene Sets Containing Differentially Bound Target Sites . . . 72
4.14 Differentially Changing Target Sites Containing Mir-19 and mir-367 Complementary Sequences Are Located in Genes Related to Neuronal Differentiation . . . 81
4.15 The Insulin Receptor Signaling Pathway as an Example of a Heavily MicroRNA Regulated Protein Interaction Network . . . 84
4.16 Brain Gene Co-Expression Network Association Analysis for Genes Containing Differentially Bound Peaks . . . 88

5 Conclusions and Perspectives **93**

5.1 Adapting Methods for Quantitative Comparison of Count Data to RISC Binding Changes in Differentiation Comparisons 94

5.2 Gene Ontology Analysis of Gene Set Differentially Bound by RISC in EBRA Indicates Enrichment in Neuro-specific and Regulatory Genes 96

5.3 Gene Set Enrichment Analysis Indicates microRNA-mediated Regulation of Major Signaling Pathways of Known Neuro-specific Roles 98

5.4 External Brain Co-Expression Studies Support EBRA Condition as Being Highly Enriched in Genes Specific to Brain 100

5.5 MicroRNA Profiles and RISC Binding Data Suggest Greater Role of MicroRNA in Neuronal Lineage Specification Than Differentiation of ES Cells 101

5.6 Published Data on Mouse Embryonic Stem Cells RISC Binding Sites Are Highly Consistent with Our Reference Set of Target Sites 102

5.7 Neuronal Differentiation of Human Embryonic Stem Cells Show Extensive Parallels to Our Results 103

5.8 Future Directions 104

5.9 Closing Remarks 107

6 Appendix **108**

References **109**

List of Figures

1 Historical figure from Francois Jacob and Jacques Monod publication on regulation of gene expression 11

2 Diagram of microRNA biogenesis and mechanism of action 15

3 Human ARGONAUTE2 in complex with miR-20a provides insight into microRNA-mediated mRNA recognition 18

4 Diagram of HITS-CLIP protocol 21

5 Flowcharts showing bioinformatics processing of sequencing libraries 37

6 Details of cloning process and sequencing results 43

7 Library filtering and its impact on reads coming from different parts of transcripts . . . 45

8 Major microRNA clusters in ES cells and expression profile for microRNAs in replicates 48

9 Experimentally derived target site counts show good correlation between condition replicates 50

10 Cross-validation of the reference set of 8989 target sites with additional antibodies and published HITS-CLIP stem cell data 51

11 Composition of the microRNA seed families and their expression in relation to target sites 53

12 Unsupervised motif search in reference peak set yields sequences assignable to microRNAs 57

13 Impact of microRNA expression changes on matching target sites 59

14 Conservation of 3’UTR target sites 62

15 Number of peak sites per gene reveal lincRNAs as transcripts abundantly bound by RISC 64

16 CLIP read counts are not simple derivative of transcript levels 66

17 Effects of transcriptome adjustment with generalized linear mode on calling HITS-CLIP peaks 70

18 Differential binding plots for uncorrected CLIP peaks (according to DESeq) 71

19 Differential binding plots for uncorrected CLIP peaks (according to GLM) 73

20 Pathways and gene groups in differentiation to neuro-specific cells 75

21 Matrix of GO terms for neuro-related genes found in LIF to EBRA comparison 76

22 GO term network constructed from genes bound differentially by RISC in LIF to EBRA transition 78

23 Detailed statistics of genes repressed or de-repressed in GO term network 79

24 Gene Ontology terms significantly enriched for miR-367-controlled genes 82

25 Gene Ontology terms significantly enriched for miR-19 and corresponding genes . . . 83

26 Signaling pathways controlled by RISC binding sites and the extent of microRNA-mediated regulation of Insulin Receptor Signaling Pathway 85

27 Genes of Insulin Growth Factor Signaling Pathway with their RISC binding sites - part 1 86

28 Genes of Insulin Growth Factor Signaling Pathway with their RISC binding sites - part 2 87

29 Western blot showing PTEN, phospho-AKT and IGFBP3 protein levels 89

30 Differentially bound sets of genes progress from non-specific to being brain-specific according to external gene co-expression datasets 91

31 Genes of TGF-beta Signaling Pathway with their RISC binding sites - part 1 105

32 Genes of TGF-beta Signaling Pathway with their RISC binding sites - part 2 106

Acknowledgements

I would like to thank Gregory Hannon for being my scientific mentor and supporting me in my research on microRNA target sites. His lab is a place of active exchange of scientific ideas and working there was for me a great learning experience on many levels.

I want to thank my thesis committee for actively shaping this dissertation during the course of its evolution. I want to thank Richard McCombie, Adrian Krainer and Marja Timmermans for sitting through all of the meetings during those years. I am very also immensely grateful to Molly Hammell for joining my committee and extensive amount of help contributed in the late part of my research. I also want to thank Dave Jackson for being my academic mentor and cheering me up every now and then during our meetings. Ihor Lemishka was so kind to agree to be my external reviewer and I would like extend thanks to him for his commitment.

More than anyone I am indebted to Fedor (Ted) Karginov, who was helping me when I was beginning my research work. Most importantly, Ted was a friend I could always count on and share my thoughts about research and other life things. I would like to thank Eugene Plavskin, my fellow graduate student and friend, who has helped me immensely in editing this thesis. Without his help, his constant support and our fruitful discussions I would have not be able to finish.

I also want to thank Simon Anders for adapting his code and developing a Generalized Linear Model applicable to our research problem. Jesse Gillis helped me in applying his methodology of functional analysis to my datasets for which I am truly grateful. I would also like to thanks Aldo Mele, Sun-Wook Chi and Robert Darnell for providing a hands-on experience with HITS-CLIP protocol. Many thanks to people who have provided reagents, cell lines and important advice: Betty Chang, Youngkyu Park, Assaf Gordon, Dawid Nowak, Simon Knott, Ingrid Ibarra, Camila Dos Santos, Sakari Kauppinen, Miao He and many others.

I would also want to thank students I had occasion to supervise: Debbie Goodman, Thomas J. Dowling and Grzegorz Sienski. It is truly rewarding to see real talent in its raw form and help to refine it.

As shaping a truth-seeking mind takes many years of interactions with mentors and teachers, I would like to thank all of them for providing me the opportunity to acquire knowledge and learn discipline in reasoning. I have always tried to make the most use of what I was presented during my

years at school and at the University. I feel that this is the proper place to thank Republic of Poland for providing free higher education for its citizens, as well as Watson School of Biological Sciences and George A. and Marjorie H. Anderson Foundation for recognizing me and providing me with the opportunity for studying here.

I am infinitely indebted to my parents Stanislaw and Jadwiga Kudla, who have gently supported me in my education throughout all those years. My mother and my brother Jacek are the most important people in my life who had to endure the separation from me for all those years. Finally, I would like to thank my wife Manuela for becoming my life partner, supporting me in my work and accepting all difficulties and challenges which came along the way. Also great thanks are going to my little daughter Milena for being a great joy in my life.

1 Abstract

The problem of differentiation of pluripotential Embryonic Stem Cells into a lineage of specific committed fate has not been studied extensively from the perspective of microRNA regulation. While studies of isolated genes and microRNAs are common, systematic studies are scarce, have emerged very recently and have been restricted in their ability of comparison between the samples due to the insufficient statistical approaches.

In our study, the HITS-CLIP methodology has been applied to the problem of microRNA function in neural fate specification in a forced differentiation of mouse ES cells into Neuronal Progenitor Cells (NPC). The high quality of data generated as a result of modifications we introduced into the HITS-CLIP protocol and subsequent data analysis allowed us to perform quantitative comparisons of target site data between various differentiation conditions; this includes the normalization of RNA-induced Silencing Complex (RISC) binding sites by transcript abundance to obtain a measure of changes in intrinsic binding strength of RISC to the target site. The reference set of RISC binding sites defined in this work shows modest support by previously published data on ES cells and target site prediction software. It also confirms the long standing notion that functional RISC binding occurs primarily in 3'UTRs and CDS regions of transcripts, showing conservation of those binding sites in the former. Of additional interest is the binding of RISC to a number of lincRNAs, which show RISC binding sites along their entire length. Futhermore, our data demonstrates that neuronal fate differentiation is associated with dramatic changes in RISC binding to target sites in many genes which cannot be explained solely by variations in the level of their underlying transcripts.

We have identified known neural function related microRNAs: miR-18, miR-19 and miR-130/301 as specifically upregulated in our terminal differentiation condition of NPC. Another microRNA which contributed a biggest fraction in this condition was miR-27. Additionally miR-367, an only member of miR-302/367 cluster with sequence different than other members of this miR-290-295 cluster homolog, is transiently upregulated during the interim stage of Embryonic Body. While more microRNAs are expressed in our experimental model, those microRNA seed families contribute more than 95% of all reads showing that the stage of microRNA-mediated regulation is set with participation of just few major players.

Those microRNAs show consequential upregulation of RISC binding in their target sites, com-

pared to rest of the sites. This effect is even more pronounced in case of miR-19 and miR-367, which show site upregulation larger than any other class of target sites.

Interestingly, genes containing target sites changing in their RISC binding in the NPC condition are enriched in functionally related genes which appear to be participating in the process of differentiation into central nervous system cells. Within this set just miR-19 and miR-367 target sites alone show statistically significant enrichment for neuro-related terms implying their specialization in this cellular context. The set of genes containing differentially bound targets is also enriched in regulation networks, such as alternative splicing, histone methylation, protein ubiquitination and microRNA-related regulation, implying that in the differentiation process, the cell has to modulate its own control systems.

Additionally, we have identified multiple signalling pathways targeted by the RISC which have their functions implicated in neuronal differentiation or maintenance of pluripotency. Among them, the Insulin Receptor Signaling Pathway seems to be heavily targeted by RISC, showing differential regulation.

Having identified a set of genes differentially bound by RISC in our NPC condition we were interested in how indicative they are of differentiation fate specific to the central nervous system. We have performed cross-validation assessment of neighborhood association derived from co-expression networks to predict membership of our genes within sets of brain and non-brain genes. In our data we can observe a transition from RISC binding to transcripts that are not specific to brain in two of our comparisons to being strongly brain-specific in NPC condition. Furthermore, the enrichment of this NPC-specific gene set in neuronal lineage specific members is so high, that it appears as an outlier in our comparison compared to other gene sets, including those associated with Gene Ontology neuronal processes. This is indicative of network-like collaboration of regulated genes captured using our methods in specifying neuronal fate.

We have successfully derived experimentally supported set of RISC binding sites and have applied statistical approaches determining sites changing specifically in differentiated NPCs. We also demonstrate that this gene set is rich in neuronal fate specification genes and that NPC up-regulated microRNAs are consequently increasing RISC binding to their target sites.

2 Introduction

2.1 MicroRNAs as Crucial Elements of Ubiquitous Mechanism of Post-transcriptional Regulation of Gene Expression

The exact nature of the regulation of gene expression has been a long-standing question. It has been known since the end of the nineteenth century that bacteria synthesize specific enzymes only if the substrate for a given enzyme is present. This phenomenon, known as "enzymatic adaptation", had waited for detailed analysis for over sixty years since its discovery. In 1961 a seminal article by Francois Jacob and Jacques Monod described in detail the regulation of genes involved in lactose metabolism. [32] Bacteria facing short supply of their preferred energy source - glucose, need to switch on specific proteins required for the transport, transformation and use of complex sugars. If lactose is present, and glucose absent in the environment, bacteria turn on several genes encoding proteins responsible for facilitating uptake and breakdown of this sugar. Those genes are located in sequence on the bacterial chromosome, are transcribed into a single messenger RNA, and are subsequently translated into multiple proteins that function together (see Figure 1).

Jacob and Monod proposed two models describing the regulation of lactose metabolism genes. In the widely known first model, also known as the "genetic operator model", they proposed that a separate regulator gene encodes a protein, which can bind and be inhibited by a small molecule - either lactose itself or its metabolite. In its inhibited state the regulator protein does not bind to DNA, while without the inhibitor molecule, it binds to the operator gene located next to the structural genes and stops production of the messenger RNA. The second model, otherwise known as the "cytoplasmic operator model", assumes that the protein encoded by the regulator gene binds an operator sequence located on the messenger RNA molecule. While originally Jacob and Monod could not exclude applicability of the cytoplasmic operator model to lac operon regulation, later studies have proven it to be invalid. In fact, the majority of gene regulation occurring in bacteria seems to fit the genetic operator model. Since then, this finding has been extended to eukaryotic organisms. For more than thirty years, the cytoplasmic operator model has been believed to be obsolete, with just a few examples such as Trp attenuation mechanism, riboswitches and RNA binding proteins which are restricted in scope to relatively small set of genes.

Only in 1993, Rosalind Lee, Rhonda Feinbaum and Victor Ambros discovered that the product

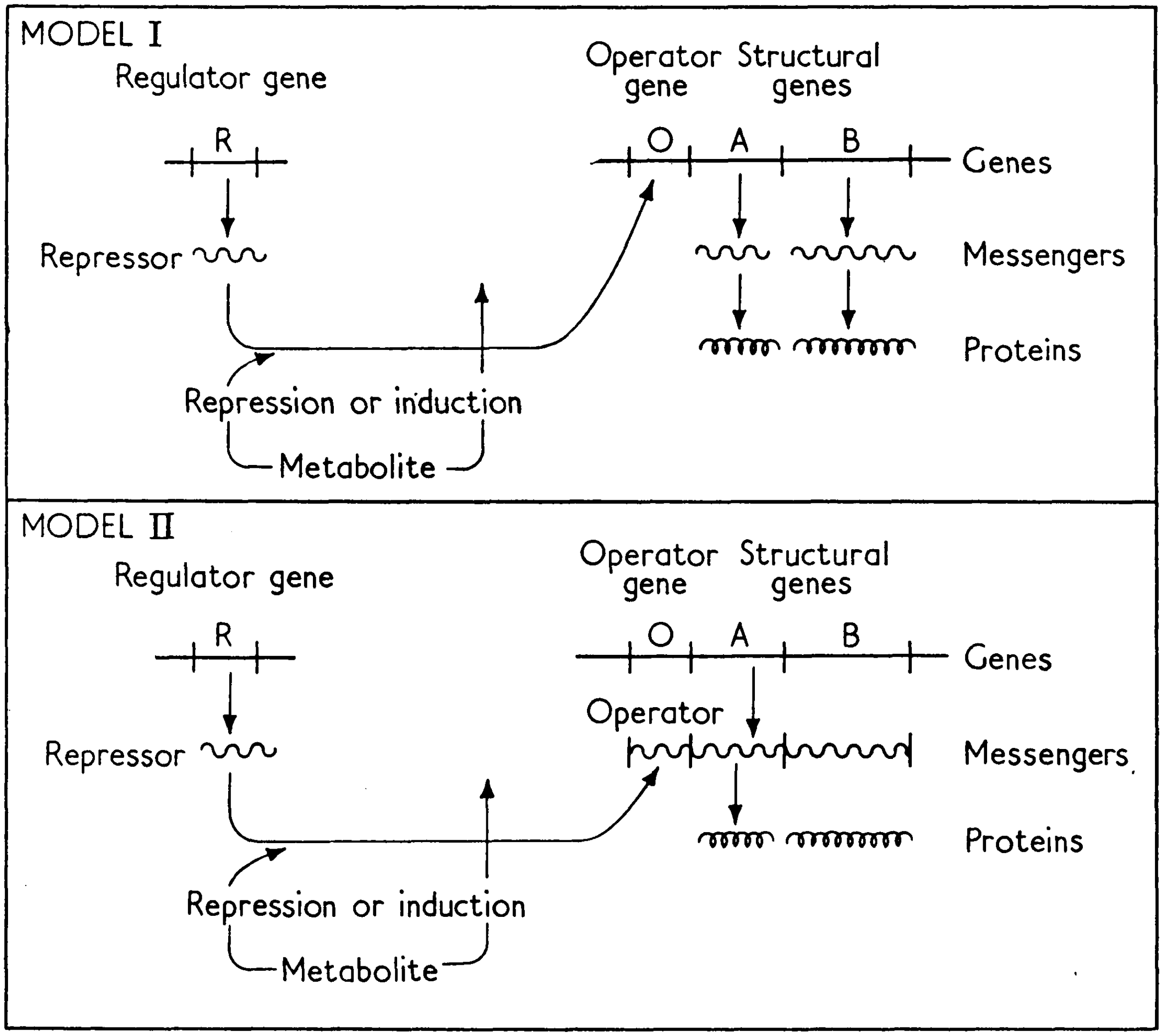


FIG. 6. Models of the regulation of protein synthesis.

Figure 1. Historical figure from Francois Jacob and Jacques Monod publication on regulation of gene expression
In their 1961 article Jacob and Monod were considering two alternative models adequately describing their results. In the genomic operator model (Model I) the expression of genes is regulated by a DNA-binding metabolite-dependent repressor protein. In second model, also known as cytoplasmic operator model, the expression of the genes is regulated by a messenger RNA binding protein in cytoplasm. While Model I is favored in Prokaryota, Eukaryota show examples of gene expression control relevant to both models.

of the lin-4, a heterochronic[1] gene from Caenorhabditis elegans was active as an RNA molecule that down-regulates levels of the LIN-14 protein. [39, 40] Lin-4 is a small transcript 61 nucleotides long. Due to its shortness and lack of start and stop codons, it does not encode any protein product, a fact that was confirmed with additional experiments. Furthermore, it can be divided into two regions of equal length with a high level of complementarity to each other, which makes lin-4 RNA fold into a hairpin structure. Interestingly, it appeared that this hairpin is further processed to a single stranded fragment, 22 nucleotides of length. Based on the fact that lin-4 is complementary to the sequence regions in the 3' untranslated region of lin-14 mRNA, researchers reasoned that the negative regulation of LIN-14 levels comes from binding of lin-4 to lin-14.

In this example, the cytoplasmic operator model has been re-established as a relevant hypothesis and, as it later turned out, applicable to a very general and ubiquitous mechanism of gene expression regulation. Lin-4 was the first discovered microRNA - an RNA molecule characterized by its small (18-24 nt) size and complementarity to targets, conventional mRNA molecules. MicroRNA action, as shown by Lee and colleagues, inhibits gene expression at the post-transcriptional level, in contrast to the genetic operator model of Jacob and Monod which was believed to be dominant, if not exclusive.

Further insight came when Craig Mello, Andrew Fire, and colleagues established that injections of double stranded RNA molecules into Caenorhabditis elegans led to down-regulation of complementary messenger RNA molecules. [16] By selecting for inhibition genes which have clear, observable phenotypes in C. elegans, Fire and Mello established a powerful experimental model for investigation of dsRNA-mediated inhibition. The effect observed by researchers has been proven to be very specific and dependent on the sequence of injected dsRNA. The injection induces a phenotype similar to deletion of the targeted gene. Interestingly, neither the sense or antisense strand injected on its own caused any effect, a stunning example of a true negative control. Only if sense and antisense strands were injected shortly in sequence after each other, they evoked an observable effect. If the separation time was longer than one hour the effect disappeared, most likely due to the fast turnover of those single stranded RNA molecules in the cell. Generality of this RNA interference effect has been demonstrated by the use of dsRNA molecule complementary to previously transfected GFP gene, a foreign DNA encoding a fluorescent protein. In this case as well, the fluorescence was extinguished upon dsRNA injection, arguing

[1] heterochronic - influencing timing of developmental changes, heterochronic gene controls timing of developmental events. Its mutations may cause abberations in the developmental process resulting in changes of size and shape of an adult organism or its juvenile developmental stages.

in favor of the hypothesis that the observed phenomenon is exclusively sequence-mediated. In the same publication, the authors also showed that the RNA interference acts in a highly non-stoichiometric way, where injection of an adult with dsRNA lead to phenotypic manifestation in all of its numerous progeny. At this stage, the initial injection would have been already diluted to just a few molecules per cell and would have to survive through a period of rapid degradation of cellular transcripts in the early stages of the development. It is unlikely that this could be robustly achieved without a signal amplification step.

In the follow-up article Mello, Fire and colleagues have identified protein-encoding genes involved in the RNA interference process. [58] One of the genes, rde-1 has been shown to encode a protein belonging to the evolutionary conserved ARGONAUTE protein family. Subsequently, Hammond, Hannon and colleagues have reported purification of the ARGONAUTE protein from RISC complexes in Drosophila cells establishing it as a catalytic engine of the RNA interference process. [25, 24] But what is the link between long dsRNAs used in RNA interference and relatively short microRNAs? The answer came with the discovery of the enzyme responsible for shortening long double stranded structures. The protein, DICER, is a ribonuclease containing two RNase III catalytic domains forming an intramolecular dimer which cleaves both 5' and 3' strands of the dsRNA. [7] The protein also contains an RNA binding domain for capturing double stranded RNA molecules and the entire enzyme acts as a ruler to measure approximately 20-nucleotide long sections of nucleic acid. DICER is a node linking different RNA interference pathways. It processes long dsRNAs, just like those produced during replication of certain double stranded viruses, transposons and other endogenous double-stranded substrates. It is also capable of processing hairpin structures characteristic of microRNAs, regulatory molecules of the transcriptome.

Interestingly, it took another seven years to discover a second example of functional microRNA: let-7 from C. elegans. [51] This microRNA is another example of RNA molecule influencing heterochronic genes controlling timing of developmental events. Since all cell divisions in C. elegans follow a deterministic schedule, the worm is an ideal model to capture such changes. While the loss of let-7 lead to the appearance of cells characteristic of larvae during the adult stage of worm's life, the overexpression caused adult cells to appear earlier in the life cycle. This pronounced phenotype has been explained by complementarity of let-7 to the 3' untranslated terminal regions of known heterochronic genes, thus establishing it as the master regulator of their activity. Another example of the short RNA molecule of the same class as lin-4 helped to realize that there may be many more which we were missing. Also important for development, potent effect of this microRNA, as well as capability of regulating several genes, primed researchers to think that there may be a whole layer of complexity in gene expression regulation to be uncovered. Soon enough, articles describing systematic search for the new class of molecules appeared, turning examples of similar molecules counted first in tens and later in hundreds. [42] Those results were soon extended to other organisms, including mammals. [1]

2.2 Introduction to MicroRNA-mediated Inhibition Mechanism and Molecular Architecture of ARGONAUTE

It is probably best to introduce at this stage a short overview of the microRNA biogenesis and mechanism of their action (see Figure 2, details in text). In the most canonical case, microRNA genes are transcribed by RNA polymerase II, which results in transcripts having a modified nucleotide on their 5' end (5' cap) and a long stretch of adenine nucleotides at the end, just like regular mRNAs. Those transcripts form a hairpin or a more complex structure, where a hairpin contains the future microRNA molecule. More than one mature microRNA molecule can be contained within such structure, however they are separated in the next processing step. In this form, the microRNA precursor is called pri-miRNA and is further processed by a DROSHA/DGCR8 protein complex to form a shorter hairpin (with 20-26 nucleotide paired region) by removal of both capped, as well as polyadenylated ends. In this form, also known as pre-miRNA, the microRNA precursor is exported to the cytoplasm in a GTP dependent process involving EXPORTIN5. It should also be noted that a novel class of microRNA (mirtrons) has been characterized, which are processed from the introns of host genes, bypassing DROSHA. While in the cytoplasm, the pri-miRNA is recognized by a DICER protein and shortened to a double stranded RNA molecule of 20-24 nucleotide length, with two nucleotide overhangs at each end. Usually, only one of the strands is loaded into an ARGONAUTE protein which led to naming it the guide strand, while the discarded strand has become a passenger or star strand. However, it turned out that the mechanism of strand choice is not as decisive as it had been thought, and there are multiple examples of microRNAs where the ratio of passenger to guide strands loaded into the Argonaute protein is close to 1:1. This led to the abandonment of previous passenger/guide terminology in databases in favor of -5p/-3p suffixes, which do not carry information about loading. [17] For example, miR-19-5p denotes a single stranded RNA coming from the 5' part of the pri-miRNA, while miR-19-3p comes from 3'

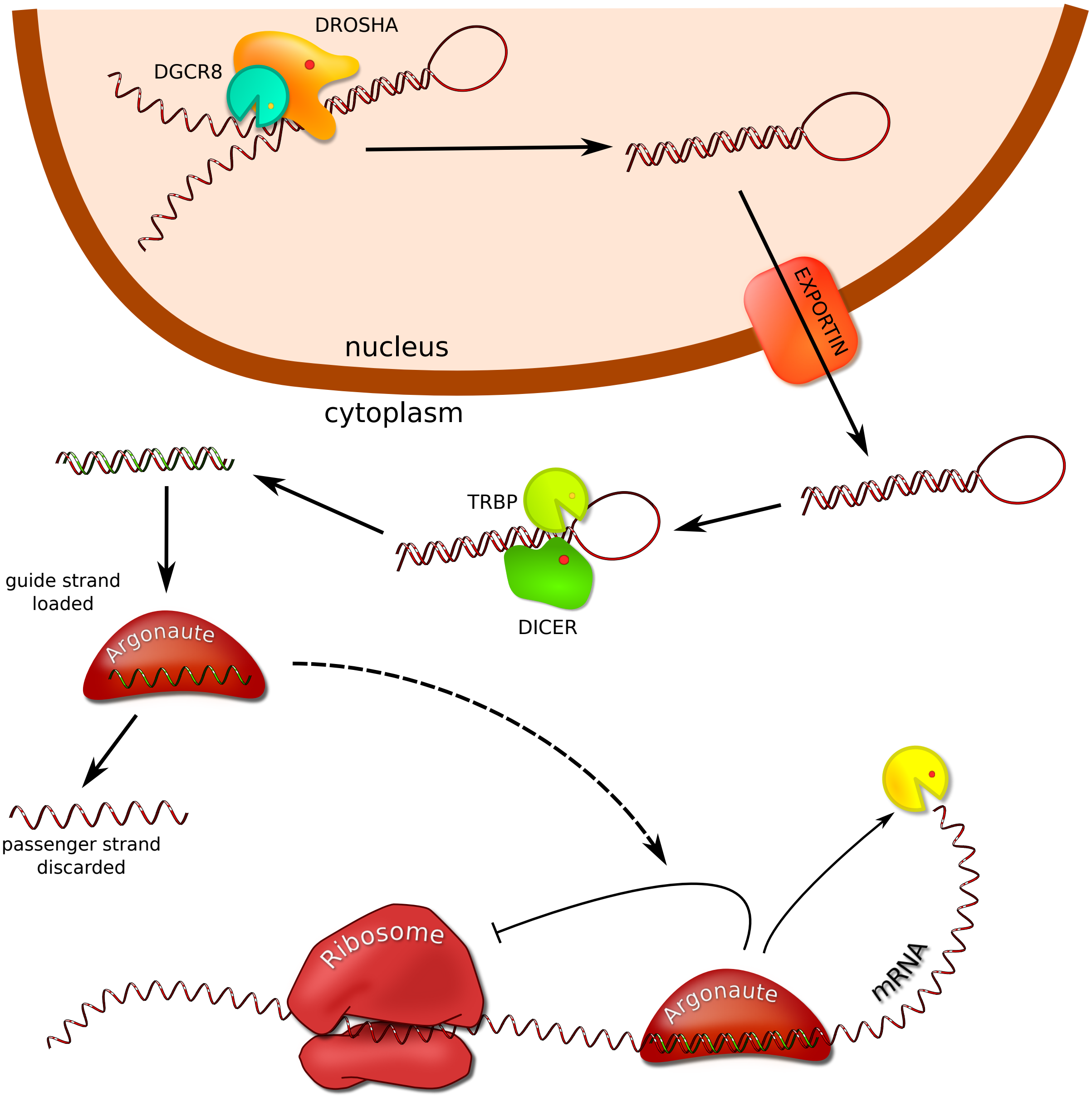


Figure 2. Diagram of microRNA biogenesis and mechanism of action

MicroRNA is transcribed as a precursor pri-miRNA by RNA polymerase III and is further processed by a DROSHA/ DGCR8 protein complex to a short hairpin: pre-microRNA. This molecule is exported through nuclear pore with participation of exportin 5 carrier protein. In cytoplasm of the cell terminal loop of the hairpin is processed by a DICER-containing complex into the guide/passenger strand duplex. Only one of the strands, known as the guide strand is loaded into one of the ARGONAUTE proteins. MicroRNA-containing ARGONAUTE protein finds its complementary target site on a messenger RNA molecule and induces its silencing through inhibition of translation and/or degradation by RNases.

part. A mature microRNA strand loaded into an ARGONAUTE protein is a specificity factor for this ribonucleoprotein complex. Part of the microRNA, exposed from its protein surroundings, pairs with other RNA molecules, thus making ARGONAUTE a programmable RNA binding protein. Interestingly, the ARGONAUTE protein has two distinct modes of action. One out of four ARGONAUTE proteins found in mouse or human, ARGONAUTE2, has an endoribonuclease activity and will cut the RNA with perfect complementarity to the loaded guide strand. In case of imperfect complementarity the mRNA remains associated with the complex which results in inhibition of translation and/or degradation of bound mRNA.

Structural characterization of ARGONAUTE (AGO) proteins succeeded initially in Archaea, with multiple structures of the protein alone and in complex with microRNA and target analogs available for study. [44, 56, 57] Those early studies established Ago as an multi-domain protein, composed of N-terminal, PAZ, Middle (Mid) and PIWI domains. Structures containing guide strands and RNA:DNA duplexes reveal that a cleft between the PAZ domain and remaining domains is the region accommodating nucleic acids. [66] The first nucleotide of the 5' end of the microRNA is held in a special pocket of the Mid domain and, given its extensive interactions with amino acids and occlusion of its base edge by the domain, it is unlikely for it to participate in Watson-Crick base pair formation and therefore in the recognition of target. This and additional studies show that U base is preferred in this position. Mid and PIWI domains form extensive interactions with the phosphosugar backbone, while almost completely avoiding hydrogen bonds with RNA bases. This is in agreement with the expectation that a heterogenous population of RNA molecules must fit ARGONAUTE, and although it is not known if there exist sequences which would be disfavored, it is safe to state that ARGONAUTE is relatively promiscuous towards diverse set of RNA sequences. The endonuclease activity characteristic of ARGONAUTE was discovered to be a property of a PIWI domain, which structurally resembles RNase H, while being dissimilar in its amino acid sequence. [57] The risk associated with inferring features of the mechanism of ARGONAUTE activity based on AGOs found in the Archaea is that those organisms do not have a known RNA interference mechanism. Additionally, investigated structures have higher affinity towards DNA targets than RNA. The function of those proteins in Bacteria and Archaea is believed to be associated with genomic defense towards rogue DNA elements, like phages. [36, 49, 50] Eukaryotic ARGONAUTES have remained elusive until very recently, when the full human structure bound to microRNA was published by two research groups. Surprisingly, archaeal and eukaryotic structures remain remarkably similar, despite more than 3 billion years of evolution separating those domains. [57, 52, 14] One notable insight coming from eukaryotic structures are the details of the microRNA seed site exposed by the Argonaute protein (see Figure 3). Nucleotides 2 to 6 of the microRNA are arranged in a form analogous to one of the strands from an A-DNA structure (favorable for 2'-hydroxyl-containing RNA) and their bases are stacked against each other. This arrangement favors the formation of an RNA-RNA duplex and is likely a key factor contributing towards enhanced affinity of the microRNA to its matching target. In fact, studies have shown that while bound to the ARGONAUTE protein, the RNA seed sequence is more than three hundred fold more likely to bind its matching sequence compared to the same seed sequence not bound to the Argonaute scaffold. It is noteworthy that for the human ARGONAUTE protein, only five bases form the nucleation seed for hybridization. The 7th nucleotide of a microRNA is separated from the seed sequence by an intercalating isoleucine from the PAZ domain. [52, 14] While inconsistent with the requirement for six matching nucleotides forming a seed sequence, an experimentally well-established fact, this finding makes sense in light of recent discoveries based on experimental target site data. While the biggest fraction of target sites can be identified as being controlled by a particular microRNA by simple complementarity to at least microRNA six-nucleotide seed site, a fraction of orphaned sites can be similarly assigned if the formation of a bulge in the targeted sequence is allowed. [11] The mRNA nucleotide forming a bulge would be falling into the kink between the nucleotide 6 and nucleotide 7 of the microRNA, while the preceding nucleotide would be pairing with nucleotide 7. This mechanism is supported by statistically significant overabundance of "bulged" motifs found in the target site sequences and the fact that those transcripts are downregulated after the transfection of a matching microRNA. While the microRNA pairing by bulge mechanism explains only a part of the orphaned sites, it is sufficiently documented to treat it as an important target site recognition mode. It is also the only one so far supported by understandable structural features of ARGONAUTE:microRNA complex, specifically the hinge between 6th and 7th nucleotide of the microRNA. There have been however many more exception rules proposed to the seed site pairing, involving bulges and G:U wobble pairs. Despite full crystallographic structure of decent quality and resolution, we only have one conformation of the complex and we cannot yet extrapolate this structure to non-canonical binding modes. [52, 14, 11]

Several observations suggest additional binding modes for microRNA target sites. "Centered sites" are one of them, being characteristic thanks to the lack of pairing within the seed as well as the 3' region of the microRNA. Instead they feature eleven to twelve consecutive paired nucleotides in the

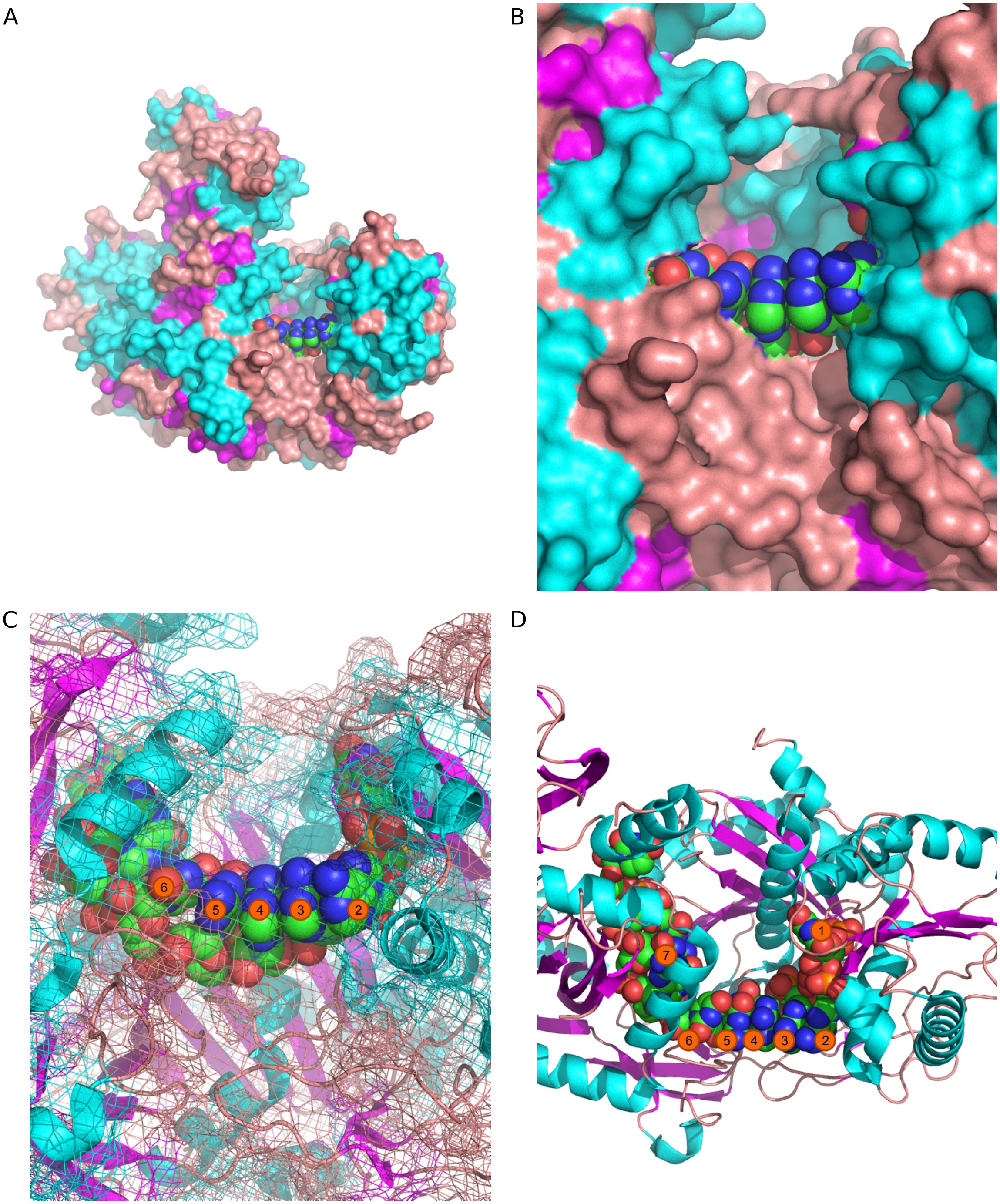


Figure 3. Human ARGONAUTE2 in complex with miR-20a provides insight about microRNA-mediated mRNA recognition.

Subfigure A shows a full model of RiSC complex with protein modelled as surface and color-coded according to secondary structure with alpha helices in cyan, beta sheets in magenta and loops in red. In the middle of the model the seed site of the microRNA is visible depicted as single atom spacefiling (blue, green and red). B shows the center of the model and C shows secondary structure as cartoon and surface of the protein as a mesh to show stacked bases of the seed. Subfigure D shows region from C from different angle. It is clearly visible that the bases of the RNA stick out towards cytoplasm (they can be recognized by their blue nitrogen atoms), while phospho-sugar backbone interacts with the protein. The bases 2 to 6 are not obstructed by protein, while base 1 goes into a protein pocket, which is a conserved feature of the complex. Interestingly, the base number 7 is separated from the rest of the seed bases by a kink in the RNA structure induced by a stacking interaction with one of the isoleucins.

center of the microRNA:mRNA duplex. Interestingly, despite perfect base pairing encompassing nucleotides 10 and 11 of microRNA, those sites control transcript levels bypassing the Argonaute-induced cleavage characteristic of a perfect match. [54] There are also reasonable indications that even in the case of no seed matching and imperfect, yet extensive level of complementarity between a microRNA and target site, the inhibition still occurs. [38, 37]

Except for the endonuclease mechanism of some ARGONAUTE proteins, the effective part of RISC action is less understood. While mass-spectrometry studies have revealed multiple proteins associating with the Argonaute protein, the exact model of inhibition and the role of participating proteins are still elusive. It is widely accepted that the GW182 protein is the primary factor, likely contributing to the translational repression mechanism through its association with the 5' cap of mRNA. Interestingly, it was also demonstrated that this protein is associated with RNA degradation proteins DCP1 and LSM1 strenghtening hypothesis about the role of GW182 in RNA degradation.

While the majority of AGO-mediated regulation is believed to be repressive, there have been a limited number of reports about translational activation induced by microRNA. [64] Interestingly, this effect has been described as being cell cycle phase dependent, with activation being characteristic to the cell cycle arrest phase. While demonstrated on two human microRNAs and a synthetic one, there are no genome-wide studies of this effect available and there is very sparse additional support from different research groups. Therefore, for the purpose of this dissertation, we assume that microRNAs act exclusively as repressors.

2.3 Experimental Methods for Determination of MicroRNA Interaction Sites

While evidence for genome-wide impact of microRNA-mediated inhibition has accumulated, researchers set out to establish a method capable of producing a map of interaction of RISC with the transcriptome. In plants, the level of complementarity of microRNAs to transcripts is sufficiently high to provide a foundation for specific bioinformatics predictions algorithms. Those methods transplanted to animals perform less than adequately, due to relaxed requirements for complementarity which were evident starting from the earliest experiments. One of the benchmark methods for investigating a specific microRNA:transcript interaction is based on a marker gene fused to the native 3'UTR of the gene under investigation. While simple and popular, this method cannot be used for sites located in the coding sequence, as well as being hard to extend into a truly global approach. Some of the early experiments involved transfection of a specific microRNA and post-manipulation measurement of transcript levels by microarrays or transcriptome sequencing. [35, 60] While measuring protein levels is the most direct way of establishing an impact of microRNA regulation on cell functions, it is limited by current mass-spectrometry technology. Most importantly, it is at this moment impossible to capture the quantities of all proteins present in the cell: only a certain fraction of the proteome is accessible to the method at any single time. Even those limited first experiments have shown that perturbation of any single microRNA under investigation impacted the levels of hundreds of transcripts and proteins. While being genome-wide, the resolution of those methods is limited to the entirety of the transcribed unit, still far from pinpointing the exact sites of RISC complex binding. The next breakthrough came from applying HITS-CLIP, a method developed in Prof. Darnell's laboratory for studying RNA binding proteins, to RISC complex. [13, 61, 62, 12] This technique is analogous to the DNA-specific CHIP method and is based on immunoprecipitation with the use of either an ARGONAUTE2-specific antibody or a tagged version of this protein in tandem with tag-specific antibody. Crosslinking of the RNA to the ARGONAUTE protein allows for its footprints on bound transcripts to be sequenced. In our extensive experience with HITS-CLIP, while it is usually challenging to apply this technique to various RNA binding proteins, the RISC complex seems to be an especially demanding case.

The basic diagram of the HITS-CLIP protocol has been depicted in Figure 4. Cells are subjected to a measured dose of 265 nm UV radiation, which crosslinks the AGO protein to the microRNA or mRNA. Subsequently, the ribonucleoprotein complex is immunoprecipitated and treated with RNase A to create a mRNA footprint, and a phosphatase reaction is used to remove phosphates from RNA fragments. This step prevents formation of undesired ligation products. Radioactive RNA linkers are subsequently ligated to the 3' ends of the RNA and the entire ribonucleoprotein complex is size-selected using protein gel electrophoresis - a specificity ensuring step. The protein is then digested, RNA extracted and subjected to 5' phosphorylation and 5' linker ligation. After the reverse transcriptase reaction and PCR amplification, the DNA product is size selected and cloned fragments are sequenced using high throughput methods.

PAR-CLIP and iCLIP are relatively minor modifications of the protocol, aiming for increased efficiency of cloning. [23, 37] PAR-CLIP relies on a modified nucleotide base, a 4-thiouridine, supplied in the cell media to be incorporated into the RNA. It is specifically cross-linked to protein using 365 nm UV radiation. While increasing efficiency of the cross-linking is the main goal of the modification,

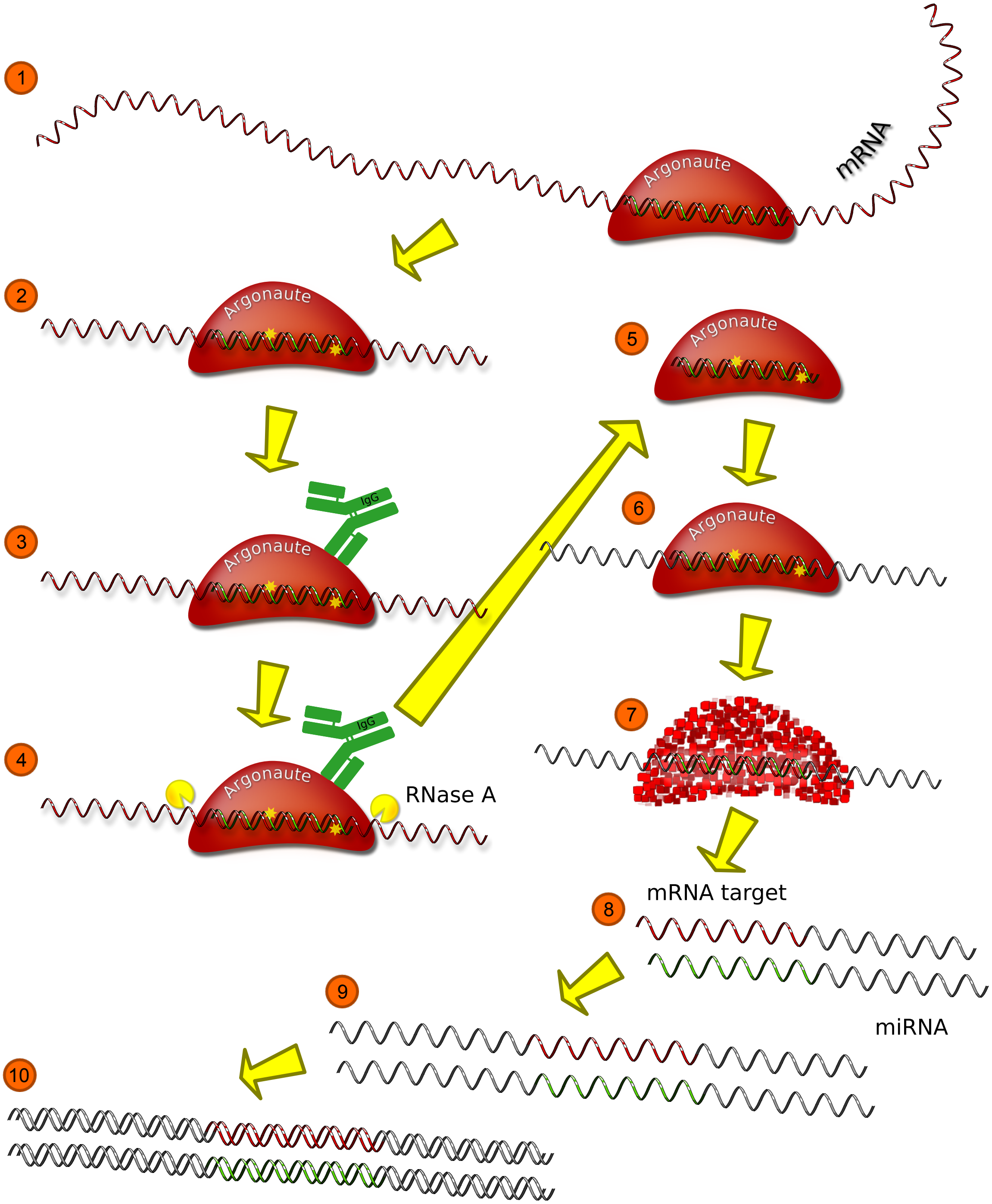


Figure 4. Simplified diagram of HITS-CLIP protocol

Step 1 shows RISC complex bound to mRNA, which is subsequently cross-linked to it using UV radiation (step 2). Immunoprecipitation with specific antibody is shown in the step 3 and the complex is subjected to RNase A treatment (step 4) to create a footprint of Argonaute protein on transcript (step 5). Step 6 involves phosphatase reaction followed by ligation of radioactive 3' RNA linker. It is notable that majority of Argonaute population is crosslinked to either miRNA or mRNA target, while double cross-linking like the one depicted on diagram is rare. Step 7 involves size selection on a protein gel followed by protein digestion which releases cross-linked fragments. The purified fragments visible in step 8 are phosphorylated and subjected to 5' end ligation. Since 3' RNA linker has a dideoxy-base on its 3' end, and the 5' RNA linker has 5' hydroxy group the formation of the product visible in step 9 is greatly favored. The reverse transcriptase reaction, followed by PCR amplification and size selection on gel, results in purified dsDNA product which is subjected to high throughput sequencing.

an additional consequence is a U to C transition of the incorporated base upon its cross-linking to the protein. This feature can be used for confirming the specificity of cloned fragments. The disadvantages of this method are the toxicity of the compound for the animals and additional variation due to incorporation randomness added on top of the cross-linking bias.

iCLIP uses only a single RNA linker and the DNA is circularized after the reverse transcriptase reaction, ligating the 3' end of the RT primer to the 5' end of the cloned DNA. This modification reduces the number of steps and also eliminates possible bias introduced by reverse transcriptase pausing on RNA base adducts coming from cross-linking. The potential disadvantage of the method is the difficulty of setting up the circularization step. While the set of possible modifications of the technique is even larger, streamlining the protocol to the point of effortless implementation does not yet seem possible. In practice all of the mentioned protocols require extensive optimizations and troubleshooting before the main experiment runs.

While successful in establishing exact sites of microRNA targeting, CLIP and related techniques suffer from lack of experimental data on microRNA:target site pairs. The latest, barely a month old, development comes from the work of Helwak, Kudla, Dudnakova and Tollervey. In their CLIP-like "CLASH" protocol involving an additional RNA-RNA ligation step between molecules cross-linked to the same ARGONAUTE protein, they are able to clone chimeric reads composed of the microRNA and its corresponding target site fragment. [27, 20] Of further interest is the fact that the target sites identified by the method fall into annotation classes with roughly similar abundance to our results and that ncRNAs are present among validated reads. While resolving pairing problem, the microRNA:target site hybrid reads captured by the method constitute only small fraction of total reads. This is most likely due to the dependence of the method on presence of both RNA molecules in the RISC complex during the ligation reaction. Simultaneous cross-linking of microRNA and mRNA to the Ago protein promotes formation of hybrids, however such events are believed to be rare and therefore the method is likely to be unsuitable for providing high-quality quantitative data on relative strength of interaction with a particular target site. However, once the microRNA:target site pairing map is produced for a specific organism, it can be used in the context of CLIP data which has potential of providing such quantitative information.

2.4 Embryonic Stem Cells and Importance of MicroRNAs in the Differentiation Mechanism

Embryonic Stem Cells are characterized by their ability to transform themselves into any specialized cell of the adult animal. This process, also known as differentiation, is the characteristic landmark of any multicellular organism (even such relatively simple organisms as Sponges, Volvox or Myxomycota show specialized groups of cells at some stages of their life). Specialization within the group of cells enhances their abilities to perform a biologically relevant function. This specialization can be visible by naked eye (nerve cells or muscle cells), but is most significantly pronounced at the molecular level. Moreover, such macro adaptations are dependent on the expression of specific proteins, thus making the control of gene expression a primary focus in the investigation of differentiation. Many animals have been evolutionarily successful, thriving to this day, with just a limited number of relatively similar cells which can revert to non-differentiated state or change identity (Platyhelminthes). However all vertebrate animals, and especially mammals, have a very high number of cell types performing very specialized functions, while their differentiation state is relatively irreversible. During the life of an animal those cells are being constantly replenished from populations of tissue-specific adult stem cells. This setup has likely evolved as a result of the necessity of tight control of cell differentiation in the context of highly specialized tissue. Activity of such a highly organised system would have been severely disrupted by random cell proliferation or type switching. Such undesired behavior is in fact the underlying mechanism of cancer pathogenicity and illustrates adequately evolutionary pressure for maintaining cell identity in highly organised animals. This tendency may be further enhanced in case of animals with long lifespan / few progeny (K-selection strategy), like Homo species.

While cellular differentiation is clearly a derivative of the cell's expression profile, it is necessary to establish the importance of microRNA-mediated inhibition of gene expression to differentiation and stem cell identity maintenance. Knockout studies in mouse have shown that disruption of microRNA biogenesis in stem cells blocks their ability to undergo differentiation. Dicer-null stem cells have been shown not to contribute to formation of chimeric mice and did not form teratomas when injected subcutaneously into nude mice. [34] They retained their ES-like morphology in culture and even in the case of forced differentiation into embryonic bodies remained aggregates of non-differentiated cells. In this type of assay, wild-type ES cells differentiate into cells from all three embryonic layers and the presence of differentiation markers for mesenchyme, endoderm or ectoderm can be easily detected. This is not the case for Dicer-null cells. Additional effects of DICER's absence included the presence of RNA transcripts from centromeric repeats, and aberrations in histone modifications and DNA methylation. [34] Some of those effects are likely to indicate the involvement of DICER in additional mechanisms in addition to microRNA-based inhibition, most probably associated with transposon or viral defense.

The Dgcr8-null mutant which disables microRNA production further upstream the microRNA biogenesis pathway than Dicer-null mutant, shows a strikingly similar phenotype, with cells maintaining their ES morphology while cultured and being unable to progress into the differentiated cells in EB formation assay or form teratomas in nude mice. [65] In Retinoic Acid induced differentiation, Dgcr8-null cells, unlike Dicer-null cells, show expression of some differentiation markers. Despite this fact, the expression of pluripotency markers remains high. Those two independent investigations show the role of microRNAs as silencers of pluripotency in ES cells. Slight differences in described phenotypes are also notable, suggesting participation of DICER and DGCR8 proteins in additional separable functions. Interestingly, a recently published article describing CLIP performed on DGCR8 shows several classes of RNA bound to the protein, including mRNAs, small nucleolar RNAs and long non-coding RNAs, like Malat1. [45] Binding of DGCR8 to mRNA may be involved in controlling of the alternative splicing mechanism. While microRNAs seem to be the most consequential subject of influence for the proteins involved in their biogenesis, they are unlikely to be exclusive. Therefore RNA products coming from additional mechanisms are likely to be detected in CLIP experiments involving miRNA-processing proteins.

ES cells have their microRNA landscape dominated by products coming from just two microRNA clusters. [30] In mouse it is a miR-290-295 cluster composed of seven microRNAs and a miR-302/367, which has five. Clusters are transcribed as a single unit and microRNAs are processed from the same pre-miRNA transcript yielding several mature strands. Interestingly, most of them share sequence similarity arguing in favor of a hypothesis of potent and robust control achieved by those microRNAs over a similar set of transcripts. One notable exception is a miR-367 with different mature strand sequence. Interestingly, in the miR-290-295 cluster the 3' part (or -3p) of the pre-miRNA is the one primarily loaded into the Argonaute and quantitatively the most abundant one. [46] However for some of the microRNAs the 5' part is present as well (291a, 291b, 292) or even dominant (290). It is unknown if this less abundant strand has any impact on the transcriptome. In the following experiments we will attempt to demonstrate that passenger strands also show a demonstrable signature in target site data.

During differentiation, additional microRNAs are transcribed and processed, impacting different sets of genes. Most notably, it is believed that the let-7 family of microRNAs is a key player in differentiation. [46] Together with its protein inhibitor LIN-28, it forms a bistable system that is believed to be capable of switching the differentiation from OFF to ON state. This role of let-7 microRNA family is well supported by its function in C. elegans as a heterochronic master control gene responsible for inducing differentiation. Additional microRNAs are implicated in reinforcing identity of any specific tissue. However, determining such tissue-specific sets of microRNAs is often problematic. While there are more than 1600 potential microRNA strands in the mouse genome, it is typical to find about 200 and more in the small RNA libraries prepared from tissues. Most of them are fairly rare, sometimes with less than a single molecule per million, a great testament to the sensitivity of our present sequencing methods. Often the levels of microRNAs between related tissues vary several-fold, while still remaining low, compared to the most abundant players. One of the recent investigations of neuro-specific microRNAs has identified 116 of them which can vary between different neuronal tissues more than five-fold. [33] One possible hypothesis is that every single mature microRNA present in the cell has its own impact on the transcriptome. The converse option is that only big players, the most abundant microRNAs, contribute to the identity of the tissue. In light that the answer to such simple question remains unclear, it is not surprising that even harder task is to judge the biological impact of microRNA expression change and and how it translates to a measurable, discrete phenotype. It is hard to imagine how those questions can at all be answered without a comprehensive and experimentaly validated set of target sites. This study represents an attempt to fill this gap and represents a first step to identifying functions for individual microRNAs.

In the following study we have decided to investigate the role of microRNA-mediated regulation in mouse embryonic stem cells as they differentiate into neuronal progenitor cells. The applicable differentiation protocol is well established and involves the use of retinoic acid which not only promotes neuronal cell fate, but also has an ability to reset some of the other states and guide them towards neuronal state. [55] In addition, there exist data about microRNA target sites obtained from mouse ES cells obtained with the use of HITS-CLIP methodology. [41] This study is interesting, since it also contrasts set of target sites from ES cells with those obtained from Dicer-null cells, however it does not address the problem of differentiation. Incidentally, a similar study performed using human ES cells, as well as derived neuronal cells, has recently been published and warrants potentially interesting comparisons between datasets with the hope of generating organism-agnostic conclusions comparing and comparing differences in microRNA-mediated regulation during human and mouse embryonic development. [43]

3 Experimental Methods and Data Analysis

3.1 Transcriptome Library Preparation Using "Not So Random" Hexamer Priming Method

This is an adaptation of the method originally published by Chris Armour and colleagues. [4] The advantage of the method is rapid amplification of transcriptome fragments suitable for direct sequencing which are depleted of ribosomal RNA. Samples were treated with Trizol reagent (Invitrogen) and RNA was separated using standard protocol. Purified RNA is subjected to RQ1 DNase (Promega) treatment for 1 hour at 37ºC, and subsequently extracted with acid phenol (300 microliters of water and 400 microliters of phenol is added) and precipitated with ethanol. In the PCR tube, 3 micrograms of isolated RNA is mixed with 2 microliters of 100 micromolar NSR forward primer pool and adjusted with water to 10 microliters. It is advisable to set up a Negative Control Sample (without RNA) and a Minus RT Sample (just like a regular sample, but reverse transcriptase is not added). The samples are heated to 65ºC (thermal cycler) for 5 minutes and then put on ice. Next, the Reverse Transcriptase Mix is added to samples followed by incubation in thermal cycler (45ºC for 30 minutes followed by 75ºC for 15 minutes and a cool down step to 4ºC), after which we add 1 microliter of RNase H (NEB) enzyme into the reaction and incubate further (37ºC for 20 minutes followed by 75ºC for 15 minutes and a cool down step to 4ºC). Subsequently QIAquick PCR purification kit (QIAGEN) is used according to standard procedure, the sample is eluted from the column with 35 microliters of Elution Buffer. 25 microliters of purified DNA is subsequently mixed with 75 microliters of Klenow Reaction Mixture and incubated for 30 minutes at 37ºC. The Qiaquick PCR purification step is repeated on the sample and 5 microliters of purified DNA is used for final amplification step. At this step different samples can be amplified with reverse primers carrying different barcodes to facilitate multiplexing samples for sequencing stage of the procedure. The PCR product is subsequently run on 1.5% agarose gel and the DNA smear between 175 and 225 bp is cut out and isolated using Qiaex II resin (QIAGEN) according to the standard protocol.

RQ1 DNase Reaction Mixture	
10 microliters	RQ1 buffer
15 microliters	RQ1 DNase
75 microliters	water

Ethanol Precipitation (for 0.4 ml sample)	
40 microliters	3M sodium acetate pH 5.5
1 milliliter	100% ethanol
Centrifuge 30 min. at 13400 r.c.f., -20ºC	

Reverse Transcriptase Mixture

4 microliters	10 mM dNTP
4 microliters	5x First Strand buffer
1 microliter	100 mM DTT
1 microliter	SuperScript III enzyme

Klenow Reaction Mixture

46 microliters	water
10 microliters	10x NEB buffer 2
5 microliters	10 mM dNTP
4 microliters	exo(-) Klenow (NEB)
10 microliters	NSR reverse primer pool

PCR Amplification Mixture

28 microliters	Accuprime Pfx SuperMix
0.5 microliter	P5-SBS3T-BC-NSR 100 micromolar
0.5 microliter	Mplex-rev-NSR-PCR, index X 100 micromolar
5 microliters	purified DNA template

PCR Program: Incubation 2 min. 95ºC, followed by three cycles of 10 sec. 95ºC, 2 min. 40ºC, 1 min. 68ºC, subsequently 26 cycles of 10 sec. 95ºC, 30 sec. 60ºC, 1 min. 68ºC, additional 5 min. of 68ºC incubation and cool down to 4ºC.

3.2 Cell Culture and Differentiation Protocol

The stem cell maintenance and differentiation protocol has been adapted from Wichterle and colleagues. [67] Stem cells were cultivated in mESC maintenance medium on gelatinized 15 cm plates. Medium was changed daily and cells were split 1:3 to 1:5 every second day. For differentiation to LIF(-) condition medium was substituted to one without LIF and cells were allowed to grow for 48 hours. For EB differentiation 2-2.5 million of LIF(-) cells were transferred to 10 cm non-adherent plate and cultivated in EB growth medium. After 2 days moderately sized embryonic bodies formed and cells were harvested or induced into neuronal cells by substituting medium with 2 micromolar Retinoic Acid and kept for additional 48 hours. If cells are maintained in the EB growth medium for 96 hours and more, they differentiate spontaneously into variety of cell types, including contracting muscle cells. This has not been observed if Retinoic Acid has been added to the medium.

mESC maintenance medium with LIF

500 ml	DMEM Knockout medium
90 ml	ES screened Fetal Bovine Serum
6 ml	L-Glutamine 100x solution
6 ml	beta-mercaptoethanol (35 microliters in 50 ml of PBS)
3 ml	Penicillin/Streptomycin 100x solution
6 ml	Nonessential amino acids 100x solution (optional)
60 microliters	LIF (ESGRO by Millipore)

EB growth medium: mix 1:1 DF medium with mESC maintenance medium without LIF

DF medium

500 ml	DMEM/F12 medium (4500 mg/l glucose)
5 ml	L-Glutamine 100x solution
2.5 ml	Penicillin/Streptomycin 100x solution
5 ml	beta-mercaptoethanol 100x solution
5 ml	Insulin Transferrin Selenium (ITS) 100x solution
20 nanomolar	Progesterone (final concentration)
60 micromolar	Putrescine (final concentration)

3.3 Cloning MicroRNA Target Sites Using HITS-CLIP on ARGONAUTE2

HITS-CLIP method has been adapted from the work of Sung Wook Chi, Robert Darnell and colleagues. [12]

3.3.1 UV cross-linking of cells

Cells are rinsed once and covered with a 5 millimeter layer of PBS. Cells are subsequently irradiated in plates with PBS, with lids off, using a total dose of 600 millijoules per square centimeter of 265 nanometer UV light (Stratalinker). Cells are scraped off the plate into the falcon tube and spin down at 400 r.c.f. for 5 minutes. After removing supernatant, the cells are resuspended in 1 ml of PBS and pelleted in an eppendorf tube by centrifugation at 500 r.c.f. for 3 minutes. After that, the liquid is removed and cells are left to freeze at -80ºC. For embryonic bodies, it is necessary to add 2% FBS to the media to avoid sticking to pipette and plates.

3.3.2 Preparing beads for IP

Prepare the beads before thawing cell lysates (amounts shown here are for a 15 centimeter plate of ES cells - approximately 100 microliters of cells). 50 microliters of Protein A Dynabeads (Invitrogen, 30mg/ml) wash three times with 500 microliters of 0.1M Na-phosphate buffer, pH 8.1. Bind 6.5 microliters of bridging antibody (Rabbit Anti-Mouse IgG, Fc, 2.4mg/ml, ImmunoResearch Laboratories, Inc.) in 500 microliters of 0.1M Na-phosphate, pH 8.1 for 60 min. at room temperature. Wash beads three times with 500 microliters of 0.1M Na-phosphate pH 8.1 and bind 26 microliters of Ago2 antibody (Abnova, clone 2E12-1C9, 0.5mg/ml) in 500 microliters of 0.1M Na-phosphate, pH 8.1 for 60 min. at room temperature. Wash beads three times with 1xPXL buffer.

1xPXL		5xPXL	
0.1%	SDS	0.5%	SDS
0.5%	sodium deoxycholate	2.5%	sodium deoxycholate
0.5%	NP-40	2.5%	NP-40
1x	PBS	5x	PBS

1xPNK

50mM	Tris pH 7.4
10mM	$MgCl_2$
0.5%	NP-40

1xPNK/EGTA

50mM	Tris pH 7.4
20mM	EGTA
0.5%	NP-40

0.1M Na-phosphate

4.66 ml	1M Na_2HPO_4
0.34 ml	1M NaH_2PO_4
0.5%	NP-40
add to 50 ml total volume	water

3.3.3 Lysis and DNase/RNase treatment

After adding second antibody to the beads take the cells out of deep freeze and thaw them on ice. Resuspend cells in 500 microliters of 1xPXL buffer (amount per 1 tube containing cells from 1x15cm plate) and incubate on ice for 10 minutes. Add 60 microliters of RQ1 DNase, mix well, incubate at 37ºC for 15 minutes with shaking 1200 rpm (Eppendorf ThermoShaker). Add 5 microliters of 1:1000 RNase A dilution (Ambion AM2272, 1mg/ml, diluted in 1xPNK buffer), mix well, incubate at 37 ºC for 5 minutes, with shaking 1200 rpm. Spin lysates at 4 ºC for 20 minutes at 13400 r.c.f. During the spin finish the preparation of the beads, so they are ready for immunoprecipitation step.

3.3.4 Immunoprecipitation, RNase digestion, Phosphatase reaction and ligation

Add cleared supernatant to washed beads, immunoprecipitate for 3 hours at 4ºC with rotation. During the incubation prepare fresh radioactive 3'RNA linker. Wash two times with 1xPXL, two times with 5xPXL, two times with 1xPNK using 500 microliters for each wash. Add 40 microlitersl of the 1:1000 RNase A dilution, incubate at 37ºC for 5 minutes with shaking 1200 rpm, wash three times with 1xPNK. Add Calf Intestine Phosphatase (CIP) reaction mixture to washed beads, incubate at 37ºC for 20 minutes shaking 1200 rpm, wash two times with 1xPNK/EGTA, wash two times with 1xPNK. Add radioactive ligation mixture to washed beads, incubate at 16ºC for 1 hour with shaking 1100 rpm for 15 seconds with a 45 second break. Add 1 microliter of 100 micromolar non-radioactive, 5‘-phosphorylated 3‘ linker, 1 microliter of 10 mM ATP and 1 microliter of T4 RNA ligase per reaction, incubate at 16ºC overnight with shaking.

3’linker radioactive labeling

1 microliter	100 micromolar 5’-OH 3’linker
5 microliters	10x PNK buffer
8 microliters	T4 PNK (NEB)
8 microliters	gamma-32P-ATP
3 microliters	RNasIN
25 microliters	water

Incubate 30-60 minutes at 37 ºC. Purify away unincorporated label by spinning through G-25 columns (GE).

Phosphatase reaction

4 microliters	10x CIP buffer (NEB buffer 3)
1.5 microliters	CIP (NEB)
1 microliter	RNasIN
33.5 microliters	water

radioactive linker ligation

4 microliter	10x T4 RNA ligase buffer
4 microliters	0.2 mg/ml BSA
4 microliters	10 mM ATP
5 microliters	radioactive 3‘ linker
1 microliter	RNasIN
1 microliter	T4 RNA ligase (Fermentas)
21 microliters	water

3.3.5 Protein electrophoresis, transfer to nitrocellulose, autoradiography

Wash beads five times with 1xPNK, resuspend in 20 microliters of Loading Buffer, incubate 15 minutes at 70ºC with shaking 1200 rpm. Load supernatant onto the 8% bis-tris NuPAGE gel (Invitrogen) with MOPS running buffer (no reducing agent in buffer) and run at 190V in cold room until the 75 kDa band is near bottom of the gel. Transfer proteins to BA-85 (pore size 0.45 micrometer) nitrocellulose using wet transfer method, preferably Novex Gel Transfer apparatus (Invitrogen) in 1x Novex Gel Transfer buffer with 10% methanol for 3 hours at 60V, 4ºC. Attach radioactive markers to the membrane (200 counts per minute), expose at -80ºC overnight using a phosphorimager screen. Scan the screen and printout the image at 100% of scale with its physical dimensions on a transparency. Overlay transparency with the membrane using radioactive markers as guides. Typically, two bands can be seen, one at 110 kDa and another at 130 kDa corresponding to microRNA and target transcript fragments loaded Ago proteins, excise separately.

Protein Gel Loading Buffer

90 microliters	1x PNK buffer
90 microliters	3x SDS sample loading buffer (NEB)
20 microliters	10x reducing agent for NuPAGE gels, Invitrogen

3.3.6 RNA extraction, 5' phosphorylation and linker ligation

Chop nitrocellulose to small pieces and add 200 microliters of 4 mg/ml proteinase K solution in 1xPK buffer, incubate at 37ºC for 20 minutes with shaking. Add 200 microliters 1xPK/7M urea, continue incubation for additional 20 minutes. Add 400 microliters of acid phenol:chloroform, continue incubation for 20 minutes with vigorous shaking. Centrifuge samples at 13000 r.c.f. for 10 minutes, collect aqueous phase. Precipitate RNA with 1 microliter glycoblue, 40 microliters of 3M NaOAc pH 4.8 and 1 ml 1:1 ethanol:isopropanol mixture. Wash once with 70% ethanol. Resuspend RNA pellet in 40 microliters of PNK reaction mixture and incubate at 37ºC for 30 minutes. Extract RNA with trizol using 500 microliters trizol with 1 microliter of glycoblue according to the standard protocol, wash once with 70% ethanol and air dry. Resuspend RNA pellet in a 10 microliter of 5’ linker ligation mixture and incubate at 16ºC overnight. Add 100 microliters of Dnase Mixture and incubate at 37ºC for 20 minutes, next add 300 microliters of water and 400 microliters acid phenol mixture, vortex for a minute, spin at 13000 r.c.f. and collect aqueous layer. Precipitate with 1 microliter of glycoblue, 40 microliters of 3M NaOAc pH 4.8 and 1 ml ethanol. Wash with with 70% ethanol, air dry.

1xPK

100 mM	Tris pH 7.5
50 mM	NaCl
10 mM	EDTA

1xPK/7M urea

100 mM	Tris pH 7.5
50 mM	NaCl
10 mM	EDTA
7M	Urea

kinase reaction

4 microliters	10x PNK buffer
0.5 microliter	10mM ATP
1 microliters	RNasIN
2 microliters	T4 PNK (NEB)
32.5 microliters	water

5'linker ligation

1 microliter	10x T4 RNA ligase buffer
1 microliter	0.2 mg/ml BSA
1 microliter	10 mM ATP
1 microliter	20 micromollar 5'RNA linker
1 microliter	RNasIN
0.5 microliter	T4 RNA ligase (Fermentas)
4.5 microliters	resuspended RNA pellet

3.3.7 Reverse Transcription, PCR, size selection and amplification

Resuspend RNA pellet in 8 microliters of water and add 2 microliters of 5 micromolar RT primer, transfer to the PCR tube and anneal at 65ºC for 5 minutes, chill on ice. Add 10 microliters of Reverse Transcriptase Mixture and incubate at 50ºC for 30 minutes, 90ºC for 5 minutes and cool down to 4ºC. Follow up with the PCR reaction using 2 to 4 microliters of cDNA. Run PCR product on 10 % denaturing polyacrylamide gel (190 V, 1x TBE buffer) with 25 bp ladder marker. Do not heat the samples or ladder prior to run (this has a specific purpose in the protocol). Visualize DNA by staining with SYBR Gold using 5 microliters for 100 ml of 1x TBE for 15 minutes. Excise bands avoiding primer-dimer products, crush acrylamide with pipette tip, soak in 400 microliters of 0.3M sodium acetate pH 4.8, 1% SDS overnight. Ethanol precipitate with glycoblue, wash with 70% ethanol and resuspend in 20 microliters of water. Use 2 - 4 microliters in next PCR amplification step, run on 2% agarose gel and extract DNA using Qiaex II resin. Purified DNA is suitable for direct sequencing using Illumina platform.

DNase reaction

11 microliters	10x DNase buffer
5 microliters	RNasIN
5 microliters	RQ1 DNase
79 microliters	water

Reverse transcription

4 microliters	5x Superscript III RT buffer
1 microliter	0.1M DTT
1 microliter	10 mM dNTPs
1 microliter	RNasIN
1 microliter	Superscript III RT
2 microliters	water

1st PCR

27 microliters	Accuprime Pfx Supermix
0.75 microliter	20 micromolar 5’ primer1
0.75 microliter	20 micromolar 3’ primer1
2 microliters	template from RT

2nd PCR

27 microliters	Accuprime Pfx Supermix
0.75 microliter	20 micromolar 5’ primer2
0.75 microliter	20 micromolar 3’ primer2
2 microliters	template from 1st PCR

3.4 Processing and Mapping Sequencing Reads

Sequencing results are processed using a combination of the FASTX-toolkit and SAMtools software packages. FASTQ files were converted to FASTA, reads not passing quality threshold were filtered and linker sequence was clipped off the 3’ end. Depending on the run, random five-nucleotide barcode was also clipped from the 5’ end and reads over 17 nucleotides were collapsed by sequence to speed up subsequent steps (abundance information was preserved in FASTA names for each sequence). Known sequencing artifacts and low complexity reads were subtracted from the dataset (see Figure 5A). MicroRNA reads were separated from the subsequent analysis by mapping to mirBase reference hairpins using nexalign with up to 3 mismatches and 1 insertion/deletion. [22] Those reads are subsequently mapped to -3p and -5p mature strands from mirBase 18 using nexalign for obtaining condition’s microRNA profile. Processed target site reads are mapped to an mm9 version of Mus musculus genome

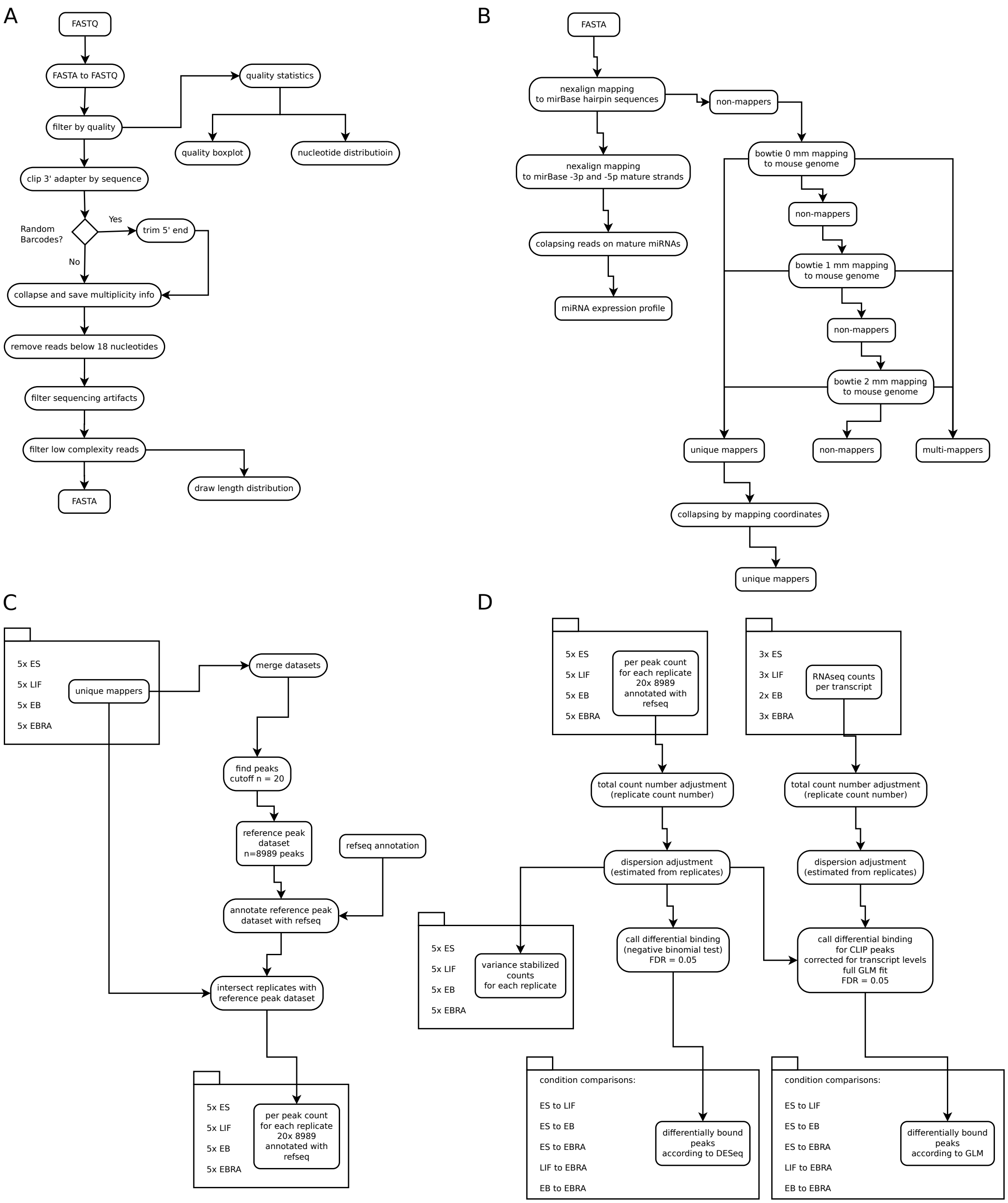

A
FASTQ
FASTA to FASTQ
quality statistics
filter by quality
quality boxplot
nucleotide distributioin
clip 3' adapter by sequence
Random Barcodes?
Yes
trim 5' end
No
collapse and save multiplicity info
remove reads below 18 nucleotides
filter sequencing artifacts
filter low complexity reads
draw length distribution
FASTA
B
FASTA
nexalign mapping to mirBase hairpin sequences
non-mappers
nexalign mapping to mirBase -3p and -5p mature strands
bowtie 0 mm mapping to mouse genome
non-mappers
colapsing reads on mature miRNAs
bowtie 1 mm mapping to mouse genome
miRNA expression profile
non-mappers
bowtie 2 mm mapping to mouse genome
unique mappers
non-mappers
multi-mappers
collapsing by mapping coordinates
unique mappers
C
5x ES
5x LIF
5x EB
5x EBRA
unique mappers
merge datasets
find peaks cutoff n = 20
reference peak dataset n=8989 peaks
refseq annotation
annotate reference peak dataset with refseq
intersect replicates with reference peak dataset
5x ES
5x LIF
5x EB
5x EBRA
per peak count for each replicate 20x 8989 annotated with refseq
D
5x ES
5x LIF
5x EB
5x EBRA
per peak count for each replicate 20x 8989 annotated with refseq
3x ES
3x LIF
2x EB
3x EBRA
RNAseq counts per transcript
total count number adjustment (replicate count number)
total count number adjustment (replicate count number)
dispersion adjustment (estimated from replicates)
dispersion adjustment (estimated from replicates)
5x ES
5x LIF
5x EB
5x EBRA
variance stabilized counts for each replicate
call differential binding (negative binomial test) FDR = 0.05
call differential binding for CLIP peaks corrected for transcript levels full GLM fit FDR = 0.05
condition comparisons:
ES to LIF
ES to EB
ES to EBRA
LIF to EBRA
EB to EBRA
differentially bound peaks according to DESeq
condition comparisons:
ES to LIF
ES to EB
ES to EBRA
LIF to EBRA
EB to EBRA
differentially bound peaks according to GLM

Figure 5. Flowcharts showing bioinformatics processing of sequencing libraries
Subfigure A shows pre-processing step performed on FASTQ files. Note that despite the collapsing step, the information about the number of sequences is preserved in FASTA file. Subfigure B shows read mapping step performed on FASTA file from previous step. MicroRNA reads are removed from main dataset to generate their own expression profile. Rest of the reads are mapped to genome using sequential mapping with increasing number of mismatches. C outlines generation of reference peak set. Subsequently replicates are intersected with the reference to generate counts for each reference target site in this replicate. D summarizes normalization steps for count data of CLIP peaks and RNA-seq transcript datasets. Calling differential binding in the dataset can be done ignoring transcript level correction (DESeq) or including it (GLM). Normalized data can also be generated at this step for numerical comparisons and plotting.

using sequential mapping by Bowtie algorithm with 0, 1 and 2 mismatches. [29] Reads mapping to the same exact genome section, but having a nucleotide mismatch are collapsed and their multiplicity is added together (see Figure 5B for diagram).

3.5 Determining a Reference Set of Messenger RNA Fragments Strongly Bound by RISC

Filtered reads from replicates for each experimental condition, namely: 5 x ES, 5 x LIF, 5 x EB and 5 x EBRA were gathered together and a peak finding algorithm FindPeaks4 was used to determine the boundaries of clusters formed by multiple reads. [15] Peaks were included in the dataset if supported by 20 or more independent sequences (which are not possible to be PCR clones), detailed diagram of this step is shown on Figure 5C. While more justified ways of establishing cutoff value could be envisioned, this value has performed adequately in filtering out noise and in reducing the total number of peaks in order to achieve significant p-values for differentially bound peaks after multiple testing correction. Additionally, closely spaced peaks which share some of the reads and thus appear as merged, were separated into subpeaks if the value between them drops to 0.6 of the maximal peak value. The resulting reference dataset consists of 8989 reference genomic intervals or peaks which represent putative RISC binding sites.

3.6 Statistical Analysis of Count Data

Subsequently, reads from each replicate of every condition are counted by intersecting them with the reference peak interval dataset and the results are saved in the matrix of conditions/replicates by peaks. In order to compare different conditions between each other it is necessary to normalize the conditions by total number of reads and also to adjust its variance in order to facilitate statistical comparison between conditions and calling differentially bound peaks. The count table has been imported into the R statistical environment and I have used DESeq package for size and variance adjustments as well as calling differential binding without correction for transcript levels. [2] Additional code authored by Dr. Simon Anders containing generalized linear model fitting has been used to call differential binding including correction for changes in transcript expression (data flow diagram for those computations is depicted in Figure 5D). [47, 3] In both cases Benjamini-Hochberg adjustment is used for multiple testing correction and False Discovery Rate of 0.05 is taken as significance cutoff. [29] The code is available as a digital file (see appendix for details). All other graphs, heatmaps, plots were generated using R statistical environment and associated visualization packages. UCSC genome browser has been used for visualization of genomic tracks.

3.7 GO Terms Analysis and Visualization of Networks

A list of genes containing differentially bound peaks was analyzed for GO term enrichment using ClueGO plugin for CYTOSCAPE network analysis software package. [53, 9, 8] Mouse GO Biological Process annotation from 23.05.2012 was used to search for statistically significant enrichment of terms with FDR of 0.05 after Benjamini-Hochberg multiple testing correction method. Groups of terms sharing similar genes were visualized as connected networks with connection length between nodes being proportional to the number of genes shared, as described in the publication on ClueGO. The connectivity was also controlled by the means of restricting network edges being drawn between nodes for which correlation is less than 0.5, as measured by the kappa score[2]. The GO term fusion setting was used along with GO term grouping, while GO term restriction setting was left at default values (between 3 and 8).

3.8 Brain Gene Co-Expression Network Association Analysis for Genes Containing Differentially Bound Peaks

The method used for this analysis has been described in the article by Gillis and Pavlidis. [19] Briefly, a gene network has been built from the co-expression data using correlations of expression profiles across samples, normalized and aggregated across experiments. This network is used to check if genes within an investigated set connect preferentially to each other to such a degree that we can recover knowledge about their identity based on their connectivity or even to predict missing members of the set sharing the same function. This property is scored by area under the curve for the Receiver Operator Characteristic curve (False Positive Rate to True Positive Rate plot). This measure represents the probability of distinguishing a positive example from a negative example, given the two. The gene networks for brain

[2] kappa score - as defined by Cohen is a statistical calculation describing agreement between classifiers which divide a set of items into a number of mutually exclusive groups. It is believed to be a better measure than a simple percentage of agreement, because it factors in agreement occuring by chance. In our case can be used as a measure of distance between two nodes of the network sharing same genes, with higher kappa representing more closely related sets.

and non-brain were constructed using co-expression data from more than fifty expression data sets, each with more than 20 samples (datasets are available at http://www.chibi.ubc.ca/Gemma/home.html). Then Gene Ontology sets of genes representing functional terms were used as a test group for validation, together with the sets of genes containing differentially bound target sites from ES to LIF, ES to EB and ES to EBRA comparisons. Genes from those comparisons were ranked by their p-value and aggregated by rank across replicates. Each condition is represented by its top 20 genes. Such constructed sets of gene names were scored using ROC metrics against brain and non-brain networks.

4 Results

4.1 Establishing a Reliable HITS-CLIP Protocol

Since its conception, the CLIP method has been applied to a wide range of RNA-binding proteins. During the time leading to the publication of the first article on microRNA target site cloning by Chi and colleagues, we have been working on our own adaptation of this protocol to ARGONAUTE2-bound transcripts. [12] The method has turned out to be surprisingly difficult to apply effectively, even despite the ongoing collaboration and a hands-on demonstration from members of the Darnell laboratory. During our optimization efforts we found several, often counter-intuitive workarounds to the problems we encountered.

It was necessary to optimize immunoprecipitation protocol used in CLIP. Interestingly, the direct binding of the antibody to protein G beads was inefficient. We were able to bind significantly more Argonaute protein using protein A bead-bound bridging antibody to which we bound our Argonaute-specific antibody, amplifying the binding capacity of the beads.

We observed that carefully controlling the amount of UV dose in conjunction with RNase A concentration used in the digestion step is indispensable for successfully carrying out the CLIP procedure. With higher radiation doses, Argonaute-containing complexes formed high molecular weight aggregates unresolvable by protein electrophoresis. To maintain reproducibility and tight footprint of the Argonaute protein on mRNA, we have introduced an additional on-bead RNase A digestion step just before the phosphatase reaction.

We also moved the 5’ phosphorylation step with polynucleotide kinase (PNK) just prior to the 5’ linker ligation reaction. Interestingly, this step has proven to be incompatible with acid phenol extraction routinely used to purify RNA in our protocol. Using scintillation counting, we discovered that during the extraction, cloned RNA was exclusively confined to the organic phase. One possible explanation is potential formation of a stable complex between PNK and RNA, and we were able to resolve this problem by substituting Trizol for Phenol at this step.

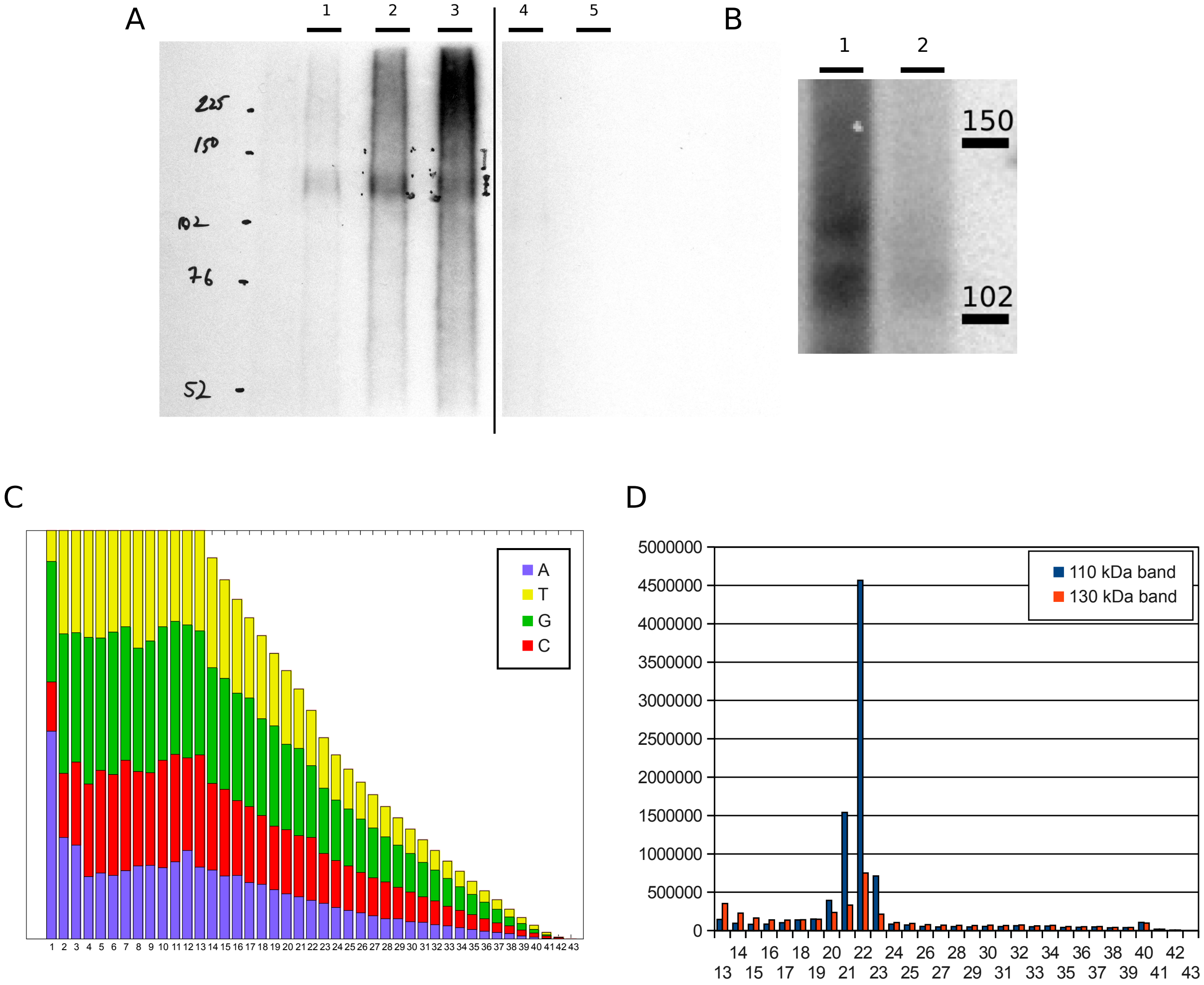
For radioactive labelling of the cloned product, we were using a ligation reaction with 5’ radioactive phosphate-labelled 3’ RNA linker. We have determined that this method of labelling is more specific than labeling cloned fragments with radioactive ATP using a direct kinase reaction. To ensure efficient cloning of target site-containing RNA fragments, after the brief period of ligation with 5’ radioactive phosphate-labelled 3' linker, the non-radioactive 5'- phosphorylated 3' linker is added to the reaction. Because the radioactive linker is 5' phosphorylated during labelling reaction by PNK, the fraction of 5' phosphate-containing RNA in this linker pool is low. The non-radioactive linker pool is phosphorylated during the synthesis and nearly all of its ends are 5' phosphorylated. This two-step ligation ensures efficient cloning, while keeping radioactivity levels within a safe range.

Our optimization experiments also show that without cross-linking, only minimal amounts of RNA remain bound to the Argonaute protein during protein gel electrophoresis phase, as can be seen in lane 7 of Figure 6A, showing one of the first successful attempts of HITS-CLIP cloning (from lanes 3, 4). In addition, none of the RNA material can be detected by autoradiography if the non-specific antibody is used, as shown in lane 8 of Figure 6A, and attempts to clone RNA material from such a non-specific immunoprecipitation trial fails. This argues in favor of the method's specificity.

Anecdotal reports have suggested that HITS-CLIP is inefficient due to a limiting cross-linking step. However, using our modified protocol, we were able to achieve success with just one 15 cm plate of Comma D cells or even a single 10 cm plate of HEK 293 cells. ES cells for quantitative experiments were cultured in 15 cm dishes and a single one was sufficient to obtain 70 - 100 ul of cell pellet. This amount generated prominent RNA:RISC complex bands, as shown on the autoradiograph on Figure 6B. For cloning from EB and EBRA cells, which are cultured as aggregates and at a lower concentration, we matched their volume to the one from ES cells by pooling material from two 10 cm plates.

It is possible to confirm that RNA being cloned co-immunoprecipitates with the Argonaute protein by performing a western blot with an Ago-specific antibody on the membrane with bound RISC following the autoradiography step. The image closely mimics the one from autoradiography with protein bands present at approximately 110 and 130 kDa. The former is believed to be enriched in microRNA cross-linked to Argonaute and migrate lower on the gel due to the discrete size of miRNAs (20 - 24 nucleotides), while the latter is mRNA-fragment-rich and its higher molecular weight and more diffuse appearance corresponds to longer and more diverse fragment lengths. This observation finds direct confirmation in our sequencing data, as plotted on Figure 6D. While the 130 kDa band still contains overabundance of reads in 20 - 23 nucleotide range, their levels are nowhere close to those in the 110 kDa band, while levels of larger fragments are slightly elevated, but still comparable to that of the lower band.

While the distribution of reads in the HITS-CLIP library is relatively uniform across almost



E

	ES	LIF	EB	EBRA
CLIP	N=5 (1214-2032k)	N=5 (142-263k)	N=5 (75-108k)	N=5 (1478-2548k)
MiRNA (CLIP)	N=5 (336- 930k)	N=5 (62-171k)	N=5 (90-161k)	N=5 (716-1099k)
RNAseq	N=3 (12.5 – 13.6M)	N=3 (13.1-14.1M)	N=2 (10.5-13.6M)	N=3 (11.8-12.5M)

Figure 6. Details of cloning process and sequencing results.

Subfigure A shows autoradiograph image of the first successful CLIP cloning attempt performed on 293S cells. While lane 1 did not have enough material for cloning, lanes number 2 and 3 contain cell material prepared according to CLIP protocol, lane 4 has non-crosslinked material and lane 5 a negative control with the material immunoprecipitated using antibody not specific to ARGONAUTE. Vertical line separates autoradiographs from two gels. Subfigure B shows CLIP cloning attempt in mouse ES cells after optimizations, as described in our protocol. Lanes differ by the amount of material used for the IP. Bands corresponding to the ribonucleoprotein complexes are visible at 110 and 130 kDa. Subfigure C shows nucleotide distribution chart of sequenced reads. D shows a distribution of read lengths cloned from lower and higher molecular weight bands. Lower band contains enrichment of reads between 20 and 23 nucleotide length corresponding to microRNA reads. Table E summarizes number of experimental replicates and amount of reads cloned for each conditions analyzed.

all positions, the first nucleotide shows enrichment for A nucleotide, possibly indicating the RNA ligase bias in 5' linker addition reaction (Figure 6C). Additionally, the number of reads longer than 13 nucleotides starts decaying fast and only half of the reads are longer than 22 nucleotides. This phenomenon is characteristic to any library sequenced on the Illumina platform and may be a source of additional noise in quantitative comparisons. Short reads tend not to map uniquely in the genome, and therefore are lost. Peak counts can be composed of any combination of different length reads and therefore be susceptible to higher noise than expected from a uniform distribution. While the method used in our comparisons is highly adaptive with respect to noise contributed in addition to Poisson counting noise, taking the length of the read into account in the library size normalization step in future algorithms may improve assessment of the count, especially in the highly dispersed low count fraction.

CLIP and microRNA profiles derived from CLIP samples have five replicates for each condition, while transcriptome libraries have three or two replicates, as summarized in Figure 6E. The number of reads in LIF and EB conditions is lower than in the two others. While this inequality is normalized for in our comparisons, it may still lead to depletion to zero in case of certain low count peaks.

4.2 Progressive Data Filtering Shows Ubiquitous Binding of RISC with Preference for 3'UTRs and CDS

In order to quantitatively compare RISC binding sites between different conditions, additionally represented by multiple replicates, it is necessary to find an adequate set of non-overlapping genomic regions over which reads will be binned. While looking at clusters of reads, it is apparent that some of the closely positioned target sites may merge, and we need to be able to find a border separating counts into two bins. Additionally, due to the multiple testing correction, we need to limit the number of such defined bins in order to be able to call differentially bound peaks. The peak detection and separation is performed by the FindPeaks algorithm, which also performs a filtering step using the number of independent sequences as a criterium. [15] This variable required adjustment, the results of which for a single sequencing library are shown on the Smad7 gene as an example.

As can be seen on a lower portion of the Figure 7A, this gene is very densely populated by HITS-CLIP reads mapping across its entire length, including intronic sequences. While we cannot make assumptions on the validity of RISC binding sites coming from any particular parts of transcript, those

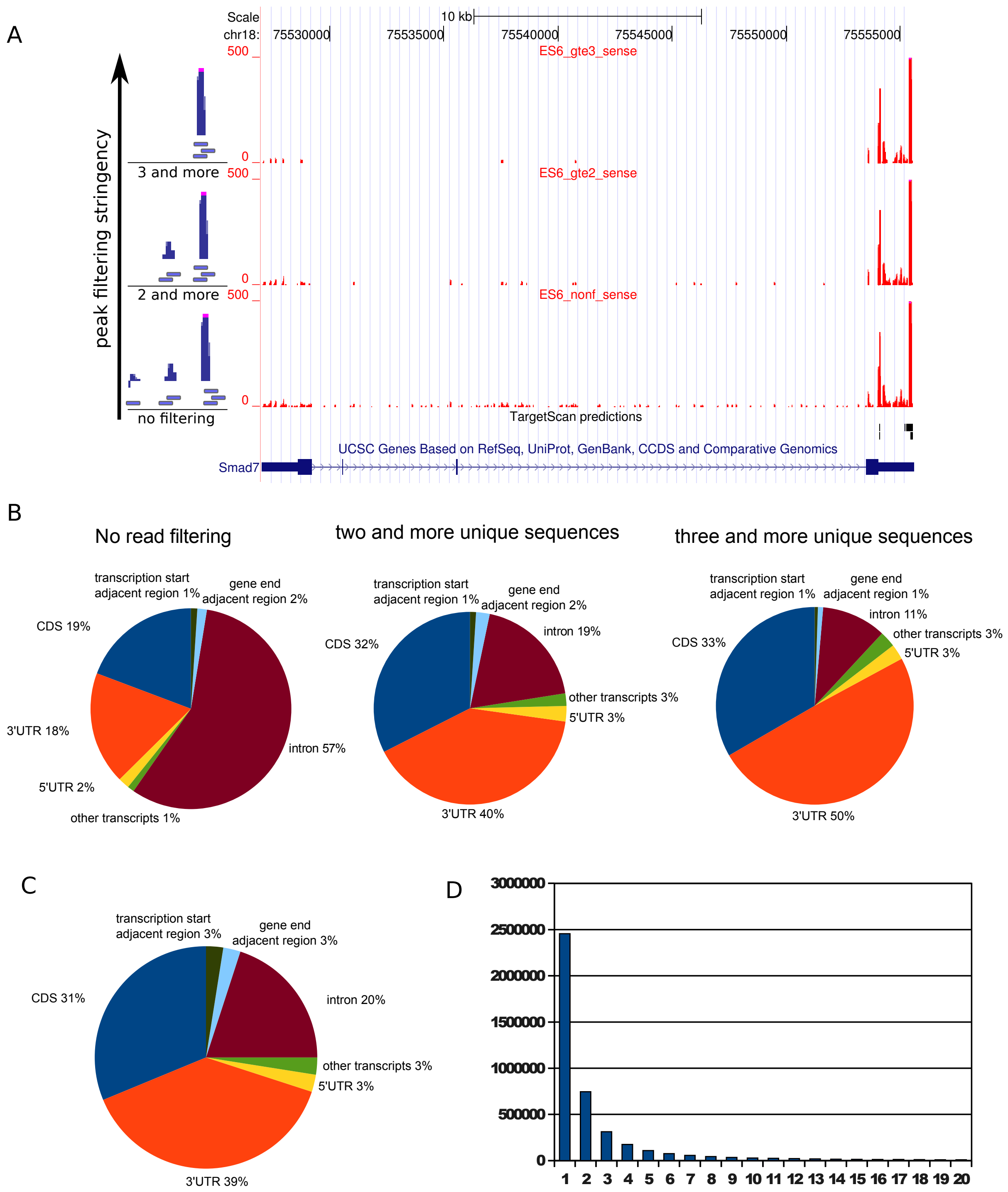


Figure 7. Library filtering and its impact on reads coming from different parts of transcripts

Subfigure A shows a Smad7 transcript extensively bound by RISC. Binding is heaviest in the 3'UTR region, in the vicinity of predicted target sites. With increasing filtering intronic reads disappear. Part B of the figure shows how annotation class breakdown changes with increased filtering performed on a single library. C shows the annotation class breakdown corresponding to the filtering level done on the set of libraries used for quantitative comparisons (20 per peak in dataset from 20 libraries). This level of filtering corresponds to the two and more unique sequences filtering level since it is remarkably similar to the middle image of panel B. Panel D shows how more gentle filtering affects number of genomic sites in the dataset of 20 libraries. The number of peaks decreases very quickly at first, but past 10 it slows down considerably.

reads do not result from random mapping, since they disappear in the regions immediately adjacent to the gene. However, it can be appreciated that truly abundantly bound regions are located within 3'UTR region, and not surprisingly, encompass predicted target sites. The Upper regions of the figure show peaks remaining after progressive filtering, requiring more than one or more than two independent reads; this filtering method results in the disappearance of most of the intronic reads and a slight decrease in the number of weak CDS and UTR reads. The consequence of filtering on the breakdown of peaks by annotation class is shown on Figure 7B. While without filtering almost 60% of peaks are located in introns and less than 40% fall into CDS and 3'UTR, with the requirement of two and more sequences supporting the peak the combined fraction of CDS and 3'UTR grows to more than 70% and with the next filtering step it grows to more than 80%, mostly at the expense of diminishing intronic read count. This is not surprising, since introns are spliced out from mature transcripts and with exception of alternatively spliced retained introns are restricted to nucleus. Additional annotation categories, like 5'UTRs, regions adjacent to transcription start site and gene end, as well as other transcripts for which partition into CDS, and UTRs is unknown, contribute only 1 to 3 % of reads. Their fraction remains relatively stable through the filtering steps. The fact that only the CDS and 3'UTR classes grow in relation to the others may indicate that those are in fact preferred, prototypical binding sites, in agreement with established knowledge.

For our quantitative comparisons, we collapsed independent reads from all twenty libraries and set a filtering requirement of twenty or more reads per peak. This resulted in filtering level comparable to the one requiring two independent reads per peak in a single library, as evidenced by comparison between Figure 7C and Figure 7B. Collapsing multiple libraries gives us the opportunity to test less strict filtering options, the results of which are summarized in Figure 7C. Without any filtering nearly 2.5 million potential RISC binding sites are found in twenty libraries; however, this number quickly falls as progressively more filtering is used. A Requirement of 5 reads gives 110 thousand potential target sites, 10 and more gives numbers below 20000 sites and from this point decreases very slowly to 8989 sites with 20 independent reads. In the end we settled on the strictest filtering criterium, believing that we can justify omitting weakly bound target sites in favor of greater sensitivity in calling differentially bound target sites for the rest of the strong sites. Rather than contemplating the validity of a given target site in context of the number of registered reads, it is arguably better to take an assumption that the abundance of reads is an indicator of how prevalent is the interaction of RISC with this region of transcript. In light of this interpretation it is better to concentrate on more highly bound target sites, since they are more likely to describe highly consequential interactions, thus justifying our stricter filtering assumptions. In addition, limiting number of sites is based on practical consideration, since statistics are typically poor for Differential Expression (Binding) of sites with fewer than 10 to 20 counts.

4.3 MicroRNA Expression in ES Cells Is Dominated by the MiR-290-295 and MiR-302 clusters

Figure 8A shows members of both microRNA clusters along with their genomic coordinates on two separate chromosomes. Despite separate locations, they share extensive homology, particularly in their seed region, sharing a consensus sequence of AGUGC (Figure 8C). Only MiR-367 shows a significantly different sequence and cannot be aligned with others. Interestingly, mature strands of the cluster can be separated into two groups based on their homology, with miR-302 cluster members grouping together with miR-291a,b, 294 and 295, as shown in Figure 8B, which depicts a phylogenetic tree calculated for cluster members. Another interesting feature of the miR-302 cluster is the first 5' U nucleotide, as opposed to the A nucleotide characteristic for rest of the group. This change is potentially of biological importance, since 5' U tipped mature strands have higher affinity towards the ARGONAUTE protein binding pocket compared to other microRNAs.

Figure 8D summarizes the expression profile of the 50 most highly expressed microRNAs across our experimental conditions. Biological replicates show very good agreement with each other, as evidenced by grouping per respective condition on the hierarchical clustering tree for the samples (Figure 8D) and corresponding high correlation coefficients (above 0.9), as depicted in Figure 8E. Incidentally, members of the microRNA clusters miR-290-295 and miR-302/367 show remarkably good coordination of their expression, as expected from the fact that they are processed from the same precursor molecule. The miR-290-295 cluster microRNAs are highly expressed in all of the conditions, particularly in the ES and LIF conditions. The miR-302/367 cluster starts to be expressed in LIF condition, however its highest expression can be detected in the EB condition.

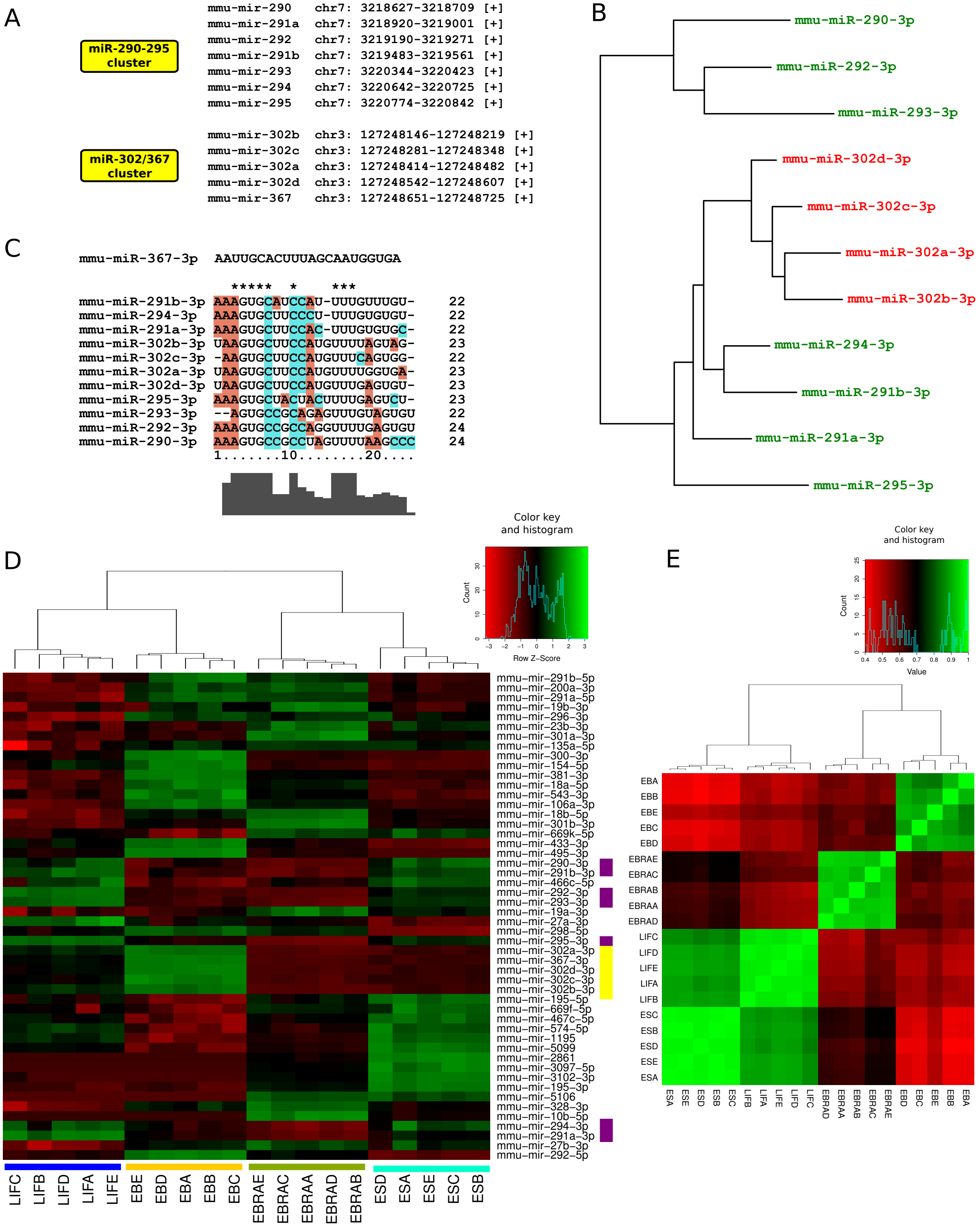


Figure 8. Major microRNA clusters in ES cells and expression profile for microRNAs in replicates.
A shows components of dominant microRNA clusters in ES cells and panels B and C show that their sequences (except for the miR-367 microRNA) are highly conserved, particularly in their seed region. D shows a heatmap of the fifty most highly expressed microRNAs, together contributing more than 99% of all microRNA reads. The numbers are normalized in respect to the row, so the heatmap shows the relative changes between conditions. miR-302/367 cluster members are emphasized using yellow color, while miR-290-295 with purple. Subfigure E shows clustering of correlation coefficients for replicates.

4.4 Excellent Correlation of the Reference Set of RISC Binding Sites Across Replicates Produces a Data Set Suitable for Differential Binding Analysis

Collapsing reads from twenty replicates over the reference set of 8989 peaks resulted in a dataset with desirable properties for quantitative and statistical comparisons. Data points between replicates show a high correlation after library size normalization and variance stabilizing transformation, which can be judged from a density scatterplot in Figure 9A. The correlation is high both for highly abundant reads, as well as for lower abundant ones. There is more dispersion in the latter, as can be judged from breadth of the distribution cloud; however its core is the most dense in reads, and remains relatively narrow, being on par with highly abundant reads. This is a desired configuration, since too much dispersion within replicates (visible as broad, cloud of points) would result in difficulties in determining statistically significant differential binding. The detailed comparison of conditions and replicates can be appreciated on Figure 9B, which shows a heatmap for the top hundred most highly bound peaks with clear examples of target sites where RISC binding consistently changes during transition from one condition to another across all replicates. Correlation between replicates and conditions is summarized in a correlation clustering plot (Figure 9C), which, similarly to the previous heatmap, recreates grouping of replicates and shows a correlation of R > 0.9 for those groups. Interestingly, LIF and EB condition appear to be modestly correlated (correlation in the range of 0.7), which may be the result to the slight under-sequencing present in those libraries.

4.5 Reference Set of Binding Sites is Well Supported By Different Antibodies and Existing ES Cell CLIP Datasets

We have tested three commercial antibodies against ARGONAUTE2 protein by performing HITS-CLIP in ES cells. All of the reads were intersected with the reference set of 8989 target sites resulting in 8250 target sites occupied by a non-zero number of reads. Both the percentage of the ES reference target site set supported by reads coming from another antibody, as well as the correlation between the samples, were assessed. Abnova antibody shows the highest level of support with 78 - 80 % of its target sites containing reads in other antibodies (As shown on Figure 10A). The correlation is the highest between Abnova antibody and either Sigma (0.67, Figure 10D) or SantaCruz (0.55, Figure 10C), while the comparison between Sigma and SantaCruz is lower (0.37, Figure 10B). Since Abnova

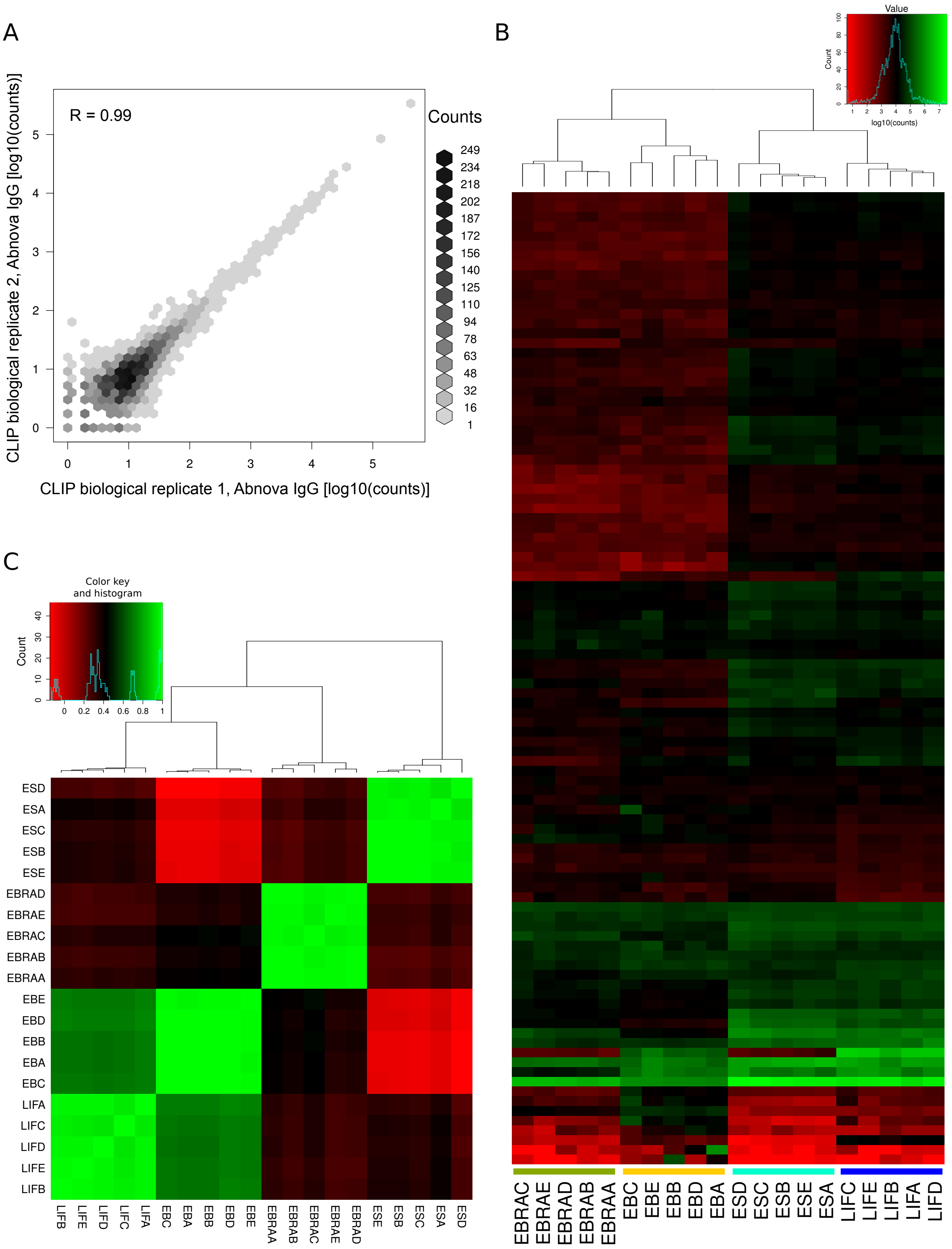


Figure 9. Experimentally derived target site counts show good correlation between condition replicates. Subfigure A shows count density scatterplot between two biological replicates of ES condition. Subfigure B presents a heatmap of a hundred most highly bound target sites in our dataset. C shows clustering of correlation coefficients between replicates.

A

	Abnova	SantaCruz	Sigma	Leung et al. data
CLIP ES (Abnova Ab) n=8250	52.4	40.3	33.3	57.3
Abnova Ab n=4383	100.0	59.7	51.5	64.6
Sigma Ab n=2763	81.7	66.6	100.0	69.8
SantaCruz Ab n=3350	78.1	100.0	54.9	65.2
Leung et al. data unknown Ab and cell line n=5047	56.1	43.3	38.2	100.0

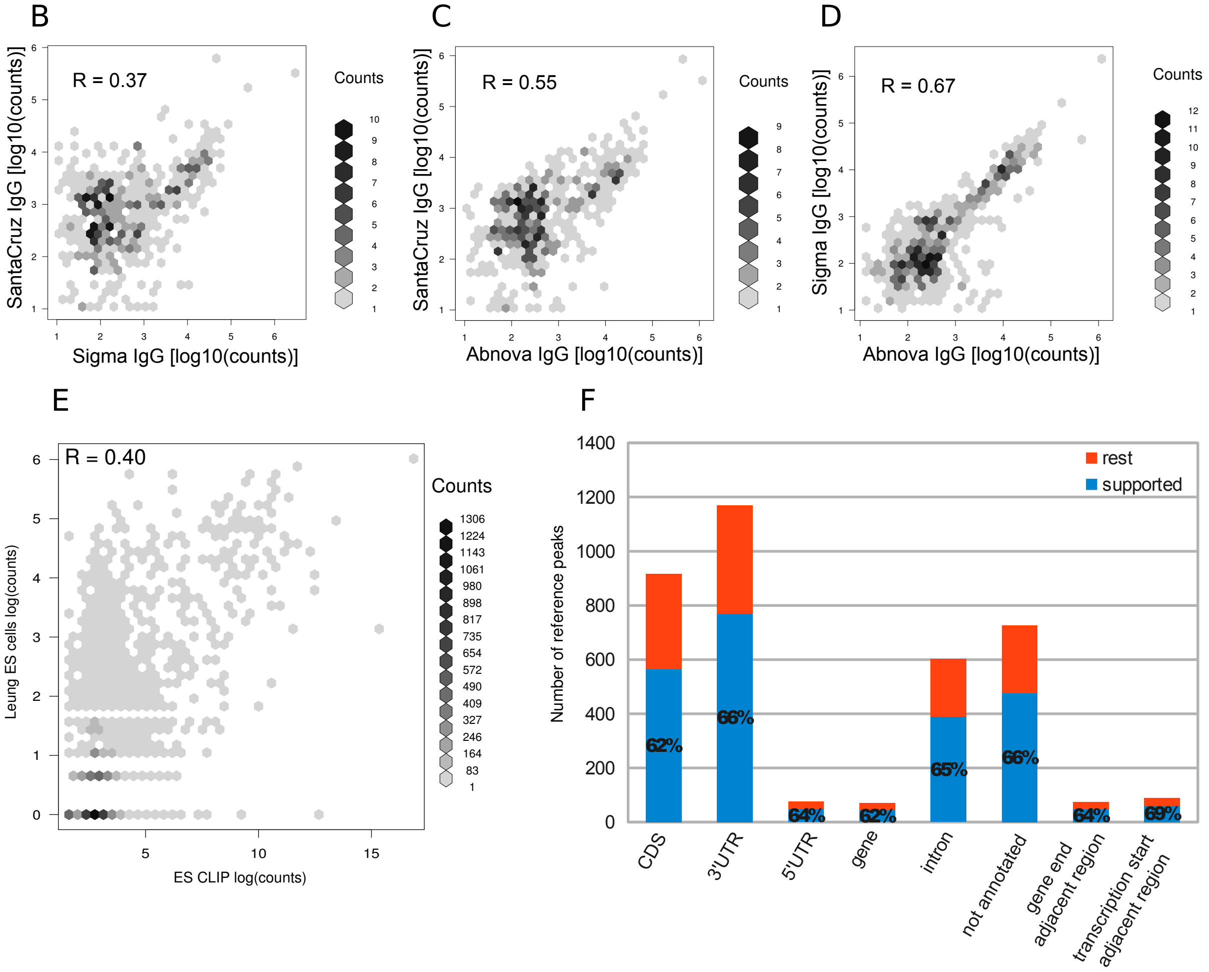


Figure 10. Cross-validation of the reference set of 8989 target sites with aditional antibodies and published HITS-CLIP stem cell data.

Subfigure A shows a table of how many reference target sites are supported by reads coming from other datasets. Cells show percentages of level of support by datasets specified in columns for set of target sites shown in rows. (target sites must be greater than zero in row datasets, hence different number of target sites in rows). B to D show count density on scatterplots for comparisons between antibodies. Subfigure E shows similar scatterplot for data published by Leung and colleagues compared to counts from our ES cell target sites, while F shows the level of support by this dataset for annotation categories in our reference target site set.

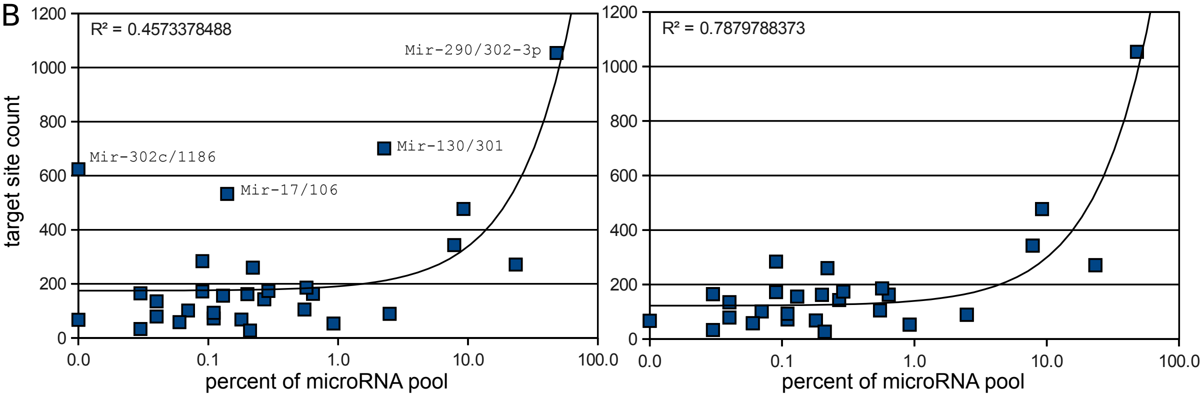
shows the highest correlation paired with two other antibodies, all other HITS-CLIP experiments were performed using this antibody. When libraries from biological replicates were cloned in parallel, a very high level of correlation was observed between replicates, as shown earlier on Figure 9A. This result also signifies that while quantitative comparison between the conditions is possible within the same experiment, the use of different cell types, antibodies and finally slightly different procedures used for HITS-CLIP cloning may preclude such comparison with experiments performed by another group of researchers. Even in this case, however, we can compare datasets qualitatively, looking for overlapping target sites, or differentially bound target sites in the same genes. We have performed such comparison against an HITS-CLIP dataset generated by Leung, Young and colleagues for mouse ES cells. [41] This comparison shows that the support for target sites in any of the antibody immunoprecipitations is high, reaching over 60% (Figure 10A). Interestingly, the support of Leung's data to the reference set of peaks does not show much variation in relation to whether the target site lies in the 3'UTR, 5'UTR, CDS, intron or other annotation classes, as shown on Figure 10F. This may again indicate that our filtering method, by changing proportions between annotation classes, ultimately arrives at a set of RISC binding sites with strong experimental support. Unfortunately, correlation between our CLIP data and that of Leung's et al., while still positive, appears not to be high, reaching only 0.4 (Figure 10E).

4.6 MicroRNA Seed Families Reduce the Complexity of the MicroRNA Expression Profile

The full count table for mature microRNA strands detected in our experiment has over one thousand positions. Many of the rows correspond to microRNAs which have been detected in a single or a few replicates with just a few counts. Additionally, many of the microRNAs share an identical seed site or have a seed site which is shifted by one nucleotide with respect to related seed sites. It is very likely that those microRNAs act together in controlling a set of target sites, possibly with slightly different affinities. In order to manage the analysis of the impact of the microRNAs on their targets I have simplified their expression profile by collapsing counts for microRNAs sharing the identical seed site. Figure 11A shows a table of thirty of the most highly expressed seed site groups across all of the experimental samples. Firstly, those families constitute between 98.11% to 99.96% of all microRNA reads present in our datasets. MicroRNA families that are not included on this list most likely contribute

A

MicroRNA seed family name	microRNAs in family	ES [%]	LIF [%]	EB [%]	EBRA [%]	seed site	number of seed sites found in the main set of target sites (n=8989)
MiR-290/302-3p	m290-3p, m291a-3p, m291b-3p, m292-3p, m294-3p, m295-3p, m302a-3p, m302b-3p, m302d-3p, m350-5p, m467a-5p, m467c-5p, m467d-5p	94.7	99.4	76.1	48.1	AAGTGC	1054
MiR-27	m27a-3p, m27b-3p, m673-5p	1.8	0.3	5.0	23.3	TCACAG	271
MiR-290-5p	m290-5p, m292-5p, m293-5p, m294-5p, m295-5p	0.3	0.2	10.2	7.8	CTCAAA	343
MiR-19	m19a-3p, m19b-3p	0.4	0.1	0.7	9.2	GTGCAA	477
MiR-367/92	m25-3p, m32-5p, m363-3p, m367-3p, m92a-3p, m92b-3p	0.0	0.0	4.1	0.1	ATTGCA	284
MiR-18	m18a-5p, m18b-5p	0.0	0.0	0.2	2.5	AAGGTG	89
MiR-130/301	m130a-3p, m130b-3p, m301a-3p, m301b-3p, m721	0.0	0.0	0.1	2.3	AGTGCA	701
	m1192, m495-3p, m7a-1-3p	0.0	0.0	1.2	0.1	AACAAA	172
	m5106	0.2	0.0	0.0	0.9	GGTCTG	53
	m298-5p	0.0	0.0	0.3	0.6	GCAGAG	163
	m291a-5p, m291b-5p	0.0	0.0	0.3	0.6	ATCAAA	186
	m106a-3p, m17-3p, m1954, m20b-3p	0.0	0.0	0.5	0.3	CTGCAG	143
	m195-3p	0.6	0.0	0.0	0.0	CAATAT	67
	m10a-5p, m10b-5p	0.0	0.0	0.0	0.6	ACCCTG	105
	m3088-3p, m433-3p	0.0	0.0	0.2	0.1	TCATGA	72
	m141-3p, m200a-3p	0.0	0.0	0.1	0.3	AACACT	174
	m543-3p	0.0	0.0	0.3	0.0	AACATT	135
	m669k-5p	0.0	0.0	0.0	0.2	GTGCAT	260
	m381-3p, m539-3p	0.0	0.0	0.2	0.1	ATACAA	101
	m200b-3p, m200c-3p, m429-3p	0.0	0.0	0.0	0.2	AATACT	68
	m100-5p, m99a-5p, m99b-5p	0.0	0.0	0.0	0.2	ACCCGT	27
	m181a-5p, m181b-5p, m181c-5p, m181d-5p	0.0	0.0	0.0	0.2	ACATTC	162
MiR-17/106	m106a-5p, m106b-5p, m17-5p, m20a-5p, m20b-5p, m93-5p	0.0	0.0	0.0	0.1	AAAGTG	533
MiR-302c/1186	m1186, m302c-3p	0.0	0.0	0.2	0.0	AGTGCT	624
	m23a-3p, m23b-3p	0.0	0.0	0.0	0.1	TCACAT	156
	m300-3p	0.0	0.0	0.1	0.0	ATGCAA	79
	m328-3p	0.0	0.0	0.0	0.1	TGGCCC	93
	m296-3p	0.0	0.0	0.1	0.1	AGGGTT	58
	m293-3p	0.0	0.0	0.0	0.0	GTGCCG	165
	m154-5p, m5118, m872-5p	0.0	0.0	0.1	0.0	AGGTTA	33
		98.11	99.96	99.81	98.29		



C

```
Mir-290/302       AAGTGC        1054
Mir-130/301        AGTGCA       701
Mir-302c/1186      AGTGCT       624
Mir-17/106       AAAGTG         533
Mir-19              GTGCAA      477
Mir-669k            GTGCAT      260
Mir-23              GTGCCG      165
```

Figure 11. Composition of the microRNA seed families and their expression in relation to target sites
Subfigure A shows a table of percentages for microRNA seed families in different conditions, along with the number of corresponding seed sites. Families with more than 0.5% of reads per condition are highlighted in grey. Short names are provided for families referenced in text. B shows plots of seed site count (linear axis Y) in respect to matching microRNA family fraction (percentage on log axis X). Linear trend line is fitted to datapoints showing improvement of coefficient of fit when three outliers are taken out. Subfigure C shows alignment of microRNA seeds with common GTG motif.

very little to post-transcriptional control of transcripts in our samples. It is important to keep in mind, however, that lowly expressed mature strands can still exhibit enormous impact in the case of perfect pairing to the target site. Secondly, the distribution of reads across the families is highly skewed towards the most highly expressed members. The table positions of 0.5% of total reads or above are shaded in gray on the figure, making it immediately clear that the bottom half of the list is poor in counts, with positions above 0.1% being sparsely distributed. Interestingly, the extent to which the distribution of represented miRNA families is skewed changes with differentiation. While ES and LIF are primarily MiR-290/302-3p seed family rich (with a small additional fraction of MiR-27 and MiR-195-3p in ES), the relative fraction of all other seed families increases in EB and EBRA.

By combining miRNA abundance data with experimental data on RISC binding positions, we can try to assess the influence of microRNA expression on corresponding target sites. One way to perform such an assessment is to count the number of target sites containing sequence complementary to the microRNA seed sequence. The importance of the seed site for microRNA targeting is well established, and while other modes of targeting were previously reported, it still explains the identity of the biggest fraction of validated RISC binding sites. [10] This suggests that this simple measure, while not exhaustive, will be a relatively robust way of judging microRNA impact and we have included it in the table as the rightmost column.

Confirming the assumption above, the dominant MiR-290/302-3p seed family has the highest number of matching complementary sites in the reference targets dataset, the upper half of the table has many positions above 200, while the lower half has most positions below 200 counts. The correlation is not perfect, with a few outliers, which become visible if the dependency of the target site counts on microRNA read fraction is plotted. Due to the high skew of the read count distribution, it is necessary to plot it on a logarithmic scale, while corresponding target site count, with more uniform distribution, can be plotted on linear scale. Figure 11B shows the resulting plots, together with the fit of linear trend line to the data points and resulting coefficient of determination. The left plot shows all data points and it is immediately clear that while most of the points fit the linear regression curve well, there are three outlying data points corresponding to MiR-130/301, MiR-17/106 and MiR-302c seed families. If we align seed sites for those microRNAs with other similar ones, as depicted on Figure 11C, it is visible that among the seed sites sharing a central GTG motif, those three share five out of six nucleotides with the most abundant MiR-290/302 family. If we remove those outlying points, as depicted in the right plot of Figure 11B, the coefficient of determination for the linear trend line drastically increases from 0.46 to 0.79. While both values implicate positive correlation between the variables, the latter is considered a fairly good fit with a corresponding Pearson correlation coefficient of 0.89. While a few high count points dominate the plot, it can be appreciated that the trend line goes across the middle of the more numerous cloud point, suggesting a linear relationship between microRNA levels and target site counts.

One minor, but surprising finding is the experimental confirmation of functionality for passenger strands accompanying some of the microRNAs. The MiR-290-5p seed family is composed of five members of the miR-290-295 cluster which together provide 8 - 10% of reads in EBRA and EB conditions than in ES and LIF where passenger reads are essentially absent. This fact is surprising on its own, since the fraction of reads of its corresponding guide strand family is lower in those conditions. The MiR-290-5p seed family has a seed site which is dramatically different from its major guide strand family and other major microRNAs present in the pool, thus making the abundance of the passenger family the only explanation for high number of corresponding seed sites.

The MiR-27 family is moderately abundant in ES and EB conditions (1.8 - 5%), while the fraction of its reads increases significantly in the EBRA condition. The corresponding seed site is unique to this family and its complementary sequence is adequately present within target site sequences. However those associated target sites do not show significant increase in their binding to RISC in any condition, the fact which may have something to do with relatively high starting level of the MiR-27 in the ES condition.

The Mir-19 family seed site comes as one of the most abundant and specific, since it is expressed primarily in the EBRA condition at about 9%, while also present at 0.7% in EB condition. As can be seen in the seed site alignment (Figure 11C), it does share a GTGC motif with the MIR-290/302 family, thus making it possible that at least part of its complementary target sites fall under joint regulation. Evidence for the contrary comes from the fact that the target site number is close to the regression line, thus arguing that the extent of inflation due to the joint regulation is not large. This microRNA is also known to be specific to neuronal tissue, making its expression in the EBRA condition functionally relevant. [33]

Mir-18 is another example of a miRNA restricted to the EBRA condition. Interestingly, its seed site is dissimilar from the MiR-290/302 family motif at its second position, despite containing a central GTG motif. This feature seems to be taking its corresponding target site away from simultaneous influence of the MiR-290/302 family motif, as evidenced by a low count of corresponding target sites. MiR-18 appears to represent another example of a neural differentiation-specific miRNA.

Another neuro-specific family is MiR-130/301, expressed to a similar extent as MiR-18 in the EBRA condition. In contrast to the latter, its seed site is very similar to MiR-290/302 and appears just shifted, sharing five out of six nucleotides. As a result, the number of its target sites is abnormally high with respect to its abundance suggesting an inability to deconvolve similar sites.

4.7 Search for Overrepresented Sequences in the Reference Set Target Site Data Identifies Motifs Corresponding to Reverse Complement of Seed Sites from Most Highly Expressed MicroRNAs

Reasoning that the most abundant microRNAs should induce enrichment of target sites with sequences complementary to their seed site, I have performed a motif search on the reference set of target sites using DREME program, a part of the MEME package. [6] It looks for over-representation of motifs in supplied sequences, as compared to the same set of sequences after di-nucleotide shuffling transformation, and calculates the statistical significance of enrichment for the identified motifs. Figure 12A shows a plot of relative abundance drawn for the most highly expressed microRNA seed families during the process of differentiation, corresponding closely to the table from Figure 11A. As can be appreciated, most of the families can be linked with a motif complementary to their seed site derived de novo from the target site data. Those motifs are composed of sequences enriched in comparison to the set of remaining possible motifs, a fact which finds correspondence in their statistical significance. One notable exception is MiR-27, which was not found to yield a matchable motif, despite 271 direct matches in target sequences. Additionally, the miR-18 motif can be found only if matching by a bulged seed is assumed. To further investigate which part of the transcript is the canonical place of action for the microRNA, we have divided the reference target site set according their annotation class: 3'UTR, CDS, 5'UTR, intron, other gene, transcription start adjacent region, and gene end adjacent region. Interestingly, only the two first classes have multiple significant motifs, as shown in Figure 12B. While it can be argued that in case of low abundant classes this may have been due to the small number of sequences, the intron class should have enough reads for statistically significant motifs to be detectable. Interestingly, there are fewer CDS motifs and they correspond mostly to the MiR-290/302-3p seed family, while motif-rich

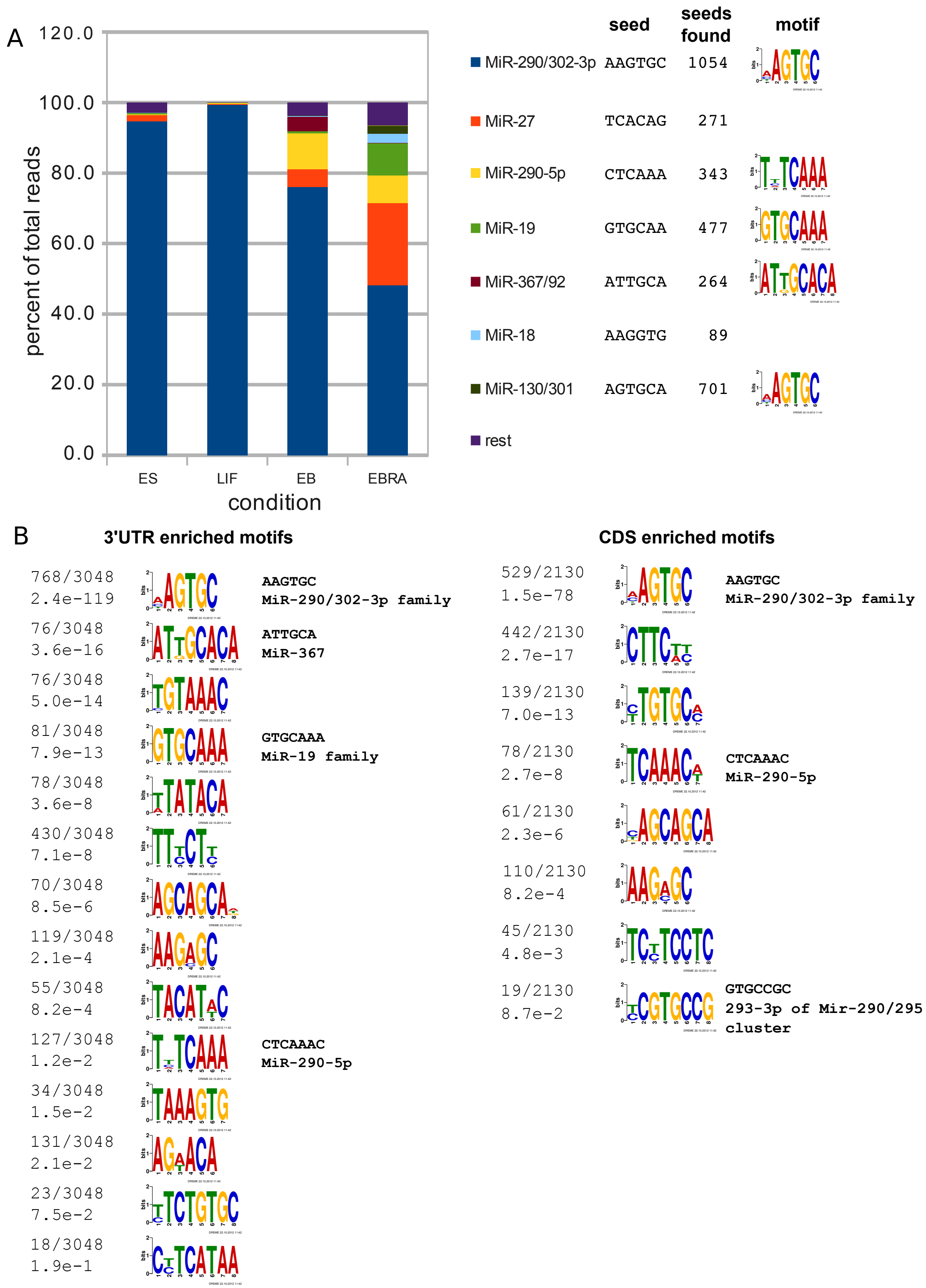


Figure 12. Unsupervised motif search in reference peak set yields sequences assignable to microRNAs. Subfigure A shows expression profiles for seven most highly expressed microRNA seed families, along with motifs matching their corresponding seed sites. Panel B shows results of de novo motif search in fragments coming from 3'UTRs and CDS. All of the motifs shown have been found to be statistically significant with corresponding p-values and number of sites found listed next to the motif.

3'UTR contain enriched sequences corresponding to at least two more major microRNA families and additional unassignable motifs.

4.8 Impact of MicroRNA Level Change on RISC Binding to Matching Target Sites

Using a table of read counts for the reference set of target sites, we can normalize libraries for size and use a variance stabilizing transformation to derive a table of variance stabilized counts to perform various comparisons. In particular, we were interested in exploring whether the changing levels of microRNAs between conditions show measurable impact on their associated target sites. Since we concluded that our dataset is rich in seven microRNA seed families and the most diversified condition is EBRA, we have performed a comparison of target sites changing between ES and EBRA. For the purpose of identifying target sites corresponding to microRNA families we have used a simple matching with a canonical seed site defined as hexamer of nucleotides 2 to 7 of microRNA mature strand. Panels A-D of Figure 13 summarize results showing ECDF[3] plots for various microRNAs in relation to target sites not matching top seven microRNA families (black line crossing 0, 0.5 point).

Interestingly, two microRNAs, namely miR-19 (panel A) and miR-367 (panel C), exhibited a highly shifted distribution of their target sites. miR-367 is an interesting case, since its expression is significant in the EB condition, preceding EBRA, however binding to its target sites is increased more in the EBRA condition. Since target sites for the miR-290/302 cluster family also showed significant shift, its distribution is also drawn for comparison. However, while the median of fold changes for miR-367 and miR-19 is close to four (ECDF lines crossing line y=0.5 close to x=2 and log2(4)=2) compared to other sites, the median for miR-290/302 cluster is just two. It is surprising that despite having a slight decrease in the total microRNA fraction of miR-290/302 family in the EBRA condition, we observe a slight increase of binding to its corresponding seed sites. This may have something to do with increased expression of miR-302 cluster with U-tipped mature strands which are believed to have higher RISC binding affinity. Another possibility is the increase of cellular microRNA pool itself in the EBRA condition, which is a simpler explanation of this phenomenon. In any case it seems advisable to use miR-290/302 family ECDF curve as a base reference for shifts since we prefer to be conservative in

[3] Experimental Cumulative Distribution Function is the function associated with the empirical measure of the sample. In our case a single value of this function can be interpreted as a probability of corresponding data point being lower than a specific $\log_2$(fold change). Therefore curves shifted more towards the right side of the plot represent distributions of datapoints having higher fraction of increasing values than those on the left.

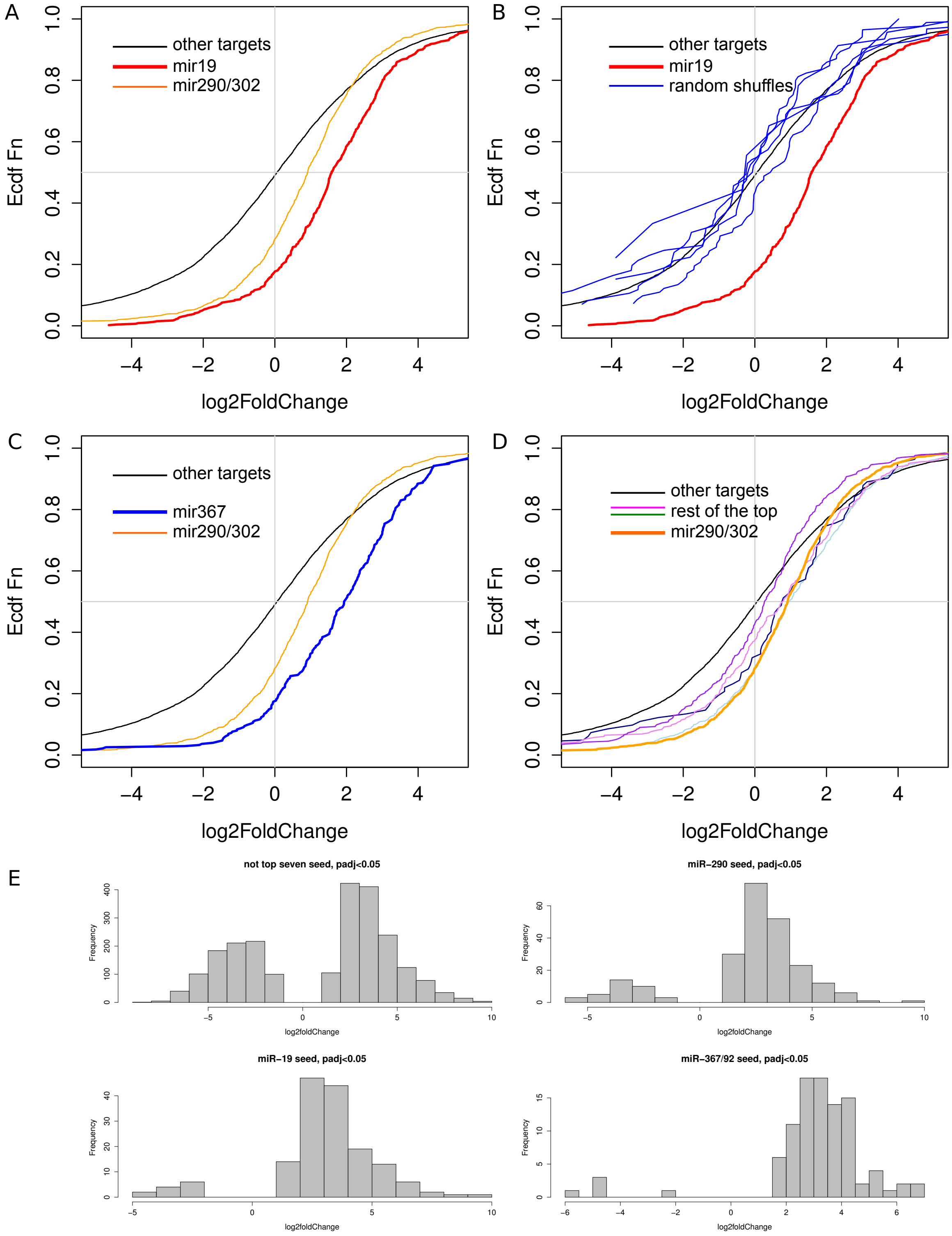


Figure 13. Impact of microRNA expression changes on matching target sites
Plot A shows comparison of ECDF functions plotted agains log2(Fold change of CLIP peak height) for target sites containing miR-19 and miR-290/302 cluster seeds, as well as sites not matching seeds from top seven microRNAs. B shows ECDF plots for miR-19 and sequences formed by randomly shuffling this seed. C shows miR-367 target sites showing similar increase in binding to miR-19, while D shows the other four microRNA families in relation to miR-290/302. Resulting fold changes for differentially bound peaks corresponding to miR-19, miR-367, miR-290/302 and rest are shown in subfigure E. Only statistically significant fold changes were used for creating the histogram.

our conclusions.

We have compared both miR-19 and miR-367 distributions to that of miR-290/302 using two-sample Kolmogorov-Smirnov test with highly significant results (p<5.342e-11 and p<4.47e-09 respectively). Since miR-19 seed sequence shares similarity with that of miR-290/302 family, we need to check if subsets of miR-290/302 family could form distribution similar to that of miR-19. For check for this distributions of samples constituting random subsets from miR-290/302 target site set were compared to their originating distribution and were shown to always score above p-value of 0.05 (0.15, 0.81, 0.53) in Kolmogorov-Smirnov test. This indicates that the target sites of miR-19 and miR-367 cannot be considered a random sample of target sites taken from the set corresponding to the dominant miR-290/302 family. Therefore shift in their distributions towards more positive values indicates increased binding by RISC which is inherent to their respective microRNA affinities and is unlikely to come from miR-290/302 cluster members. Panel B shows the ECDF plot for the distribution of target sites matching the miR-19 shuffled seed and miR-19 itself. Those sequences are found much less often, hence the jagged appearance of their ECDF curves which follow closely an ECDF function for target sites without matches to top seven microRNA families. We conclude therefore that the higher binding of miR-19 target sites to RISC registered by us is not due to the influence of related motifs and is unique to miR-19 seed.

Finally, panel D shows miR-290/302 sites plotted together with miR-290-5p, miR-18, miR-130/301 and miR-27. The miR-290-5p plot is almost identical in shape to that of the miR-290/302 family, but is significantly shifted to the left (KS test: p<3.215e-06). Other microRNA ECDF curves run very closely to the miR-290/302 ECDF, slightly jumping ahead of it only in the part of the plot containing most highly RISC bound sites. Due to this closeness, p-values for comparisons of those distributions to that of miR-290/302 are relatively high, being: p<0.677 for miR-18 (not significant), p<0.014 for miR-130/301 and p<0.019 for miR-27. Although some of them are significant, closeness to the miR-290/302 ECDF makes it hard to interpret this difference as having measurably stronger biological effect than the generic increase of count visible in the case of miR-290/302, which does not have an underlying positive increase of its microRNA counts. However one significant difference from miR-290/302 target site distribution seem to be higher changes registered in the target sites with highest RISC binding, as mentioned before.

While ECDF plots show distribution of fold changes for all target sites it was interesting to check if just the set of target sites considered to be differentially bound show consequences of increase for EB/EBRA specific microRNAs. Panel E of Figure 13. shows histograms of statistically significant (FDR <0.05) fold changes for target site-matching microRNAs. While all of the plots appear as asymmetric, in case of miR-19 and miR-367 depletion of data points representive negative logarithm of fold changes is particularly severe, visible as disapperance of the bars from the left side of plot.

4.9 Conservation and TargetScan microRNA Target Site Prediction Support For Reference Target Site Set

Additional support for a 3'UTR bias among our reference peak sites comes from intersecting them with the widely used TargetScan target site database. [17] This resource contains the locations of the phylogenetically conserved microRNA target sites predicted by matching transcript fragments with microRNA seeds. In our reference dataset, 972 of 3048 3'UTR sites were supported by this prediction. While the prediction algorithm requires exact seed matches, our experimentaly derived target sites may contain examples of seedless sites which may pair with microRNA using alternative binding modes. Additionally, our set of target sites does not have restrictions on how evolutionary conserved target sites need to be. Those properties may explain why not all of our sites are supported by corresponding prediction.

We can take advantage of the opportunity provided by our experimentally derived set of target sites and investigate whether they really show co-localization with conserved regions of the genome. The results of this analysis are shown in Figure 14A, which shows three examples of transcripts, plotted together with TargetScan predicted sites, as well as the experimental data and a plot of conservation score derived from thirty mammalian species (PhyloP track). [48] Those example plots show that all three annotation tracks often overlap. In order to find out whether the trend for increased conservation is true for our target sites, we have examined the distribution of their average PhyloP scores. Figure 14B shows the result of the comparison between the scores for our target sites and a matched set of random intervals coming from 3'UTRs. The distribution of scores for target sites is visibly shifted towards more positive values. Target sites with scores above the first quartile are positive, indicating conservation of those sites. Therefore, we can conclude that about three-fourths of our experimentally determined target sites are located in relatively conserved regions. In comparison, the mean of the random sample, which is our negative control, is close to zero. The Kolmogorov-Smirnov test indicates high significance of

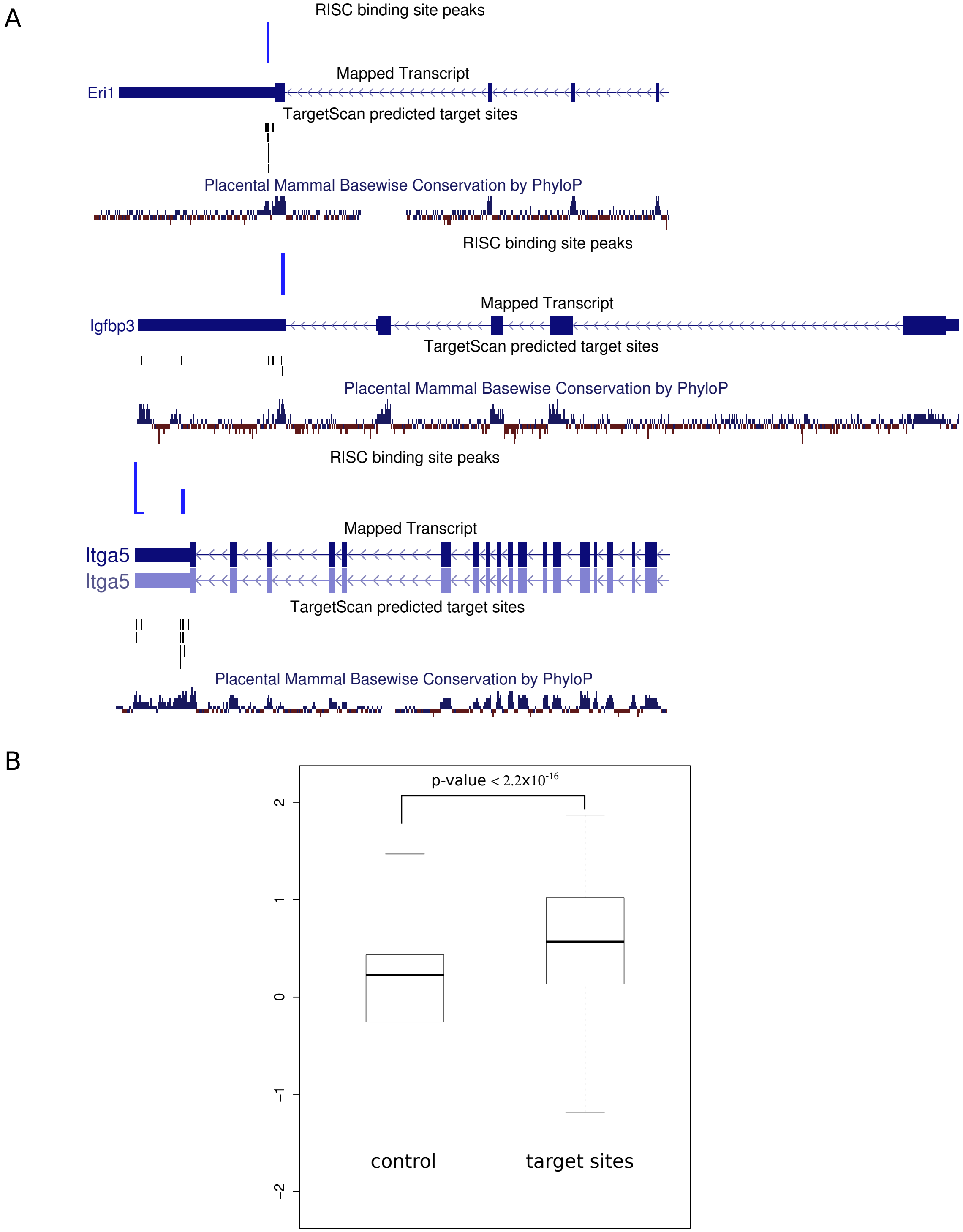


Figure 14. Conservation of the 3'UTR target sites.
Subfigure A shows a set of example target sites located in 3'UTR regions, along with their PhyloP score plot showing conservation of the underlying nucleotide sequence. While most of the 3'UTR regions have lower conservation score compared to CDS, those falling within the region of experimentally located microRNA target sites appear to be conserved. This notion appears to be confirmed by our experimental data, with target site regions scoring significantly higher than random sample of 3'UTR regions, as can be seen in subfigure B.

this shift with the p-value of the test being less than 2.2e-16.

4.10 Long Non-coding RNA are Among Targets of RISC

Figure 15, panel A shows the distribution of experimental target sites per transcript. Transcripts controlled by a single target site are the largest group identified and virtually all of the transcripts have less than ten sites. Interestingly, some of the annotated transcripts featuring the highest number of microRNA binding sites belong to the long intergenic non-coding RNA (lincRNA) class. Most notably MALAT1 and NEAT1 are two lincRNAs located in close proximity on chromosome 19. They feature multiple reads along their entire length (see Figure 15, panel B), unique feature differentiating them from transcripts with at most a few, punctuated and separated sites. We have searched for similar examples of read clustering and were able to find a handful of examples. Interestingly, some of them belonged to the Scarna (small cajal body-specific RNA) class of non-coding RNAs and we systematically checked members of this group, identifying additional examples. This group of transcripts forms extensive secondary structures, as pictured on panel C of Figure 15, which makes it possible for them to be recognized and processed by DROSHA containing complex. Incidentally, MALAT1 is also known to be bound by DGCR8, a DROSHA complex member, which potentially suggests that it is processed and in consequence also incorporated into the RISC complex. [45] The histogram of read lengths for MALAT1 (Figure 15, panel D) shows a broad distribution of reads, and although the reads from nucleotides 18 to 24 form the peak of the distribution, it is not highly enriched and longer reads are also present in substantial amounts. This suggests that detected MALAT1 reads cannot be exclusively derived from processing by DROSHA and DICER and are at least in part the result of RISC binding to MALAT1 transcript. It is known that MALAT1 is processed by RNase P to yield a MALAT1-associated small cytoplasmic RNA, an ~60 nucleotide long fragment of secondary structure similar to that of tRNA. This is an unlikely target for processing, while it needs to be established what kind of secondary structure rest of MALAT1 transcript form in vivo and if it can be target of DROSHA-containing complex. Interestingly, while MALAT1 transcript is uniformly covered with CLIP reads, our attempts of predicting microRNA binding sites have failed to yield corresponding enrichment. Finally, MALAT1 reads were found in comparable abundance in our CLIP experiments performed on human cells, suggesting it as a conserved target for RISC complex.

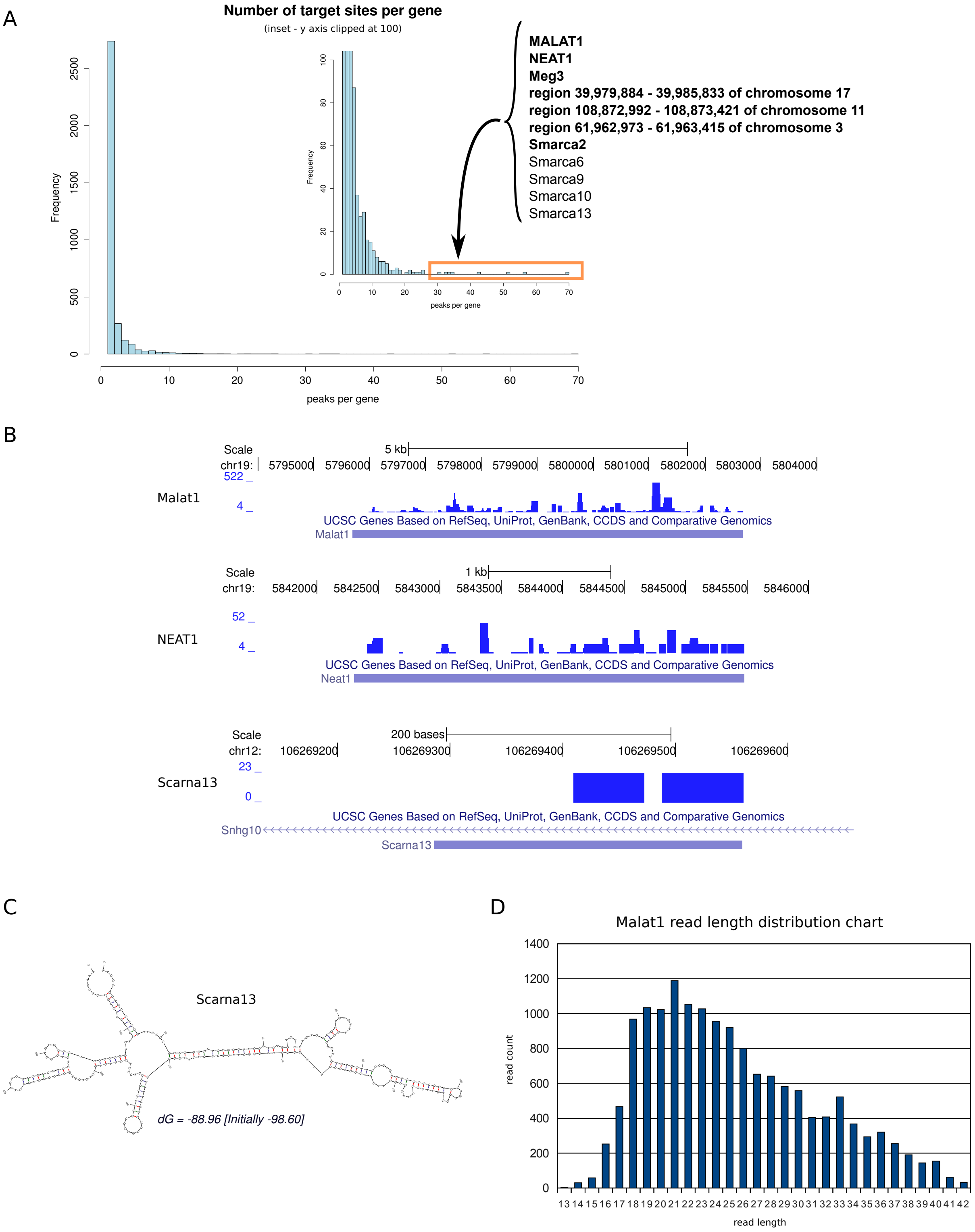


Figure 15. Number of peak sites per gene reveal lincRNAs as transcripts abundantly bound by RISC
Subfigure A shows distribution of the number of peaks found per gene. Some of the uncommon, most highly bound genes belong to a lincRNA class. Subfigure B shows three examples of such transcripts. C shows an outline for putative secondary structure of Scarna13 non-coding RNA. Subfigure D shows distribution of read lengths over the MALAT1 transcript.

4.11 Changes of Underlying Transcript Levels Are Not a Major Factor Influencing Target Site CLIP Reads

Figure 16, panel A shows CLIP reads binned over a length of transcript plotted against its RNA-seq counts in one of the ES samples. Transcripts highly expressed in ES cells have few corresponding CLIP counts and the highly RISC-bound population is restricted to relatively low RNA-seq counts. Therefore, there is no evidence of a strong linear dependence or even correlation between those two variables, as evidenced by the very low Pearson correlation coefficient (0.049). Even if we compress the scales and plot the same data points on a log-log scatterplot (Figure 16, panel B), the result is a spherical cloud of points, closely resembling a distribution produced by two independent variables. To further check this independence, I have plotted the fold changes between ES and EBRA conditions in CLIP and RNA-seq. The result is a cloud of data points in the form of an oblate spheroid with longer axis running parallel to the x axis, as can be seen in panel C of the same figure. Binning over a hexagonal grid was used to show the density of the cloud and emphasize its approximate symmetry over x=0 line. CLIP data points vary more dramatically than RNA-seq data. While the distribution of changes for the former exceeds the range between -5 and 5 log2 fold change, the RNA-seq changes are mostly contained between -2 and 2 log2 fold (slightly shifted towards positive values) causing the distribution to appear elongated. The parallelism of this cloud to the x axis is of importance, because it signifies that changes between x and y are uncoupled from one another. This is clearly visible, as the highest density of the cloud runs along the y=0 line, meaning that despite no changes in transcript levels, the CLIP reads still can vary considerably (reaching beyond the -5 and 5 range). Additionally, the symmetry of the cloud informs us that regardless of whether RNA-seq reads increase or decrease, within those two separate transcriptome level trends we have a largely equal number of data points which either increase or decrease in CLIP reads. Had it been otherwise and the two variables depended on each other, the cloud of points would have its major axis angled, showing either a negative or a positive correlation. Together, these data suggest that transcript level is not the major factor inducing changes in CLIP reads, with changes in microRNA levels bound to RISC or other factors influencing RISC affinity being more important.

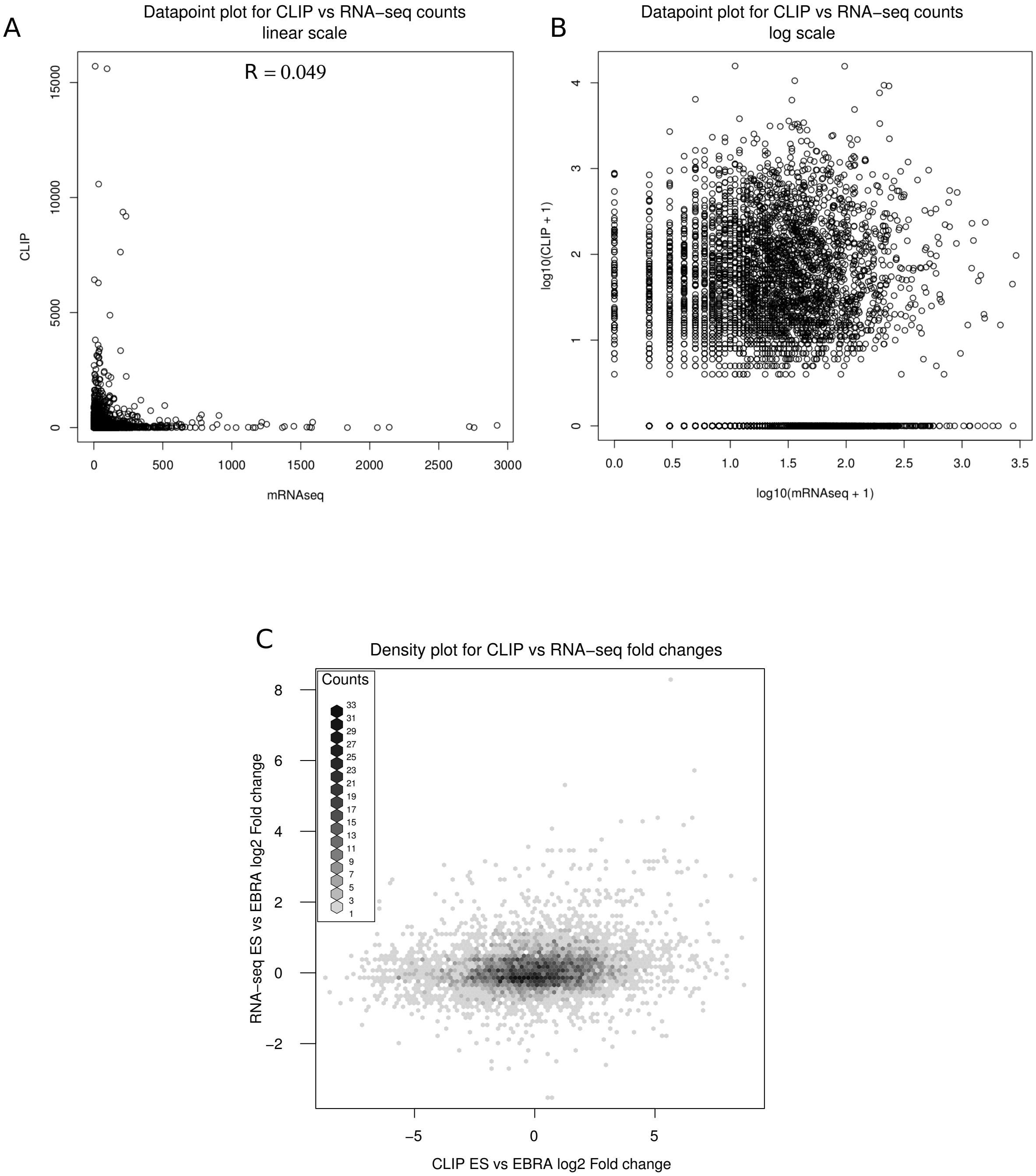


Figure 16. CLIP read counts are not simple derivative of transcript levels.
Subfigure A shows CLIP peak counts plotted in linear space against their corresponding transcript counts. Interestingly, medium and high CLIP reads are restricted to low abundant transcripts. This results in a very low correlation between variables. Even in the log-scale plot, as depicted in B, the dependence is comparable to those of two independent variables. Subfigure C shows the plot of fold changes corresponding to transition from ES to EBRA condition for both CLIP and RNA-seq reads.

4.12 Calling RISC Binding Sites Differing Between Conditions Using Negative Binomial Distribution Statistics and a Generalized Linear Model

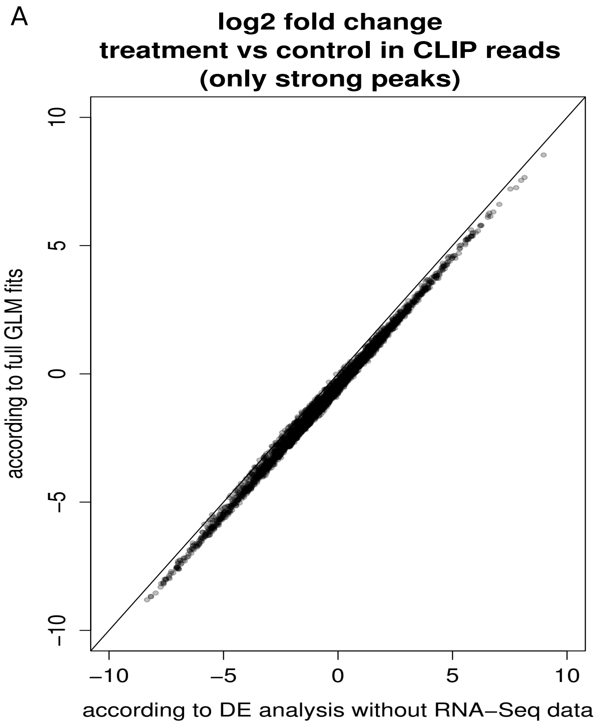
Having a table of counts per RISC binding site for each experimental condition, we face a problem of determining which sites are bound more or less. This problem is similar in essence to the one in microarray data analysis, where we need to distinguish true changes of the transcript levels from the noise introduced by imperfection of the method, as well as the biological noise. For microarrays, however, the value of transcript abundance is read out as intensity of fluorescence, while sequencing based methods produce sequences, which can be binned according to matching transcript sequence and counted. The process of counting introduces a different model of error associated with the measuring process compared to microarrays. It is described by a Poisson model and is characterized by variance that is not constant (like in the microarray intensity readout), but increasing, being equal to the mean. This means that for low counts, the standard deviation (square root of variance) is high in proportion to the count, while for high counts it is proportionally smaller (because count increases linearly, while standard deviation increases like a square root of the count). This causes the problem in considering a significant change for a value, because even a three-four fold change for the low count values is highly probable to have arisen by chance. The Poisson distribution would fit the data well if we created replicates by repeatedly sequencing the same population, performing technical repeats on the same biological sample. However, to be able to compare conditions, we need to have multiple biological replicates for them, thus introducing another source of noise which needs to be modeled. Therefore, if we plot variances derived from replicates of the experimental condition, they appear over-dispersed compared to the Poisson theoretical distribution. This problem has been realized just recently, with the advent of high-throughput sequencing methods and attempts at comparing biological samples. It has been successfully solved using a negative binomial distribution, which can be parameterized to behave like a Poissonian, but additionally takes as its input a variable which accounts for the biological noise. This variable can be estimated from replicates of the data, and the fitness of this variance estimation grows if more replicates are used. Additionally, we chose the conservative parameters for this model, which uses the maximal amount of dispersion available for a given data point, either a theoretical value from fitted standard deviation or the experimental value estimated from replicates available for this point. We have also noticed that increasing the number of replicates and comparing samples produced at the same time in a single experiment dramatically improves the number of called differentially bound peaks. Therefore, we feel that multiple replicates, preferably more than three, is an important new standard for statistically analyzed count data. This statistical method, initially developed for calling differentially expressed transcripts, has been applied by us to cloned RISC-bound transcriptome sections. To our knowledge, it is the first example of applying a statistical approach to quantitative comparison of transcript sections bound by a RISC complex.

Adding to the complexity of our problem is the fact that not only can the binding strength of RISC to a transcript section change, but also the transcript level may change between experimental conditions. Therefore, the number of reads per peak which is registered in the HITS-CLIP procedure is the combination of the microRNA level-induced changes and the transcript level changes. The method described above calls the peaks as being differentially bound if either of the two factors changes. To decouple those two we need another set of data, transcript levels for underlying RISC binding sites for all experimental conditions. Having those, we need to extend our statistical model to take into account data from both datasets. Most importantly, by introducing another source of data, we introduce additional noise into the system - both biological and statistical fluctuations. Most simply, in the case of continuous variables, a linear model is used in such a situation. It treats two variables as if they could be multiplied and presents them as one with larger normal distribution of errors. Since both our datasets are counts, we cannot use such an approach, as our errors are not normally distributed. There is, however, a solution which incorporates all necessary elements in the form of a generalized linear model. It combines two variables, just like the linear model, but between the resulting value and final value used in statistics, there is a link function which transforms one into another. Interestingly, for the Poisson distributions, the link function is simply log, while for the count data derived from biological replicates it is the negative binomial function. Since there was no available algorithm for such comparison, we reached out to a statistician, Dr. Simon Anders, who prepared for us the appropriate code to perform this computation. Finding differentially bound target sites using those two methods together provided us with an interesting and powerful combination of possibilities. We can either investigate a summary impact of transcriptome and RISC binding changes or the RISC binding affinity alone. Both approaches have their advantages, and while the former is based on fewer assumptions and thus safer, the latter provides the advantage of investigating more biologically consequential features of microRNA-directed RISC binding.

Correcting for transcript abundance induces a modest shift in many of the CLIP values, as can be judged in the fold-change plot in Figure 17, panel A. Most of the reads are shifted in one direction compared to the transcript-uncorrected data. What is important, however, is the fact that the transcript induced changes are small (maximally 2-fold) in comparison to the spectrum of changes exhibited by CLIP reads (which can go as high as 256 fold increase or decrease). With the introduction of the additional data source present in the generalized linear model, we were expecting the sensitivity of the method to drop, resulting in a smaller number of statistically significant hits. Luckily, this effect was not as pronounced and while this effect can be visible when comparing various conditions to EBRA, the converse is true for ES to LIF and ES to EB comparisons, which yield a higher number of hits in transcript adjusted data (Figure 17, panel B). Interestingly, hit lists produced by both methods overlap partially, particularly in the LIF to EBRA and ES to EBRA comparisons. The last one is notable, since the majority of the hits found in the transcript adjusted data are also found in non-adjusted (see panel C). Both methods seem to have their own sets of hits, which are specific to either of them. In further sections we will investigate if those hit lists complement each other functionally.

The hits produced with the use of either of the methods must have their initial p-value for the differential change between experimental conditions estimated and subsequently adjusted for multiple testing using the Benjamini-Hochberg method. Therefore, all hits listed are at 0.05 level of False Discovery Rate per comparison. This 0.05 value can be considered as strongly conservative, in light of certain statistical studies which advise considering values of up to 0.5 (Prof. Naomi Altman, personal communication).

As can be expected, with increasing progress in differentiation, the number of differentially bound peaks increases as well. This can be appreciated in Figure 18. depicting the logarithm of fold changes of peak counts plotted against the logarithm of average counts in two conditions. This type of plot makes it easy to spot differentially bound data points by situating them close to the peripheries of the distribution; however, the ultimate decision regarding whether a peak is differentially bound or not relies also on variance present in data point. This is the reason why we can see extensive presence of non-significantly changing data points among the hits. It can be appreciated that our method calls reasonable data points in all of the comparisons and in particular, hits in the highly bound target sites are also present. The ES to EBRA comparison confronts the two most distantly separated conditions and is the richest in differentially bound RISC binding sites. What makes us cautious is the fact that the



B

	Analysis without RNAseq data	Transcript adjusted
ES_EBRA	1359	465
LIF_EBRA	479	269
EB_EBRA	190	100
ES_LIF	23	46
ES_EB	128	270

C

	Without RNAseq only	Common	With RNAseq only
ES_EBRA	991	368	97
LIF_EBRA	406	73	196
EB_EBRA	173	17	83
ES_LIF	22	1	45
ES_EB	92	36	234

Figure 17. Effects of transcriptome adjustment with generalized linear model on HITS-CLIP peaks
Subfigure A shows a target site fold change plot using CLIP data (DESeq, X axis) and with the transcript level correction (generalized linear model, Y axis). Subfigure B shows numbers of differentially bound target sites called for either transcript adjusted or unadjusted data. Interestingly, while some of the conditions show significant overlap between differentially bound target sites called using either method (ES_EBRA and LIF_EBRA), others are more separate.

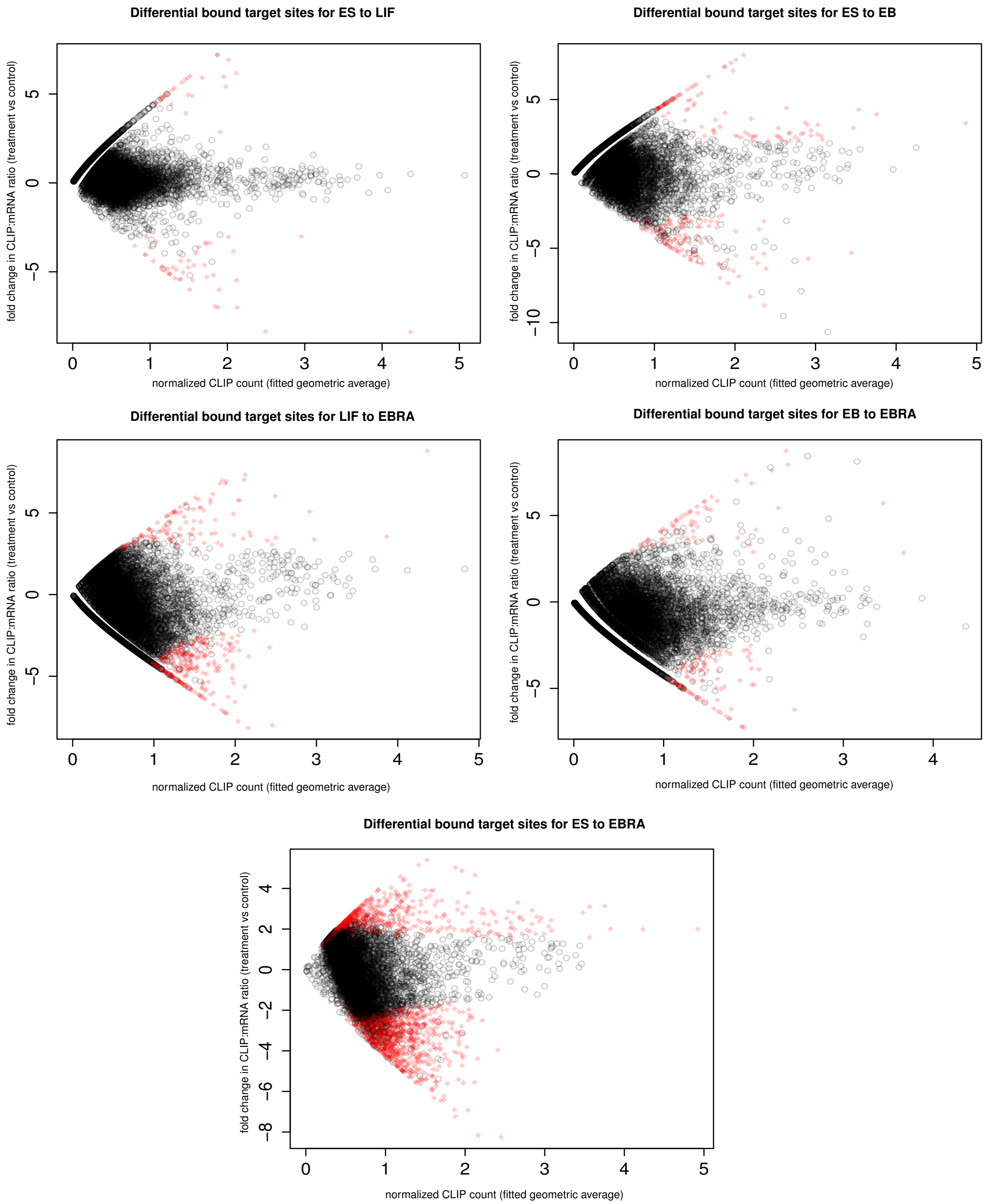


Figure 18. Differential binding plots for uncorrected CLIP peaks (according to DESeq)
Five plots show differential binding, as detected by DESeq package (no transcript level correction). The significance threshold for detection is less than 0.05 False Discovery Rate.

fraction of differentially bound peaks goes deep into the main distribution of data points, without a clear separation that can be hypothesized in some of the lightly bound comparisons.

Differentially bound target sites called using the Generalized Linear Model method appear to fall in the same regions of the plots compared to DESeq, which is anticipated and speaks about proper behavior of this model (Figure 19). One notable feature is slight asymmetry present in the LIF to EBRA and EB to EBRA comparisons, appearing as a shift in the y axis in either direction. This resulted in asymmetric calling of differential binding for data points, with one of the directions being poor in differentially bound reads. We are unsure about the cause of this problem, but it may have something to do with the lower abundance of counts in the LIF and EB datasets which may have consequences on normalization or convergence of GLM algorithm.

Altogether using both statistical methods we are able to successfully call differentially changing RISC target sites at a conservative False Discovery Ratio. As it can be appreciated, either method results in numerous hits passing statistical significance level, therefore making possible functional studies using gene set enrichment criterion.

4.13 Functional Analysis of Gene Sets Containing Differentially Bound Target Sites

Our statistical models produce lists of target sites bound differentially between two conditions. Genes in which those target sites are located are likely to have something to do with differentiation and can give useful clues about cellular mechanisms important for this process. Those genes are likely to act together forming groups of interacting proteins in the form of pathways and regulatory circuits. The rationale justifying this approach is similar to the one for gene co-expression studies and is based on the assumption that although genes behaving as master regulators do exist, they still have downstream effectors forming functional group, while even more processes are controlled by a more subtle cooperation of several factors forming a network where responsibility for function is distributed between its members. Appearance of such groups in the data is easy to spot, and more importantly, is amenable to statistical analysis to assess how non-random the enrichment of a certain group of genes is.

We have used methodology which uses the Gene Ontology Biological Process dataset to identify cellular functions with enriched gene groups. [5] Since this dataset has a hierarchical structure, our first attempts at generating lists of cellular functions were troubled by numerous entries containing redundant information. In order to organize those results better, we have used ClueGO plugin for Cytoscape, which

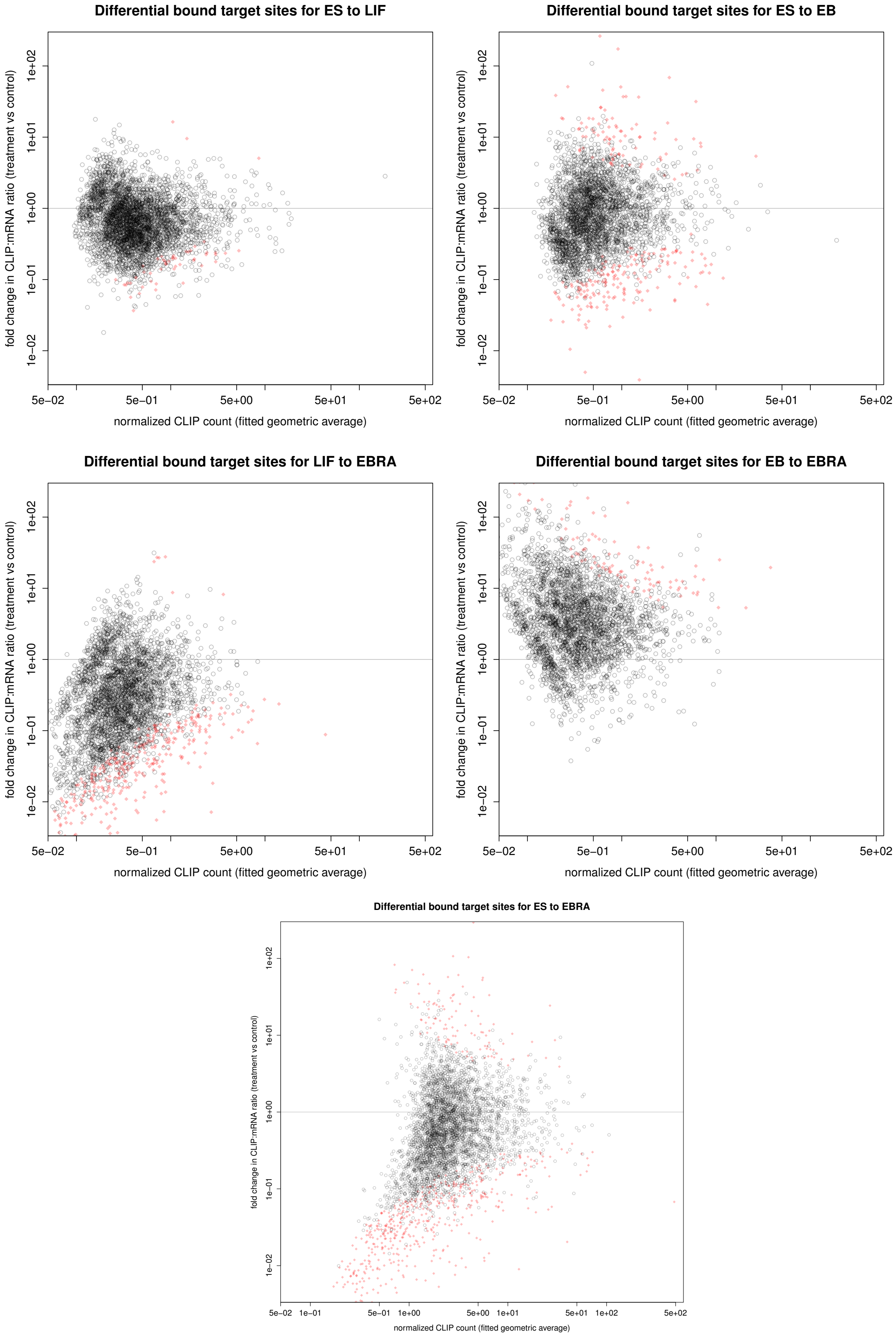


Figure 19. Differential binding plots for transcript corrected CLIP peaks (according to GLM)
Five plots show fold changes for CLIP peaks and mark differentially bound sites, as detected by generalized linear model using transcriptome corected data. The significance threshold is set at less than 0.05 False Discovery Rate.

creates a network with GO terms represented as nodes and the number of shared genes is used to connect them with edges. [9] It is possible to manipulate the connectivity of the network by adjusting the kappa score in order to arrive at networks grouping functionally related sets of genes.

We were successful in finding multiple significant terms and gene groups in our comparisons. We have decided to focus on the problem of differentiation into a neuro-like lineage for several reasons. First of all, only the EBRA condition is enriched in several specific microRNA families and as we have shown before, some of them enact measurable impact on target sites. Secondly, gene ontology terms associated with neuro-like processes are easily identifiable, while terms associated with ES cell pluripotency are more generic. Additionally, the EBRA condition looks as the most distant and the most specific compared to three others, as judged by high number of differentially bound target sites.

Figure 20A presents a general overview of the results, with percentage breakdowns for highlighted functional groups in relation to rest of the GO terms found in the data. We wanted to highlight the fact that several functional groups appear in all of our comparisons to the EBRA condition, most notably a group of neuro-specific terms and the TGF-beta/BMP signalling pathway. While the ES to EBRA comparison is the richest in number of differentially bound genes, both LIF and EB used as a starting point for comparison to the EBRA condition may also result in identifying genes specific for neuronal differentiation.

Panel B of the figure shows the sub-network corresponding to the highlighted neuro-specific set of genes in the LIF to EBRA comparison from panel A. The inset on this figure shows the same sub-network color-coded according to the number of genes per node being either repressed or de-repressed in the EBRA condition, and demonstrates that the majority of nodes contain genes of both types. The connectivity within this network suggests that gene groups associated with GO term nodes share common members, and this fact can be appreciated in full by looking at Figure 21. depicting the matrix of GO terms and their associated genes. The 65 genes that represent nearly 10% of differentially bound genes fall into 23 neuro-related GO terms, and 18 of them are members of the connected sub-network described above. The following GO terms were most enriched in our differentially bound data set: neuron development (43 genes), neuron projection development (41), neuron projection morphogenesis (30), cell morphogenesis involved in neuron differentiation (29), and regulation of neuron differentiation and axonogenesis (24). It is important to note that most of them allude not to the functions of the neurons themselves, but rather to neuronal development and specification, in agreement with the fact

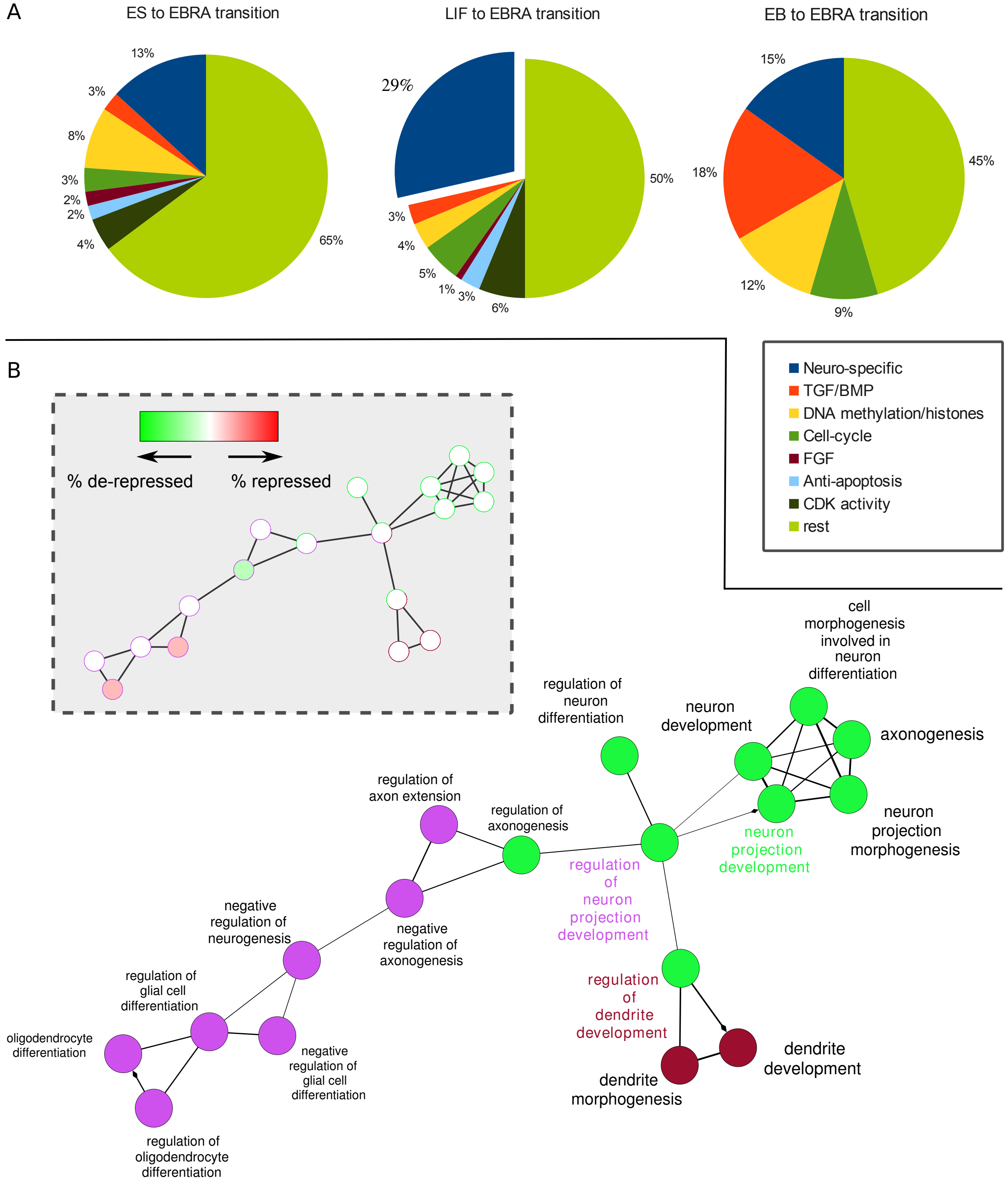


Figure 20. Pathways and gene groups in differentiation to neuro-specific cells
Subfigure A shows percentages for recurring statistically significant groups in all transitions leading to EBRA. Panel B shows a selected sub-network corresponding to the neuro-specific terms in LIF to EBRA comparison. The main figure shows annotation of the nodes, as well as color-coded divisions into main groups based on the degree of overlap between node gene sets. The inset shows how many repressed or de-repressed genes are in the particular node.

	axonogenesis	regulation of neuron projection development	positive regulation of neuron projection development	primary neural tube formation	dendrite development	regulation of axon extension	axon regeneration	neuron projection development	regulation of neuron differentiation	negative regulation of neuron differentiation	positive regulation of neuron differentiation	regulation of glial cell differentiation	negative regulation of glial cell differentiation	neuron development	cell morphogenesis involved in neuron differentiation	oligodendrocyte differentiation	regulation of oligodendrocyte differentiation	neuron projection morphogenesis	dendrite morphogenesis	negative regulation of neurogenesis	regulation of axonogenesis	negative regulation of axonogenesis	regulation of dendrite development
Abi2					■			■						■									
Abl2		■	■					■	■					■									
Actb	■							■						■	■			■					
Adnp	■	■	■			■		■	■					■	■			■			■		
Apoe	■						■	■						■	■			■					
Atf1		■	■					■	■					■									
Bnip2									■		■												
Btg2								■						■									
Calr									■	■													
Caprin1		■			■			■	■					■	■			■	■				■
Ccnd2									■														
Cdkn1c														■									
Cntnap2								■						■									
Ctnna1	■						■	■						■	■			■		■			
Ctnnb1												■	■			■	■			■			
Cxcr4	■							■				■		■	■	■	■	■					
Cyfip1	■							■						■	■			■					
Dlc1				■																			
Dnmt3b									■		■												
Dst	■							■						■	■			■					
Ect2									■		■												
Enah	■			■				■						■	■			■					
Exoc4																■							
Fkbp4		■						■	■					■									
Gas1	■							■						■	■			■					
Gata3	■							■						■	■			■					
Gja1								■						■				■					
Hectd1				■																			
Hes1									■	■		■	■	■	■	■	■			■			
Igf1r	■							■						■	■			■					
Itgb1		■	■					■	■	■	■			■									
Jun	■						■	■						■	■			■					
Lifr								■						■				■					
Lin28a									■		■	■	■							■			
Mtap2	■				■			■						■	■			■	■				
Mycn												■	■							■			
Nckap1				■																			
Nedd4		■			■			■	■					■	■			■	■				■
Nefl	■	■					■	■	■					■	■			■			■		
Nodal				■																			
Ntn1	■	■				■		■	■					■	■			■		■	■	■	
Olig2									■	■		■				■	■						
Picalm	■				■			■						■	■			■	■				
Ppp1r9a								■						■									
Prickle2								■						■									
Pten	■	■			■			■	■					■	■	■		■	■	■	■	■	■
Ptprf		■			■			■	■					■	■			■	■				■
Rap2a		■			■			■	■					■	■			■	■				■
Rest									■	■										■			
Rnf6	■	■				■		■	■					■	■			■		■	■	■	
Rtn4	■	■				■	■	■	■	■		■		■	■	■		■		■	■	■	
Sall1				■					■		■												
Sema4c				■																			
Sept2								■						■									
Sipa1l1	■	■			■			■	■					■	■			■	■		■		■
Stk24	■	■					■	■	■					■	■			■			■		
Tcf12									■		■												
Tgif1				■					■		■												
Top2b	■							■						■	■			■					
Trim71				■																			
Uchl1	■							■						■	■			■					
Usp33	■							■						■	■			■					
Vldlr		■	■		■			■	■					■									■
	24	17	5	9	10	4	6	41	29	6	8	7	4	43	29	7	4	30	8	10	8	4	7

Figure 21. Matrix of GO terms for neuro-related genes found in LIF to EBRA comparison

Those 63 out of 675 differentially bound genes detected in LIF to EBRA transition are found to be neuro-specific according to GO terms associated with them. Majority of those terms assemble into a connected network on the virtue of number of genes shared between them.

that we are sampling cells on the path of neuronal differentiation and not adult brain tissue.

A full view of the network associated with LIF to EBRA differentiation is shown in Figure 22. It can be appreciated that it contains many other functional groups, like alternative nuclear mRNA splicing via the spliceosome, protein ubiquitination, and mitotic cell cycle checkpoint, which is functionally related to negative regulation of cyclin-dependent protein kinase activity and negative regulation of protein phosphorylation. There are also 'histone lysine methylation' and 'gene silencing by miRNA' groups, showing that epigenetic regulation is also affected during differentiation. Of note are several GO terms disjunct from main neuro sub-network with annotation which can support its role, like primary neural tube formation, positive regulation of neuron differentiation, negative regulation of neuron differentiation, axon regeneration, and positive regulation of neuron projection. Such separation from the main neuro sub-network is caused by having a set of genes which is not shared with those from main sub-network. Also notable is presence of genes involved with membrane depolarization - a hallmark of neuronal cells. It is important to note that attempts of network construction based on Gene Ontology terms derived from gene sets differentially expressed found in RNA-seq data were unsuccessful. While we were able to obtain a number of statistically significant GO terms, they did not form connected networks and what is more important, they did not show enrichment in neuro-related terms.

Two insets on Figure 22 show color-coded versions of this network. The upper inset indicates whether the specific node consists of genes that are mostly depleted or enriched for RISC complex binding following differentiation. The majority of the nodes have mixed constitution, disfavoring a naive model of differentiation causing microRNAs to either repress or to cease repressing every gene associated with a functional group, except for few rare nodes. This can be appreciated in detail on Figure 23 which shows a summary breakdown of the network with mixed nodes visible in panel A, ones bound by RISC in response to differentiation in panel B, and ones released by RISC in panel C. What is remarkable in mixed nodes is equipartition of their genes between increased and decreased binding of RISC. Most of the nodes are between 4 and 10 % of respective gene group associated with GO term; however, neuro-specific terms are in fact supported by a large number of genes, as compared to the rest. Interestingly, several neuro-related terms are exclusively present in panels B and C. While this opens the possibility for a straightforward interpretation of the results, our hopes are quenched pre-maturely if we compare repressed GO term positive regulation of neuron projection development with de-repressed negative regulation of axonogenesis. While the former implies inhibition in formation of projections, the

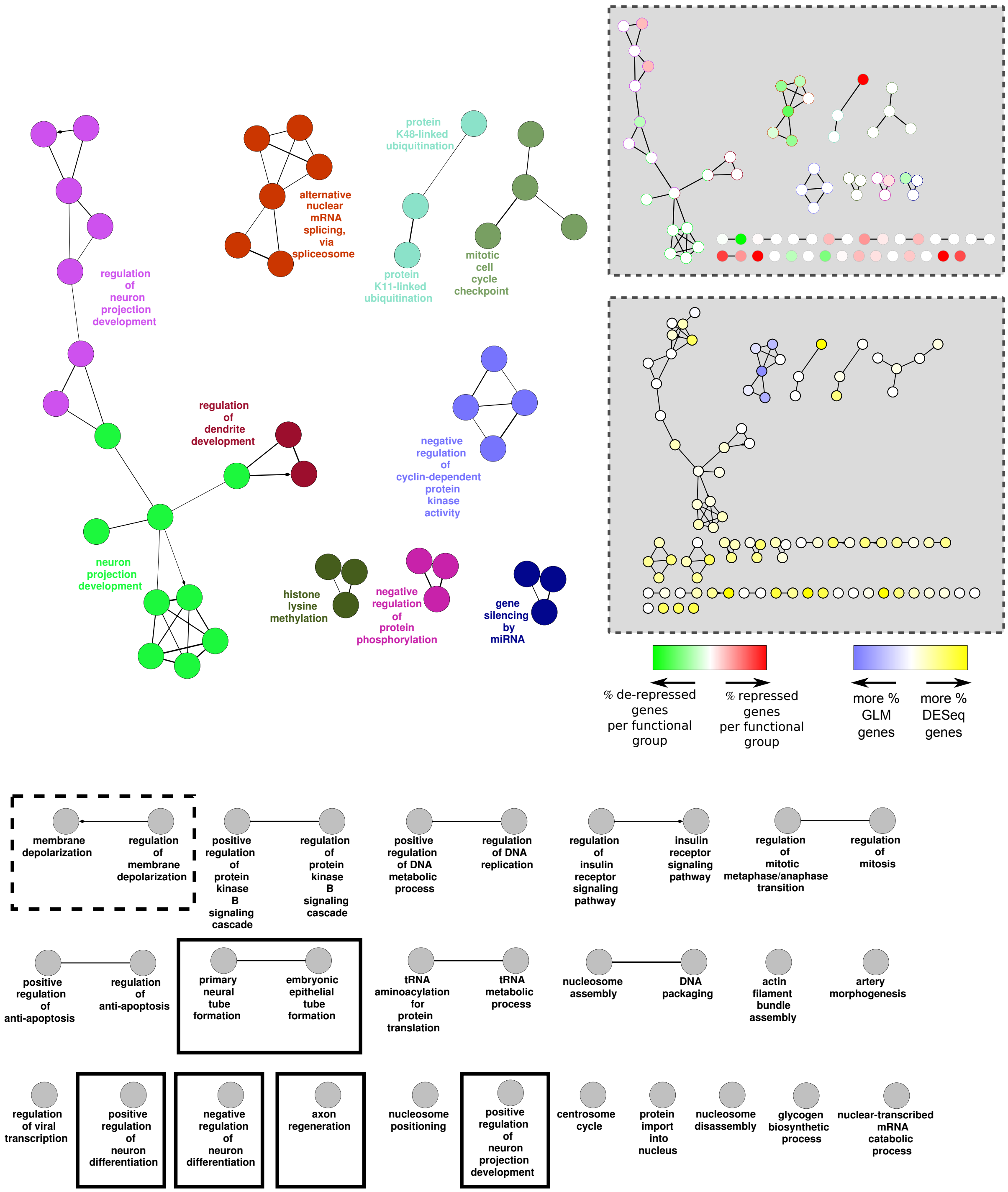

regulation of neuron projection development
alternative nuclear mRNA splicing, via spliceosome
protein K48-linked ubiquitination
protein K11-linked ubiquitination
mitotic cell cycle checkpoint
regulation of dendrite development
negative regulation of cyclin-dependent protein kinase activity
neuron projection development
histone lysine methylation
negative regulation of protein phosphorylation
gene silencing by miRNA
% de-repressed genes per functional group
% repressed genes per functional group
more % GLM genes
more % DESeq genes
membrane depolarization
regulation of membrane depolarization
positive regulation of protein kinase B signaling cascade
regulation of protein kinase B signaling cascade
positive regulation of DNA metabolic process
regulation of DNA replication
regulation of insulin receptor signaling pathway
insulin receptor signaling pathway
regulation of mitotic metaphase/anaphase transition
regulation of mitosis
positive regulation of anti-apoptosis
regulation of anti-apoptosis
primary neural tube formation
embryonic epithelial tube formation
tRNA aminoacylation for protein translation
tRNA metabolic process
nucleosome assembly
DNA packaging
actin filament bundle assembly
artery morphogenesis
regulation of viral transcription
positive regulation of neuron differentiation
negative regulation of neuron differentiation
axon regeneration
nucleosome positioning
positive regulation of neuron projection development
centrosome cycle
protein import into nucleus
nucleosome disassembly
glycogen biosynthetic process
nuclear-transcribed mRNA catabolic process

Figure 22. GO term network constructed from genes bound differentially by RISC in LIF to EBRA transition. Figure shows full GO term network with annotated functional groups in color and single/double nodes shown in grey. Neuro-related terms are highligted by encircling with black line, while dotted line shows potentially involved terms. Top inset shows minature network color-coded according to the percentage of repressed or de-repressed genes found in the node. Bottom inset shows a related network color-coded according to percentage of DESeq or GLM genes contributed to the node.

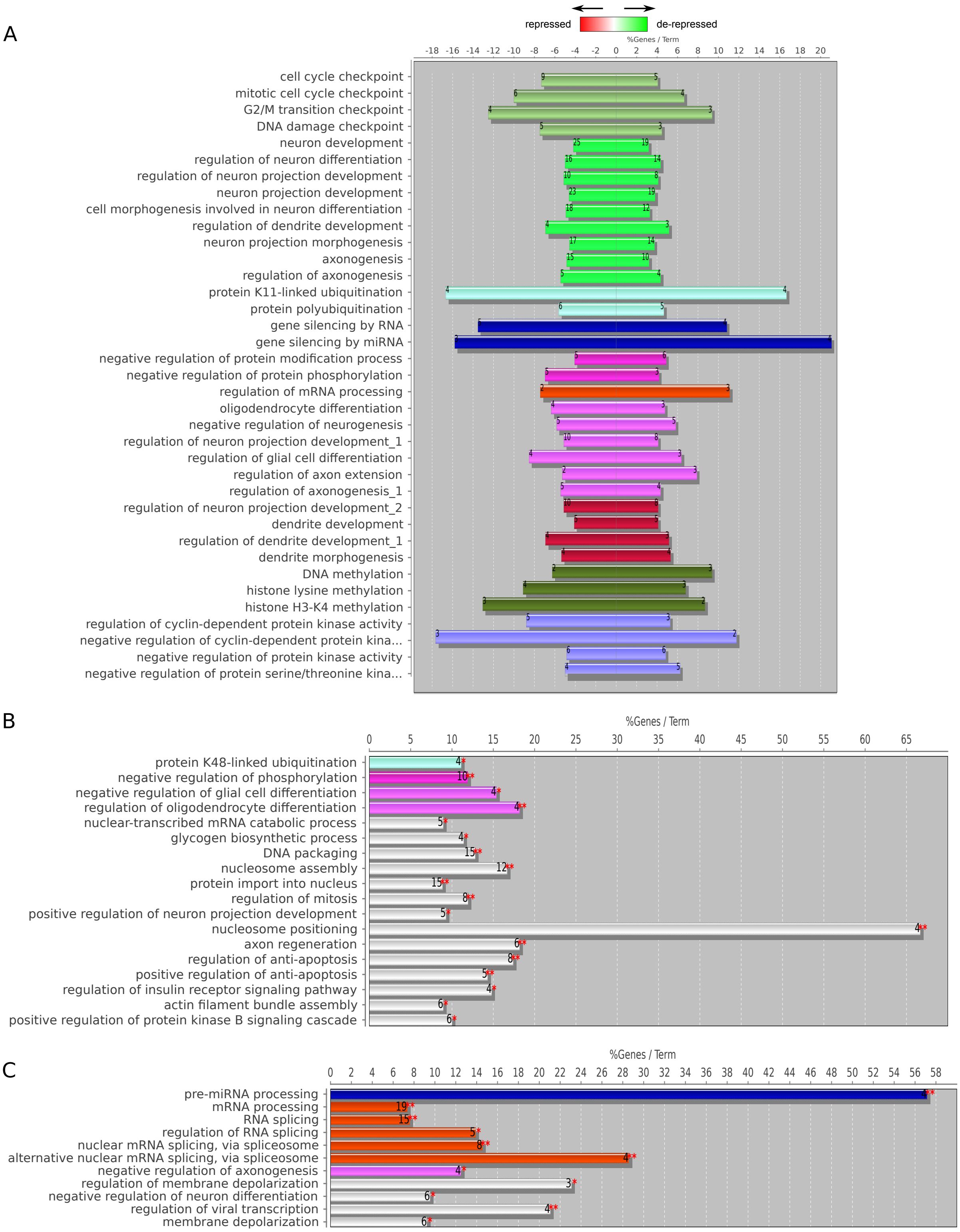


Figure 23. Detailed statistics of genes repressed or de-repressed in GO term network.

Subfigure A shows the summary for GO term network for LIF to EBRA transition divided according to whether genes are up or down-regulated in detected significant GO terms and groups. Subfigure B shows specifically repressed terms while C shows de-repressed terms.

latter suggests the stimulation in formation of axons, which themselves are a form of projections. With only these data available, we can conclude only that these kinds of processes are regulated by microRNAs; however, we cannot guess the outcome of this regulation on differentiation. The problem is even more complicated due to the presence of additional mixed terms like neuron projection morphogenesis, and it appears that without a more detailed and preferably mathematical/numerical model involving all the genes affecting a given process, we will be unable to predict the outcome of miRNA-mediated regulation solely with these analyses. It seems that such interpretations are only possible in rare cases, when the entire group of related GO terms falls under either regulation. In our case this could be possibly attempted in case of the alternative splicing regulation group and membrane depolarization terms, which are both released from microRNA inhibition upon differentiation. Even in this case caution needs to be observed, since many genes are pleiotropic and therefore their function as defined by GO terms is valid only for a specific functional context. Some of the genes identified through Gene Ontology may have functions different than those implied by analysis due to the different functional context than the one present at the time of experiments leading to defining function for Gene Ontology. Therefore GO term analysis can only be used to roughly define gene set potentialy involved in a particular function. As we can also see, rarely it is a specific function and more often a range of functions relevant to the group. This vagueness seems as a severe handicap of the method, however it needs to be noted that it is still superior as a starting point analysis and can be extended by literature search to form a base for further experiments aimed at deciphering cooperation of genes in the context of suggested specific function.

The lower inset on Figure 22 shows the constitution of the nodes according to whether genes come from the hit list identified through a transcript-adjusted method (GLM) or the unadjusted one (DESeq). It turns out that genes from both sources contribute to the formation of the network. Although transcript unadjusted data provides arguably more target sites, if we attempt to create a network using either data source alone, the resulting network does not appear to have as many connections between functionally relevant terms. This signifies the fact that each method contributes its own set of unique genes, which fall into the same functional terms, enriching them to the extent that they contain genes common with other related terms.

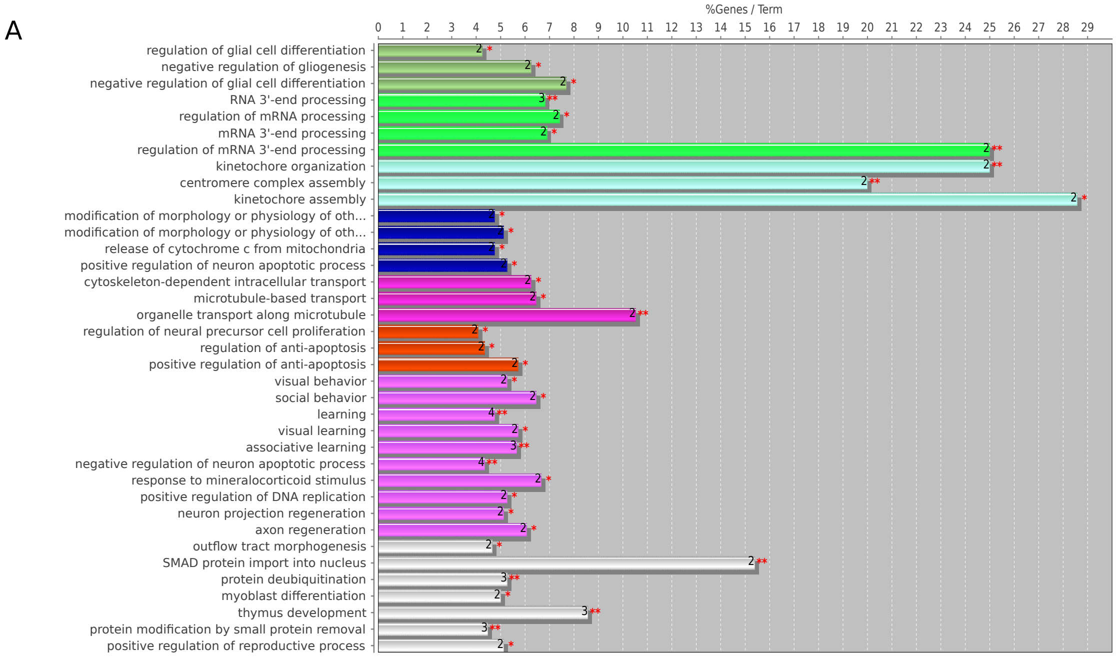
4.14 Differentially Changing Target Sites Containing Mir-19 and mir-367 Complementary Sequences Are Located in Genes Related to Neuronal Differentiation

Encouraged by highly pronounced enrichment of neuro-related terms in our comparisons, we investigated whether EBRA-enriched microRNAs have a causal role in this observation. We concluded previously that target sites containing complementary sites for miR-19 and miR-367 microRNA families show significant up-regulation in the EBRA condition as compared to others. While the resulting enrichment for neuro terms in EBRA is likely to have arisen through the co-operation of multiple microRNAs, it appears that even single microRNAs are capable of targeting genes to a sufficient extent necessary for the detection of significant GO terms.

Out of 284 total complementary target sites for miR-367, 113 are differentially bound, and they map to 89 genes. Of these, only 18 are associated with significant GO terms, which can be appreciated in detail in Figure 24, panel A and B. Eight out of those genes are associated with terms related to neuronal development, and surprisingly, Btg2, Htt, Jun and Kras are associated with higher cognitive functions, like learning or social behavior. However those functions should not be interpreted literally, rather the presence of those genes in the differentially bound set is simply the consequence of them having pleiotropic involvement in many neuronal functions depending on context.

Mir-19 has corresponding 477 target sites, out of which 163 are differentially bound within 138 genes. 26 of them are found to be associated in our analysis with functional terms (see Figure 25, panel A and B). Ultimately, only six genes are associated with neuro-related terms, aggregating into two broad functions of neural tubule formation and nerve development. Neural tube formation is an especially interesting category, since it is a part of the early developmental program necessary for the appearance of the central nervous system. Additionally, the association of this gene group with some signalling pathways is apparent, including with insulin-like growth factor receptor signaling, FGF, Notch, MAPK.

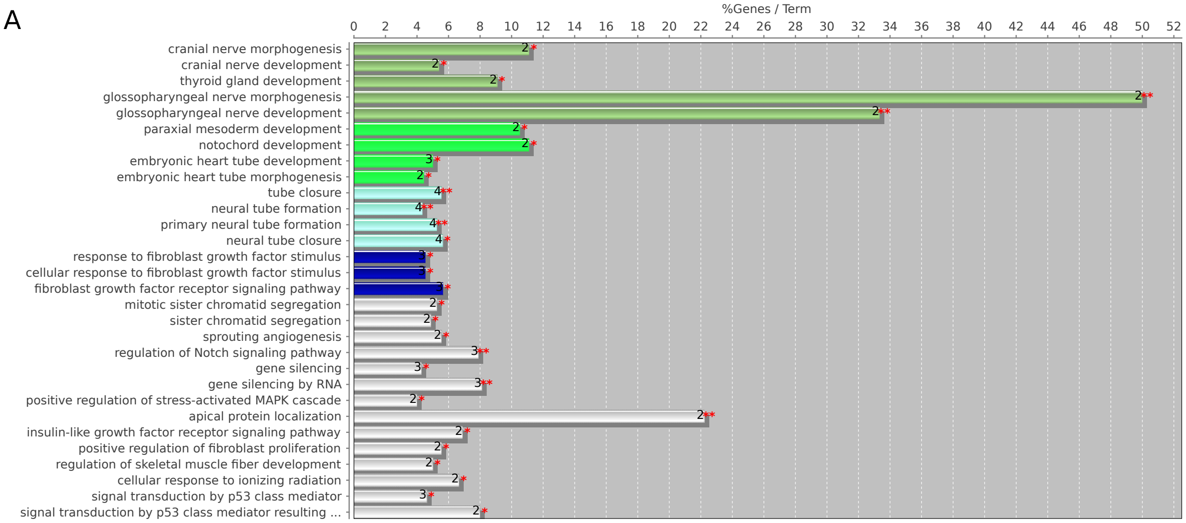
While all of the functional terms detected are statistically significant at False Discovery Rate of 0.05, the low numbers of genes associated with them are somewhat disappointing. This is a dramatically different situation from that shown in Figure 23, panel A, where neuro-related terms have significantly higher numbers of genes associated with them. This may signify that cooperation between several microRNAs, may be necessary to attain such a high number of sites, as well as participation of target sites with unidentified microRNAs controlling them, which were ommited from those analyses.



B

	Ahcyl1	Bcl2l11	Bicd2	Btg2	Ctnnb1	Hoxa3	Htt	Jun	Kras	Lin28a	Mis12	Nefl	Otud7b	Rnf4	Tob1	Usp28	Usp33	Zfhx3
release of cytochrome c from mitochondria		■						■										
outflow tract morphogenesis					■			■										
SMAD protein import into nucleus								■							■			
learning				■			■	■	■									
visual behavior							■		■									
associative learning				■			■		■									
visual learning							■		■									
microtubule-based transport			■				■											
negative regulation of gliogenesis					■					■								
protein deubiquitination													■			■	■	
cytoskeleton-dependent intracellular transport			■				■											
neuron projection regeneration								■				■						
axon regeneration								■				■						
RNA 3'-end processing	■			■						■								
mRNA 3'-end processing	■			■														
regulation of mRNA 3'-end processing	■			■														
centromere complex assembly											■			■				
social behavior							■		■									
modification of morphology or physiology of other organism		■						■										
negative regulation of neuron apoptotic process							■	■	■			■						
positive regulation of neuron apoptotic process		■						■										
myoblast differentiation					■													■
regulation of glial cell differentiation					■					■								
negative regulation of glial cell differentiation					■					■								
positive regulation of DNA replication								■	■									
regulation of anti-apoptosis				■	■													
positive regulation of anti-apoptosis				■	■													
thymus development		■			■	■												
regulation of mRNA processing	■			■														
kinetochore assembly											■			■				
kinetochore organization											■			■				
response to mineralocorticoid stimulus									■			■						
modification of morphology or physiology of other organism involved		■						■										
protein modification by small protein removal													■			■	■	
organelle transport along microtubule			■				■											
regulation of neural precursor cell proliferation				■	■													
positive regulation of reproductive process					■			■										

Figure 24. Gene Onthology terms significantly enriched for mir-367-controlled genes
Subfigure A shows a summary statistics for GO terms, while B shows genes responsible for GO terms, with neuro-specific terms highlighted in yellow.



B

	mitotic sister chromatid segregation	sister chromatid segregation	neural tube formation	neural tube closure	sprouting angiogenesis	embryonic heart tube morphogenesis	fibroblast growth factor receptor signaling pathway	regulation of Notch signaling pathway	primary neural tube formation	gene silencing	cranial nerve development	glossopharyngeal nerve development	cranial nerve morphogenesis	glossopharyngeal nerve morphogenesis	thyroid gland development	notochord development	gene silencing by RNA	positive regulation of stress-activated MAPK cascade	embryonic heart tube development	cellular response to fibroblast growth factor stimulus	apical protein localization	insulin-like growth factor receptor signaling pathway	positive regulation of fibroblast proliferation	paraxial mesoderm development	regulation of skeletal muscle fiber development	tube closure	cellular response to ionizing radiation	response to fibroblast growth factor stimulus	signal transduction by p53 class mediator	signal transduction by p53 class mediator resulting in induction of apoptosis
Ccdc99	■	■																												
Clock																											■			
Ctgf							■											■		■								■		
Dll1								■																						
Ephb4					■																									
Fosl2																							■							
Gigyf1																						■								
Gja1						■													■										■	
Hipk1																													■	■
Hoxa3											■	■	■	■	■															
Hoxb3											■	■	■	■	■															
Igfbp3																						■			■					
Il6st								■																						
Ino80	■	■																									■			
Nckap1			■	■					■							■			■		■			■		■				
Nptn							■													■								■		
Nrarp					■			■																						
Rab14							■													■	■							■		
Sall4			■	■					■																	■				
Sema4c			■	■					■									■								■				
Sik1																									■				■	■
Tead1						■										■			■					■						
Tgif1			■	■					■														■			■				
Tnrc6a										■							■													
Tnrc6b										■							■													
Tnrc6c										■							■													

Figure 25. Gene Onthology terms significantly enriched for mir-19 and corresponding genes
Subfigure A shows a summary statistics for GO terms, while B shows genes responsible for GO terms, with neuro-specific terms highlighted in yellow.

4.15 The Insulin Receptor Signaling Pathway as an Example of a Heavily MicroRNA Regulated Protein Interaction Network

Figure 26, panel A shows some of the signalling pathways for which genes can be found in our comparisons, together with corresponding significance figures for GO terms and groups (groups of terms identified as having related function with their genes together tested for likelihood of appearing together and not by random chance). It can be appreciated that there are several signalling pathways present, known to be involved in neuronal fate specification, namely TGF-beta, the insulin receptor pathway, nodal and VEGF pathways, as well as the Retinoic Acid receptor pathway. [26, 63, 28, 59] The best positive confirmation that our method identifies processes specific to neuronal differentiation is presence of the last pathway. Since we are adding Retinoic Acid to the medium in order to promote differentiation into the neuronal lineage, it is logical to see a regulation of this process.

While the pathways listed above rarely formed groups, and therefore were somewhat less visible, the knowledge that even the perturbation of a single gene within the pathway may have consequences on its signaling output encouraged us to conduct a more detailed investigation. While not a fore-runner in the number of enriched genes, the insulin receptor signalling pathway appeared as the one having its major players under the influence of microRNA regulation. Figure 26, panel B shows a simplified diagram of this network. Pathway members shown in yellow boxes were detected to have RISC binding sites incidentally located exclusively in the 3'UTR regions of their transcripts. Detailed diagrams depicting RISC-binding levels for those target sites in all five replicates per condition are shown in Figures 27 and 28.

Levels of RISC binding to at least a single microRNA target site in the transcript change dramatically in majority of those genes, primarily increasing in the EB or EBRA condition. This is probably best seen on the example of Igfbp3, where a single peak jumps from near zero level in ES and LIF to more than 7000 reads in EB in a very consistent way. Contrary to the rest, Pten is heavily bound in ES and LIF conditions, while transcript of its downstream effector Akt seems to be more uniformly bound. Panel B of Figure 26 shows increased binding of RISC to transcripts in EBRA condition (arrows pointing down), which can be deduced from Figures 27 and 28. In principle, it appears likely that the pathway is being shut down, since most of its components have transcripts bound more strongly by RISC, including its effectors: p21 (Cdkn1a) and p27 (Cdkn1b), as well as receptors (Igf1r, Igf2r) and ligand (Igf2).

A

signaling pathway	Group p-val	Term p-val	Implicated in neuronal fate	Implicated in maintenance of pluripotency	Supported by multiple comparisons
TGF-beta	9.40E-003	2.70E-002	Y		Y
FGF	6.50E-006	6.30E-005		Y	Y
insulin receptor	-	9.30E-003	Y		Y
activin receptor	-	2.30E-002		Y	
nodal	-	1.20E-002	Y	Y	
retinoic acid	-	3.60E-003	Y		
VEGF	-	3.30E-003	Y		

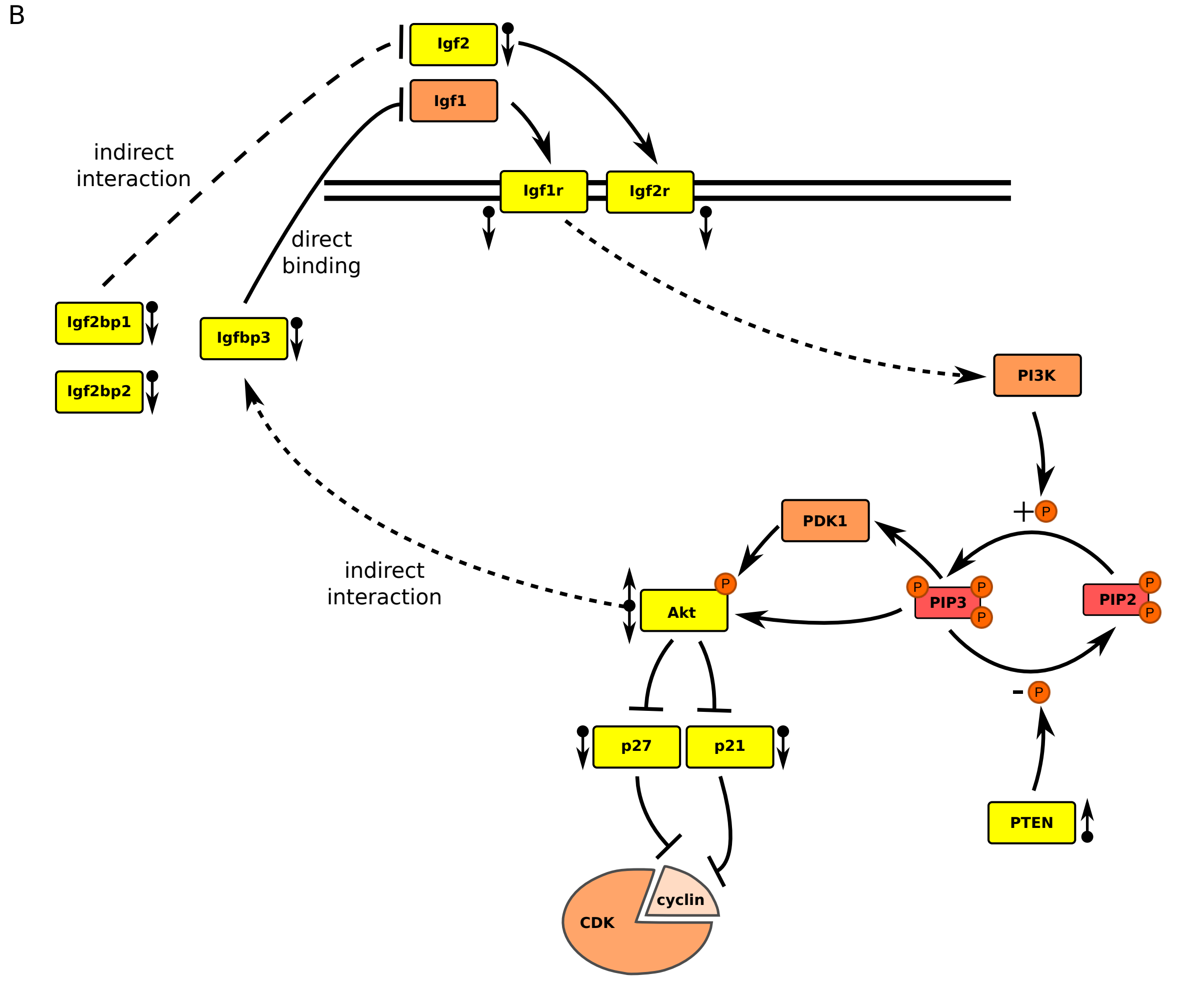


Figure 26. Signaling pathways controlled by RISC binding sites and the extent of microRNA -mediated regulation of Insulin Receptor Signaling Pathway

Subfigure A shows table of signaling pathways which can be detected with the help of GO search. Also listed is possible involvement in our differentiation model, as supported by the literature. P-values listed are multiple testing adjusted. Panel B shows relations in the Insulin Receptor Signaling Pathway. Proteins marked in orange were not detected to have RISC binding sites. Arrows with a dot next to the protein involved show inhibition or de-repression by microRNA.

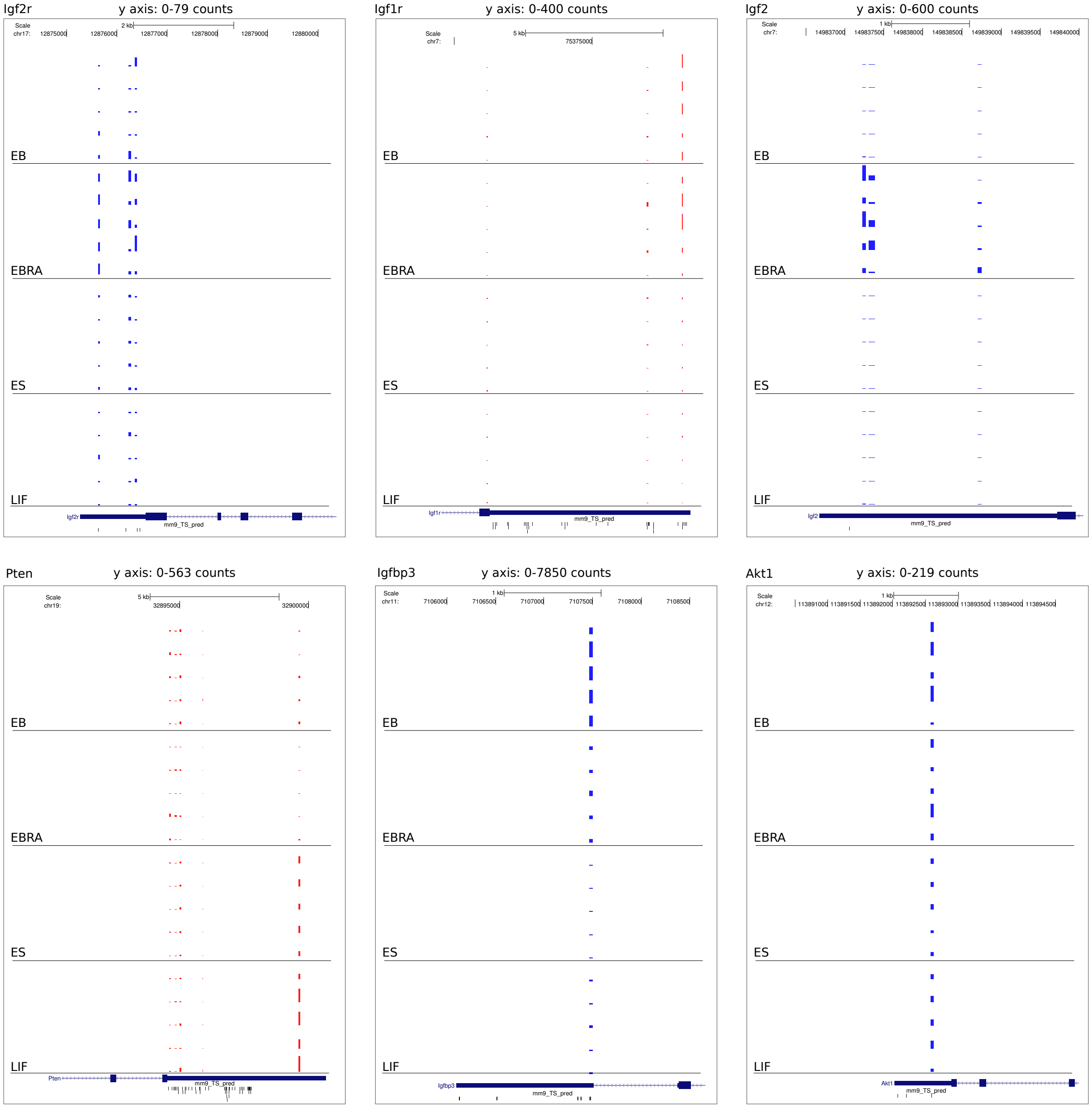


Figure 27. Genes of Insulin Growth Factor Signaling Pathway with their RISC binding sites.
Figure shows overview of the RISC binding portions of the genes, along with the TargetScan prediction track. RISC binding sites show bars proportional to the numbers of reads registered with text above each image identifies maximum value for the highest bar in all replicates. CLIP peaks from Watson strand appear in red, from Crick strand in blue.

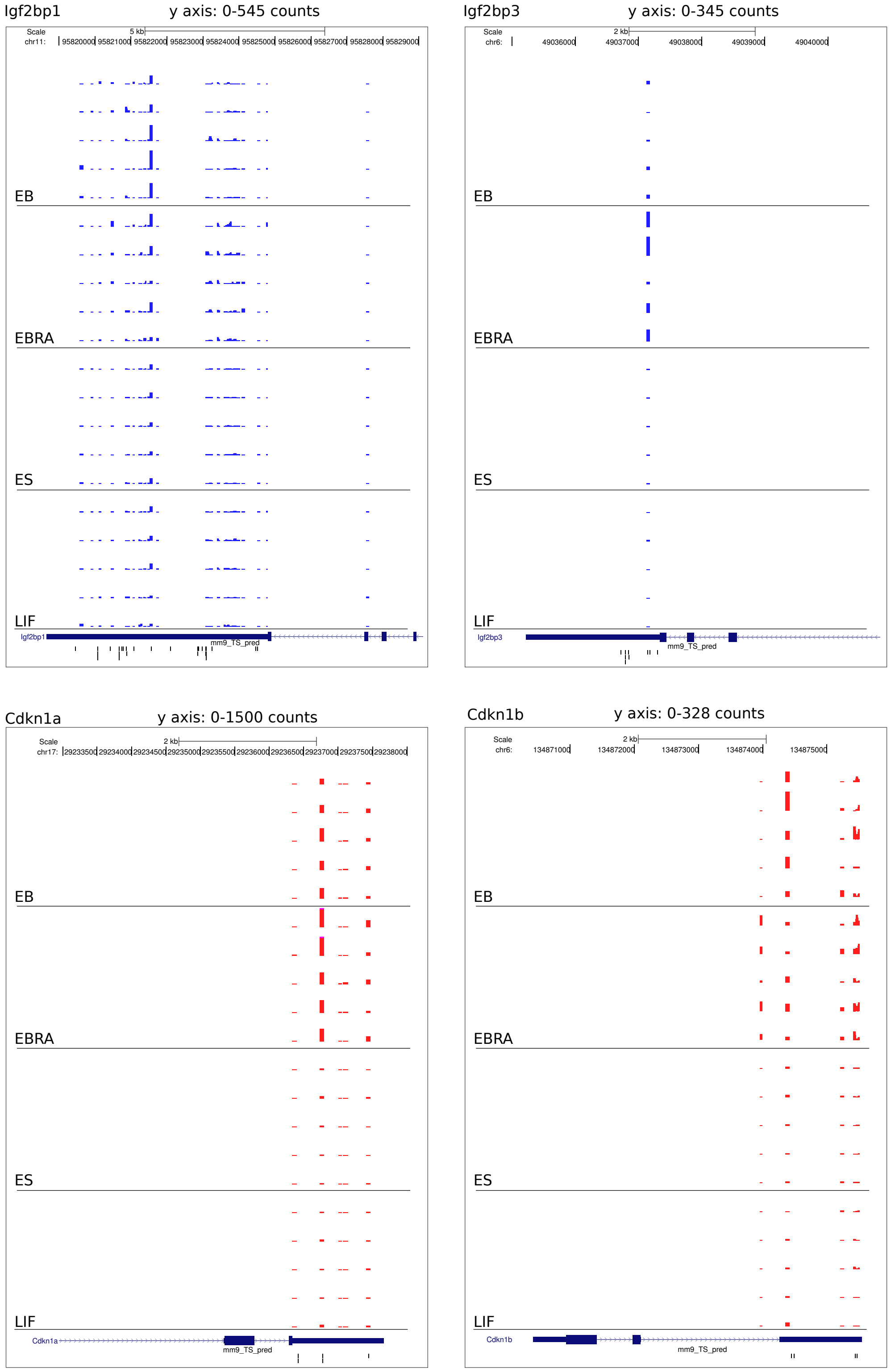


Figure 28. Genes of Insulin Growth Factor Signaling Pathway with their RISC binding sites.
Figure shows overview of the RISC binding portions of the genes, along with the TargetScan prediction track. RISC binding sites show bars proportional to the numbers of reads registered with text above each image identifies maximum value for the highest bar in all replicates. CLIP peaks from Watson strand appear in red, from Crick strand in blue.

Interestingly, inhibitors of the ligand are also shut down, including Igfbp3, which binds Igf1 directly. In an attempt to clarify this convoluted image, I have measured levels of protein for PTEN, phosphorylated AKT and IGFBP3 (Figure 29, panel A). Consistent with the level of RISC-mediated inhibition, PTEN level increases in EB and EBRA conditions, in which it is poorly bound by RISC, while staying low in ES and LIF. Interestingly, this change may have biological consequence on the levels of AKT, which is severely depleted in the EB condition, only to be replenished in EBRA. Surprisingly, levels of IGFBP3 were not changing as dramatically as their RISC binding would suggest, with only a minor dip visible in EB condition. This may have something to do with the outcome of PTEN up-regulation, because it is known that it can transcriptionally up-regulate levels of IGFBP3. [68]

4.16 Brain Gene Co-Expression Network Association Analysis for Genes Containing Differentially Bound Peaks

We were interested to what extent genes identified through our experiments as having dramatically changing RISC binding sites are characteristic of neuronal lineage. [19] Since this property does not have a defined metric, we have settled for measuring the level of association of a specific gene with others in a particular condition. This can be compared to looking at groups of people frequenting specific places, such as a bar and a library. If I know that John, Josh, Jeremy a Rajesh are friends who frequently meet at the bar, then when I am called by Rajesh and being told that he just saw Jeremy, I can infer that it has happened at the bar and Josh and John were also there. If Rajesh also happens to be an avid reader and participates in the book club with Joan, Janice, Jennifer and Judy, then the information that he is somewhere with Janice identifies the place of the meeting as a library with certain probability. Interestingly, while it is imaginable for Rajesh to be somewhere else with Janice, if I am provided with information that he is also with Joan and Judy, the probability of the meeting place being the library increases (as well as the probability that Jennifer is also there). It is important to note, that a single person (for example Rajesh) may not be enough to identify the place, but a group makes this more likely. Also, it can be argued that it is the group which really decides on the character of the place or meeting, which is certainly true both for people, as well as for genes.

Due to the availability of numerous datasets on gene transcript levels in the brain, we have decided to infer networks for condition-specific genes from their transcriptional co-expression. Two

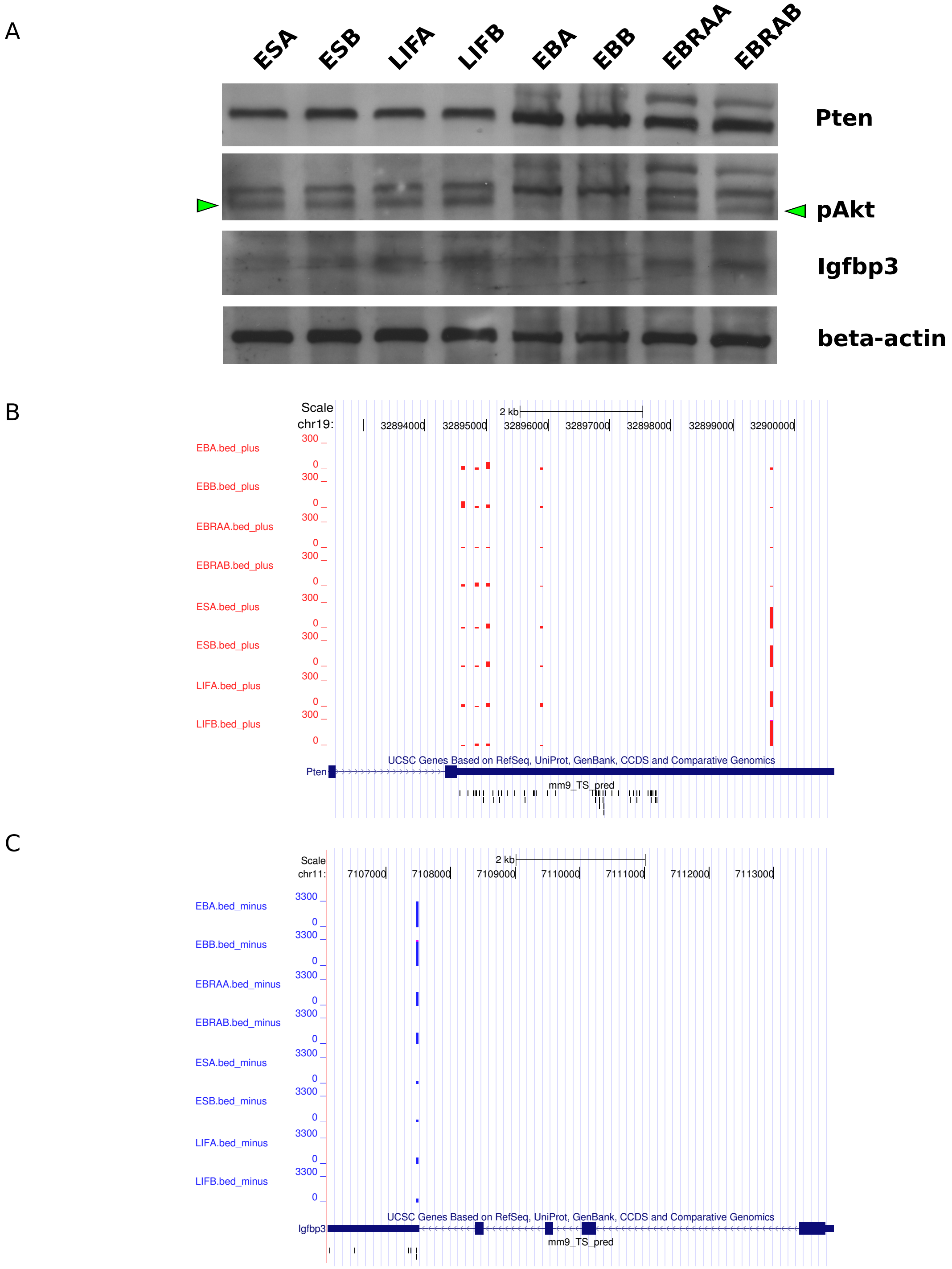


Figure 29. Western blot showing PTEN, phospho-AKT and IGFBP3 protein levels
Subfigure A shows protein levels of Pten, phosphorylated Akt (pAkt) and Igfbp3 protein in comparison to beta-actin. Two replicates of each condition were run and Panel B and C show CLIP peaks for PTEN and IGFBP3 from replicates which were used in western blots. CLIP peaks from Watson strand appear in red, from Crick strand in blue.

networks were built for this purpose - a brain-specific network and non-brain tissue network. Gene Ontology term associated gene sets were used as a reference, while the experimental sets are represented by twenty genes of highest significance in their change of CLIP peaks in transitions from ES to LIF, EB or EBRA conditions. While the highest significance may not be the highest change, those two measures are related and those peaks can be considered strongest consistently changing peaks. Such defined gene sets were scored against both networks and the resulting ROC scores were plotted against each other. ROC scores are normalized to a value range between 0 and 1 with the latter being equivalent to the ideal classifier, which is always right in judging the affiliation of a gene with the proper class. Not surprisingly, such ideal examples are distanced from real-world and classifiers with ROC of 0.9 are uncommon, while ROC of 0.7 is considered typical for succesful classifier, while 0.5 represents a random choice. Values below 0.5 represent an interesting and potentialy useful example of classifier biased against a right choice.

The results of the analysis are shown in Figure 30 showing that the top genes in the ES to LIF transition scored highly (0.67) in the non-brain data network, while their score was significantly less than that of a random classifier in brain (0.36). ES to EB transition is similar, scoring 0.64 in the non-brain network, but being on par with a random classifier in the brain derived network. Stunningly, the ES to EBRA transition was like the mirror image to the ES to LIF transition. The genes associated with this transition scored highly in the brain gene network, slightly above 0.7, while being below random classifier performance in the non-brain network. Compared to GO term gene sets, the ES to LIF and ES to EBRA come out as clear outliers on the ROC plot, while ES to EB gene set resides within the cloud of points. This outlier status is achieved by a combination of uncommonly low score in one of the class and a moderately high score in the other, resulting in largest distance from the diagonal of the plot, a feature characteristic of specificity. It is noteworthy that there are points present with an ROC of 0.9, however they achieve such high value in both networks and therefore are useless from the perspective of classification. Those points represent ubiquitous genes, important for metabolism of any cell and therefore being not characteristic to either of the cell-type specific networks. Of interest is also the bias visible especially in the ES to LIF transition, which signifies that this gene set is depleted for brain-specific genes. Together, those data points show a stunning example of gradual transitioning between two identities during the differentiation, whose final point is confirmed by external data to be brain-specific. Arguably, the ES to LIF and ES to EBRA gene sets are some of the most specific in

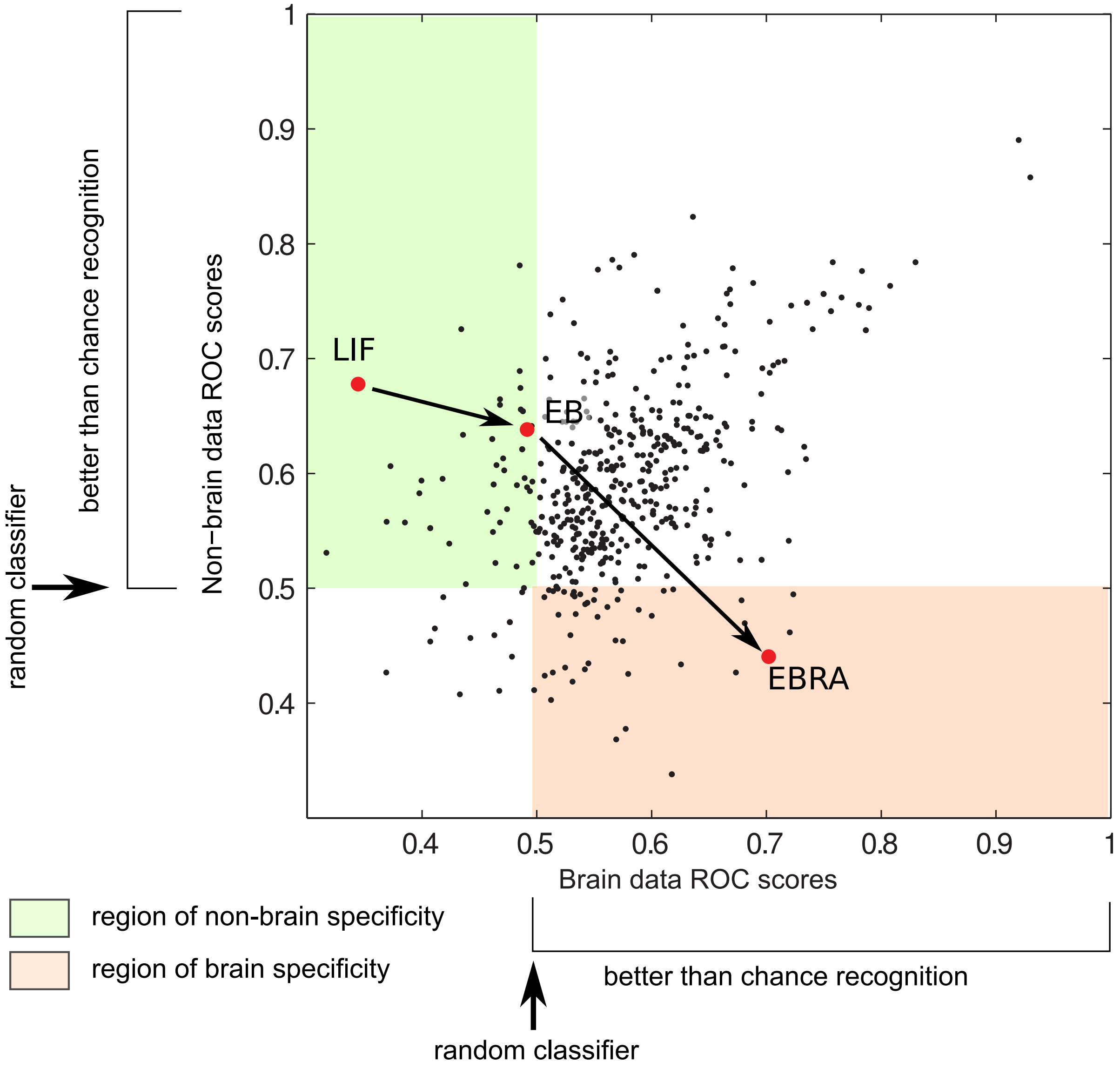


Figure 30. Differrentially bound sets of genes progress from non-specific to being brain specific according to external gene co-expression datasets.
The values of the X and Y for points on the scatterplot represent area under the curve for Receiver-Operator Characteristic (ROC) for gene sets in brain (X) and non-brain (Y) co-expressed gene networks. Black dots correspond to gene sets linked to Gene Ontology terms, while red dots represent gene sets associated with transitions between ES condition and LIF, EB and EBRA. While being significantly better than a random classifier in their respective categories, LIF and EBRA show an interesting bias in complementary category. Both properties are indicative of specificity towards dominant category, more information in the text.

their respective categories. This is particularly interesting in case of the latter, since among numerous plotted neuro-specific GO terms it still appears as one of the most significant outliers. This argues in favor of the hypothesis that our experimentally derived set of genes whose RISC binding changes in EBRA condition is characteristic of neuronal tissue.

5 Conclusions and Perspectives

Not until the beginning of 2000s was it realized that microRNA-mediated regulation is a widely adopted mechanism of influencing protein levels in the cell. [1] Shortly thereafter, examples of gene regulation by microRNAs began to accumulate, and the need for producing a global overview of RISC binding sites in the transcriptome became apparent. In 2003, Ule, Jensen and colleagues developed an experimental method for cloning transcriptome fragments associated with RNA binding proteins and in 2009 this method was successfully applied to the problem of RISC-bound transcripts by members of the same research group, directed by Prof. Robert Darnell. [12, 62] Our efforts in determining target sites began shortly before those developments, leading to a collaboration with Prof. Darnell's group, which helped us to adopt their method. Today, global surveys of RISC binding sites are beginning to accummulate, and we are hoping that through the application of the technical approaches and statistical analysis methods described in this work, they will become a standard method of investigation for questions of microRNA regulation.

The HITS-CLIP method produces data which are direct evidence of RISC binding to specific places in the transcriptome. It can be used to create an ultimate map of microRNA binding sites in the transcriptome improving on existing predictions. With rich, transcriptome-wide experimental data on target sites it is now possible to develop new generation of target site prediction algorithms with improved accuracy due to incorporation of both documented and un-documented ARGONAUTE binding modes. CLIP and related methods have an added advantage in that they can be transformed into semi-quantitative data, namely counts per transcriptome section. This gives us the experimental information about strength of the binding, similar to the score calculated by some of the predicition algorithms. These data can be used to create a snapshot of RISC binding sites in a particular cell line, condition, or time point. Therefore CLIP, unlike bioinformatics predictions can be used to track changes in strength of RISC binding to target sites during transitions between biological timepoints. Additionally, CLIP captures RISC interactions as they are influenced by other RNA-binding proteins, something which cannot be reliably predicted at this time. Summarizing, the difference between this kind of experimental snapshot and predictions is akin to the one between a recent satellite photo and an old map from an atlas of the world.

Despite filling a need for experimental method for determining target sites HITS-CLIP has some significant caveats. One of the most consequential is lack of data about influence of RISC binding on protein levels. This makes discussion about functional impact of CLIP-measured RISC binding difficult, forcing us to adopt assumption that increased RISC targeting of specific biological pathway must have functional consequences. The other disadvantage is the inability to identify from experimental data which specific microRNA is responsible for binding to a target site. This problem can be solved to some extent by use of sequence based target site prediction algorithms to generate this information in an indirect way. From this perspective CLIP can be considered as a method for narrowing regions for prediction algorithms to the extent when prediction itself becomes trivial. Altogether those shortcomings restrict the usability of the method, however the data on the location and strength of RISC binding provided by CLIP is a necessary prerequisite for further research on the impact of microRNA-mediated inhibition.

5.1 Adapting Methods for Quantitative Comparison of Count Data to RISC Binding Changes in Differentiation Comparisons

Our efforts have focused on the problem of using the data that can be produced with CLIP to compare different biological conditions. Biological data is inherently variable, and the extent of RISC binding to a particular site exhibits random fluctuations. However, using several replicates of the same biological conditions, we are able to assess the range of regular fluctuations for a particular condition or time point and use this information to produce a list of RISC binding sites which change as a consequence of underlying functional change. This necessitates the introduction of a statistical method that is appropriate for handling counts data in biological samples. For this purpose we have adapted a negative binomial based method, which was initially developed for comparison of transcript counts in sequencing data. Since cloned CLIP sequence tags are just sections of transcripts this adaptation comes naturally.

Counts of the cloned CLIP sequence tags for a particular binding site depend both on the numbers of specific microRNAs incorporated into the RISC, as well as the number of physical copies of this transcript available for interaction. In comparison, the RNA-seq method has its readout method, the transcript read count, more directly related to the number of physical transcript molecules. This combination of two major variables in CLIP requires a separate method for statistical comparison, and we have adopted a generalized linear model for this purpose. This method was developed for us by our collaborator Dr. Simon Anders, the author of the negative binomial based method for calling differential expression in counts data also applied by us. [2, 3]

While we were able to demonstrate a measurable impact of the change in microRNA levels on target site counts, as evidenced by shifts of ECDF functions for miR-19 and miR-367 (see Figure 16); changes of transcript levels, on the other hand, had only a minor impact on CLIP target site counts, as evidenced by the close to zero correlation coefficient measured between CLIP and RNA-seq reads. Consequently, fold changes in RNA-seq and CLIP show only a minor correlation, and the resulting correction for transcript abundance in the generalized linear model is only slight, as can be appreciated in Figure 17. However, its results on the outcome of calling differential change for target sites depends on the pair of conditions compared. While for the ES to EBRA comparison, nearly 80% of differentially bound target sites determined by GLM are also called by DESeq, this fraction decreases to 17 - 20 % in other comparisons to EBRA. At the same time, both methods bring their own set of additional hits. We expected that additional noise introduced by the RNA-seq data would diminish the number of differentially called hits, and while this is true for the comparisons to EBRA, those additional target sites considered by GLM were surprising to us. Our functional analyses show that merging hits coming from both methods results in additional enrichment for neuro-specific genes with functional relationships. This is visible in our GO term network analysis, where inclusion of GLM-derived genes results the in emergence of a connected network of neuro-specific terms, which would otherwise appear as disconnected separate sub-networks. Initially, we did not intend to merge gene sets from those two sources, expecting them to be functionally separate (i.e. assembling into separate functional sub-networks); however, when we input them as separate sets into the program in order to plot the resulting terms and create a map color-coded according to the contributing source, the resulting network was much more connected, especially in the case of neuronal term sub-network. This merging of the networks shows that both methods produce sets of neuro-specific genes that complement those from the other method in creating functional groups. This is evidence for the usefulness of both methods in detection of functionally relevant changing genes. The possible explanation for additional target sites brought by the GLM method of calling differentially changing sites is illustrated by situation where smaller changes in CLIP counts go against changes in transcript levels. While CLIP changes on their own would not be counted as significant by DESeq, when corrected by transcript level change they become large enough to be statistically significant.

Since our methods produce significant numbers of genes (from tens to above a thousand), it is difficult to translate those changes into biological consequences for the tissue assayed. As mentioned previously, we have used GO term search to detect genes the co-regulation of which may indicate a specific biological function or pathway affected. An advantage of this approach over picking a couple of examples, or so called "interesting genes", and building the model for consequence of their regulation with a manual literature search is that it is unbiased and also has an existing statistical model. While it is permissible not to restrict the search for GO terms using statistical significance and see if a group of not significant terms form a functional group, we have decided to be more conservative and restrict the search to GO terms which pass the 0.05 level of multiple testing adjusted p-values.

The analysis we performed is relatively conservative, after filtering weak sites by requiring more than twenty independent hits per peak, we are using genes with peaks passing a FDR level of 0.05 for further analyses. We require those genes to form sets of functionally related genes, which pass the same level of significance in another test, this time for the sets. Even more, most of the networks built from those significant functional nodes also pass statistical tests formulated for statement that they form a functionally connected network. Finally, while only the comparison of LIF to EBRA has been plotted in detail for the sake of simplicity, we have analyzed all of the comparisons: ES to EBRA, LIF to EBRA and EB to EBRA, looking for gene groups which are present in all analyses. In fact, in the functional analyses, the problem we were facing was not the lack of signal, but rather that the last step of grouping nodes into sub-networks was sensitive to the grouping value used (kappa) and could result in grouping additional terms, making the network confusing and difficult to interpret. We have opted instead for having a larger number of sub-networks with clearer functional relationships by setting kappa value to 0.5, higher than default 0.3. It is notable though, that the sub-networks shown in Figure 22 can further merge, indicating further functional cooperation between the genes.

5.2 Gene Ontology Analysis of Gene Set Differentially Bound by RISC in EBRA Indicates Enrichment in Neuro-specific and Regulatory Genes

Since our terminal differentiation point should force cells to adopt neuronal identity, we were looking for functionally related terms. Genes associated with neuro-specific terms were in fact the most highly significant and at the same time the most numerous group detected among the regulated genes in EBRA condition. This result was persistent for all of the comparisons to the EBRA condition, although the LIF to EBRA transition presented the most clear image of this.

Of interest is also the fact that, of the neuro-related terms, most actually suggest involvement in differentiation, as opposed to functions of adult neurons. Some of the gene groups also suggest specific cell fate specifications, such as glial tissue or oligodendrocytes. While this information can be used to suggest the identity of the cells, we are cautious not to over-interpret it, since some of those detailed specifications seem counter-indicative. This may indicate a possibility of assaying a mixed population of neuronal cells or may also indicate that those gene groups have unknown pleiotropic functions. Of interest is also the presence of genes associated with the embryonic developmental process of neural tube formation; since this is a natural way of specifying neuronal tissue in the embryo it may indicate that cells forced to undergo directed differentiation need to recapitulate this developmental process on the molecular level. Ultimately, the presence of genes involved in membrane depolarization suggests that at least a fraction of differentiated cells adopted this function, since it is restricted to a small set of specialized cells that includes neurons.

Additionally, groups of genes involved in various forms of regulation are also present among changing genes in EBRA, and we speculate that their enrichment may be highly consequential to differentiation process. We observe groups which regulate cellular functions through histone modifications, ubiquitination of proteins for degradation, microRNA-related control, alternative splicing and additionally genes/groups influencing a major cellular mechanism, such as cell cycle control. While its absence is not conclusive, regulation of transcription is surprisingly not present on this list. The presence of those groups may indicate that a cell undergoing differentiation and restricting its multiplication potential may have to modulate the strength of its own regulation mechanisms. Additionally, specific up-regulation of the alternative splicing terms may indicate that in the differentiation process, the cell has to increase its repertoire of alternative splicing isoforms or use this mechanism to modulate transcript levels in order to formulate its terminal identity. In fact reports confirming this notion do exist, making involvement of microRNAs in its regulation even more consequential. [18, 21]

It appears, however, that differentiation is a complex process involving many cellular mechanisms for regulation of transcripts and protein levels. This cooperation of various processes makes us expect that the differentiation is achieved through network-like summation of small changes rather than one or few major drivers. In light of this fact, it is highly unlikely for us to achieve the ultimate goal of modeling this differentiation process without a dedicated effort to assay all of those mechanisms involved.

However, this is the first study of this type revealing a large degree of microRNA regulation on the differentiation process. Arguably, microRNAs form another layer of regulation, with its intricate network capable of extending control over large number of transcripts. Until very recently, the identity of regulated genes was locked away from us barring global assessment of microRNA regulatory potential. This study shows that it is possible to find the sites of RISC interaction with transcripts and furthermore demonstrates that it is possible to find the target sites for which RISC affinity changes in the differentiation process.

5.3 Gene Set Enrichment Analysis Indicates microRNA-mediated Regulation of Major Signaling Pathways of Known Neuro-specific Roles

In our efforts to identify some major players in the differentiation process we have switched our attention to signalling pathways, since they are often believed to be drivers of differentiation. Genes belonging to signaling pathways often came to light of our analyses as being differentially bound by RISC and as it turns out, entire sets of genes participating in signaling are present in our datasets. Since we reasoned that pathways, being composed of a smaller number of genes, are a worse target to assess enrichment we were looking for any enriched terms relating to pathways, not for the formation of groups, as we did for broad categories described above. Looking at various comparisons, we have found several signaling pathways for which genes appear to be enriched. Interestingly, those pathways seem to be involved either in the process of pluripotency maintenance (Activin, Nodal, FGF) or neuronal cell specification, like the TGF, Insulin receptor, Nodal, VEGF, and Retinoic Acid pathways. [63, 59, 26] The last one is the best example of a positive control for the methodology used for assigning function to gene groups presented in this research, since retinoic acid is the molecule which we used for inducing differentiation, and it is logical to expect some form of response in the pathway itself. [28]

Interestingly we discovered that Insulin Growth Factor Signalling Pathway, which is present among the enriched gene groups, is especially targeted by RISC, having most of its genes differentially bound between the conditions. Interestingly, all of those genes appear to harbor RISC binding sites in their 3'UTR region, consisting with the hypothesis that part of the transcript as a primary place for microRNA regulation. Most importantly, both of this pathway's downstream effectors, CDKN1A and CDKN1B, are targeted in the EBRA condition. Since both of those genes are inhibitors of cyclin, their downregulation by microRNA may have a stimulating effect on the cell cycle. We were interested in whether microRNA inhibition of Insulin signaling components such as Pten is in fact consequential to its function. A Western blot performed on protein samples from material used in the CLIP experiment shows that the PTEN level is in fact negatively correlated with the level of RISC binding in one of its 3'UTR sites. While we observe lower levels of PTEN in ES and LIF replicates, it goes up signficantly in EB and EBRA conditions. The functional effect of this change can be assayed by measuring the level of phosphorylated AKT. While Akt transcript has a RISC binding site, it does not register as differentially changing, and furthermore, upon manual inspection the level of binding seems matched in all of the conditions. Interestingly, p-AKT goes down to almost to zero in EB condition, just as it should if down-regulated by increased PTEN levels, but returns to ES phosophorylation levels in the EBRA cells, despite a high PTEN level. This may indicate that instead of being a decisive factor controlling steady state levels of proteins, in this case microRNAs may just be one of many factors involved in a dynamic system. Additionally, we have also assayed the levels of IGFBP3, a secreted protein which can bind and sequester IGF1, an IGF receptor ligand. This protein was interesting for us, since it appeared in our datasets even in some of the previous analyses which were not described in this work, and attracted our attention as having a single RISC binding site, whose occupancy changes remarkably between conditions. Interestingly, the result of this remarkable change in RISC binding transforms just to a small dip in protein levels in the EB conditions. While disappointing, this may have something to do with the fact that increasing PTEN levels up-regulated the production of the IGFBP3 via indirect transcriptional activation. Overall, we can speculate that microRNAs may have the primary aim of shutting down, or what is more probable, temporarily down-regulating this pathway for purposes of subtle regulation of cell cycle control. While we were disappointed in not providing the definitive answer for the role and effects of this aggregation of microRNA-mediated regulation on this pathway in neuronal differentiation, we have also realized that in order to study this in detail, we would have to design a specialized and extended study. We would have to assay protein levels and known phosphorylation states involved in the transduction of signal in this signalling pathway. More importantly, rather than taking single snapshots of defined conditions, we would have to take a time course that includes transition stages between conditions. This sequential sampling would make it possible to observe dynamics of changes and possibly trace cause and effect in this network of interactions. To put it in mathematical perspective, instead of table of steady state levels, we would like to have differential equation describing their changes in time. Such temporarily fine grained experiment carries astounding amount of workload and money requirements if performed using CLIP and RNA-seq. However with this interesting gene set defined through our experiments, we can attempt to perform such a time course using more established methods, like western blots, qPCR specific to transcripts of interest and reporters for 3'UTRs of the genes involved. One remaining requirement for such study, especially when performed using focused methods, is detailed knowledge of all relevant players and interactions between them. This may be difficult to achieve without further investigations and collaboration with scientists specializing in particular pathway. On the upside, the incompleteness will be apparent when the model will fail to replicate measured protein level changes warranting search for missing or misinterpreted element of the model.

Witnessing the consequence of miR-19 and miR-367 upregulation in the increase of their target sites, we were interested if the affected genes specific to either of the microRNA on their own contain a group of genes which possibly cooperate in a specific cellular process. Our predictions were confirmed by finding neuro-specific gene groups in the sets regulated by both of the microRNAs, however restricted to just few genes in each case. Interestingly, miR-19 also contained genes associated with signalling pathway responses, which were absent in miR-367 targets. Altogether this is suggesting that microRNAs have varying specificity towards different functional groups of genes, thus possibly being responsible for targeting slightly different processes or co-regulating processes together with other microRNAs in combinatorial fashion.

5.4 External Brain Co-Expression Studies Support EBRA Condition as Being Highly Enriched in Genes Specific to Brain

We felt encouraged by the presence of neuronal fate-related genes in our set of genes bound differentially by RISC, because it confirmed specificity of our method in the NPC context. However, those genes constitute only about 10% of the total number of differentially called genes. And while we expect the rest of those genes to have functions related to neuronal fate specification, we are stuck with their identification as related to processes of basic cell function control without the possibility to assess their significance for the former. We were looking for a method to assess the specificity of our set in its entirety towards neuronal tissue. In collaboration with Dr. Jesse Gillis, we have employed cross-validation assessment of neighborhood association derived from co-expression networks in order to evaluate whether our differentially bound genes belong to a set derived either from brain tissue or non-brain tissue. [19] The method results in a single value, the Receiver Operator Characteristic, which measures a likelihood of our gene group belonging to either a network of genes derived from brain or non-brain tissue. Our gene sets show an interesting progression from being specific to non-brain tissues in case of LIF, through being in the middle between both and therefore not specific to either (EB), to being very brain-specific in case of EBRA. The extent of this specificity is worth being emphasized, since it appears that our EBRA-derived gene set is better at characterizing brain tissue than any of the GO term-associated gene sets, including those describing neuronal functions.

5.5 MicroRNA Profiles and RISC Binding Data Suggest Greater Role of MicroRNA in Neuronal Lineage Specification Than Differentiation of ES Cells

Despite published evidence for the importance of microRNAs for ES cell differentiation away from the pluripotent state, we were unable to find a good candidate for microRNA responsible in this process. Since the LIF factor is important for maintenance of pluripotence of ES cells, we were hoping to pick up specific microRNAs changing during the transition from ES to LIF negative cells. This however turned to be impossible, since microRNA expression profiles for both conditions look strikingly similar. Only the EB and EBRA condition show significant appearance of other microRNAs, namely: miR-27, miR-367/92 seed family in the EB, and miR-18, miR-19 and miR-130/301 family in EBRA. Incidentally, the three EBRA-specific microRNAs are implicated to be neuro-specific in the literature, thus adding credibility to the hypothesis of their importance to neuronal fate specification. [33]

Suprisingly, while being implicated in the process of ES cell differentiation as a master switch controlling differentiation process, we were unable to find evidence that let-7 changes significantly between our experimental conditions. [46] The microRNA family of let-7 remained lowly expressed in all our samples, without any of its members or the family as a whole changing considerably between conditions. Since the Gene Ontology analysis of genes associated with our target sites implicates neuronal lineage specification, this may indicate that in cases like ours ES cells may differentiate without significant involvement of this microRNA family. Another possibility is that our specific conditions simply miss the differentiation step with ransient up-regulation of let-7 family, however the progression of differentiation is typically described as the one with the fraction of let-7 steadily increasing in the microRNA pool.

Outside of the few microRNAs which achieve high expression, our ability to determine which changing microRNA has impact on cellular processes of differentiation is severely limited. We have used ECDF plots (see Figure 13) to assess impact of changing microRNAs on their target sites and determined that only in case of miR-19 and miR-367 we can observe a considerable increase in RISC binding. Our method sufers however from imprecise definition of the set of target sites associated with specific microRNA. Our expression profiles show many candidate microRNAs changing by several folds of magnitude, however since their expression remains inherently low, we are cautious to indicate their involvement in neural fate specification. It is possible that by improving definition of specific microRNA associated target site sets we will achieve an improved capability of calling microRNAs with considerable impact on their binding sites. Even at this stage it is an useful method of narrowing down the list of microRNA candidates for experimental validation.

5.6 Published Data on Mouse Embryonic Stem Cells RISC Binding Sites Are Highly Consistent with Our Reference Set of Target Sites

Leung et al. have published a study describing a set of microRNA target sites for ES cells, thus providing us with a natural verification for one of the conditions used in our experiments. [41] We have concluded that we observe the experimental support in form of reads intersecting with our reference set of target sites present in ES for over 60% of sites, coming from any annotation category. The correlation between the data is positive, but not great, reaching only ~0.4 Pearson coefficient and therefore being rather unsuitable for numerical comparisons. This is hardly surprising, given the fact that different cell lines, antibodies and finally different methodology for filtering the results were used. The results of Leung et al. mirror our observations in motif prediction, finding a [A/C]AGTGC as one of the most highly enriched, which corresponds to our most highly enriched [A/C]AGTGC motif complementary to the miR-290/302 family seed site. In both cases, the motifs were enriched to a similar extent, reaching 18% of the total microRNA target sites in theirs and 25% in our datasets. Additionally, they report a GGCTGG motif, which we have not identified in our dataset, and a low complexity C/T rich motif, which is variable and present in high numbers. Interestingly, we believe that we can match this motif with corresponding highly variable and very abundant motifs from our datasets, namely TTTCTT and CTTCTT, also present in considerable excess. Similarly to our data, Leung et al. demonstrate shift of the binding strength for target sites matching the main motif using ECDF function in comparison between wild type and Dicer-null cells. We did the same in our condition comparisons registering highly significant shifts, particularly for miR-19, miR-367, but also for other microRNA families to a lesser extent. Interestingly, the miR-290/302 family cluster also shows a shift in the same direction as those, while reducing its overall percentage; however, we interpret this as the outcome of switching from the miR-290-295 sequences with A on 5' end to miR-302 mature strands, having U as the first base. The 5'U mature strands are proven to have greater affinity towards the ARGONAUTE proteins.

5.7 Neuronal Differentiation of Human Embryonic Stem Cells Show Extensive Parallels to Our Results

There is an interesting parallel study by Lipchina et al. on neuronal fate specification in ES derived cells performed in human using PAR-CLIP method. [43] Reassuringly, our results show certain level of agreement with this data, particularly in identifying miR-18, miR-19 and miR-103 as microRNAs enriched in neuronal progenitor cells. However, while in our data we see the high extent of miR-290-295 cluster expression in ES cells, which slowly lowers its percentage ratio in favor of miR-302/367 and others, the human genome only has miR-302/367, but there is no close equivalent of miR-290-295 cluster. Lipchina et al. see downregulation of the miR-302/367 cluster in favor of neuro-related miRs: miR-9, miR-15, miR-16, miR-17,miR-18, miR-19, miR-20, miR-21, miR-103, miR-153. Interestingly, like in our data, the inhibition of the miR-302/367 cluster corresponds with up-regulation of their predicted target sites, a situation which is similar to our miR-290/302 target sites. In contrast to their results, we have detected a shift in ECDF curves for most of the major target site families for which microRNA expression increases in NPC. The study of Lipchina et al. also identifies a number of genes with functional target sites and they are also present in our data. Most notably, these include CDKN1A, PTEN, LEFTY1 and LEFTY2 and TXNIP, but also others. Based on their reference set of targeted genes, they suggest a heavy involvement of TGF-beta/BMP and Activin/Nodal signalling in the process of neural lineage induction, which mirrors the results of our GO term enrichment searches. While we note those pathways as enriched in our data, we also identify genes belonging to other pathways, whose role in neuronal differentiation is supported by literature data. Additionally, they propose three new inhibitors for TGF-beta activity, namely TOB2, DAZAP2 and SLAIN1, the first two of which seem to be supported by our comparison data; the full extent of differentially bound genes related to TGF-beta pathway can be seen in Figures 31 and 32.

Our approach has an advantage of quantitatively comparing target sites between conditions, as could be seen in the case of SMAD2, where they have determined it as specific to differentiation, while our data indicate equivalent binding in all of the conditions.

5.8 Future Directions

Thinking about extending our research, we have noted that comparably high-throughput protein level assessment method to match the data from RNA-seq and HITS-CLIP is necessary for evaluating the outcome of the RISC binding on protein levels and therefore the real impact of microRNA-mediated inhibition on cellular functions. Since mass-spectrometry appears to still be limited and unable to produce an exhaustive snapshot of condition-specific protein levels, we propose that recently developed high-throughput ribosome footprinting technique is a suitable candidate to fill this gap. [31] This method has an advantage of assaying current levels of protein synthesis and therefore may be even better suitable for assaying the impact of microRNA expression change on protein translation.

Demonstration of microRNA impact on function can be also done on the smaller scale of a cellular circuit. Our studies point to the Insulin Receptor Signaling Pathway as being a good candidate for such investigations, thanks to the demonstrated RISC binding to its major components, as well as its link to important cell cycle control effectors. In fact, the time course involving the transition between LIF stage through EB to EBRA, while focusing just on the proteins of this pathway may be the least complicated way of assessing microRNA impact on its regulation.

While we have demonstrated that differentiation into neuronal progenitor cells is accompanied by upregulation of specific microRNAs, which have a consequence for RISC binding to their target sites, it would be interesting to see if this could be recreated by transfection of those microRNAs, and if the resulting RISC binding changes would be enough to drive cells into differentiation. This study would have an advantage of potentially solving a long standing problem of finding the definitive set of target sites specific for a given microRNA, a task which at this moment can be accomplished only in

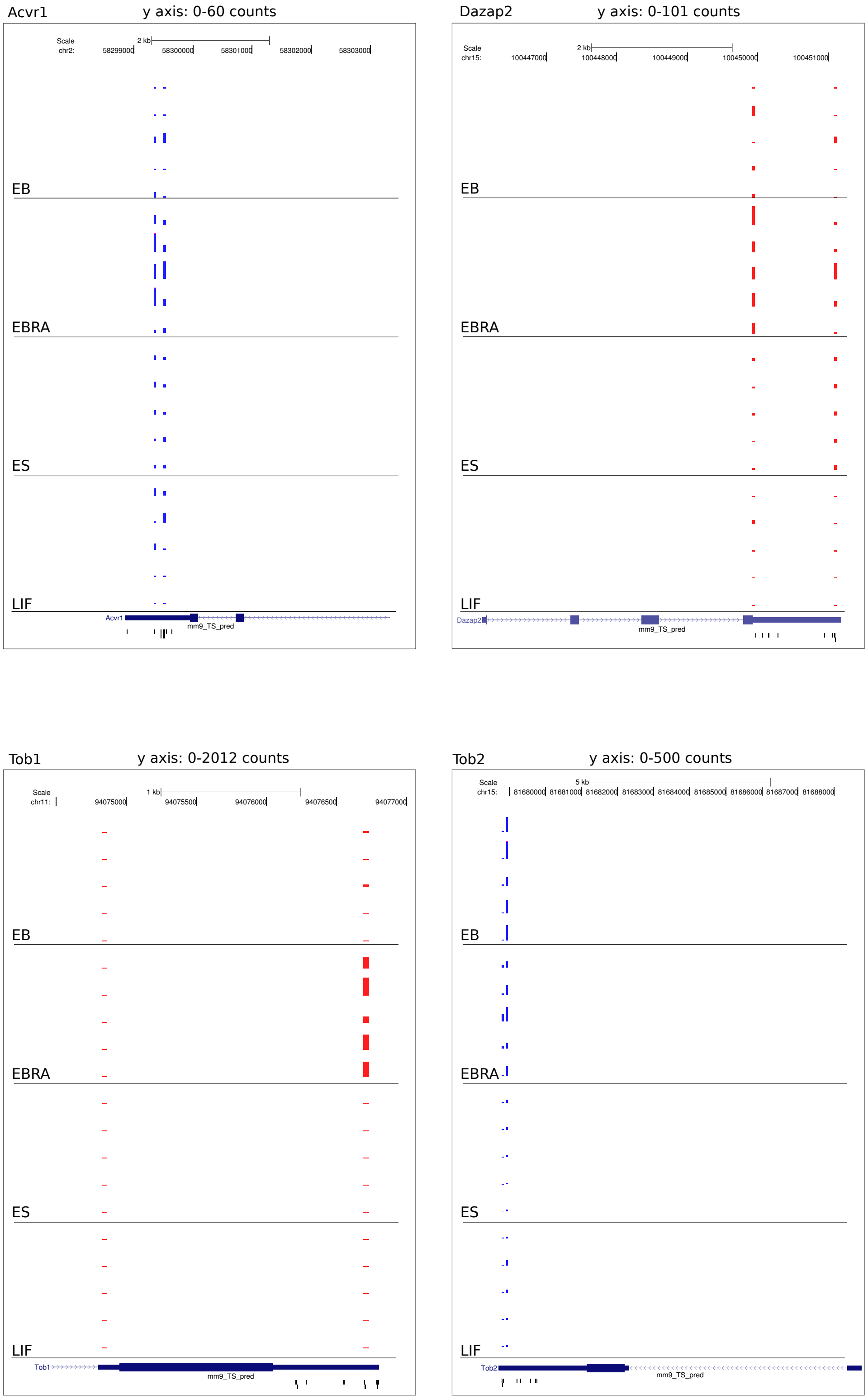


Figure 31. Genes of TGF-beta Signaling Pathway with their RISC binding sites - part 1.
Figure shows overview of the RISC binding portions of the genes, along with the TargetScan prediction track. RISC binding sites show bars proportional to the numbers of reads registered with text above each image identifies maximum value for the highest bar in all replicates. CLIP peaks from Watson strand appear in red, from Crick strand in blue.

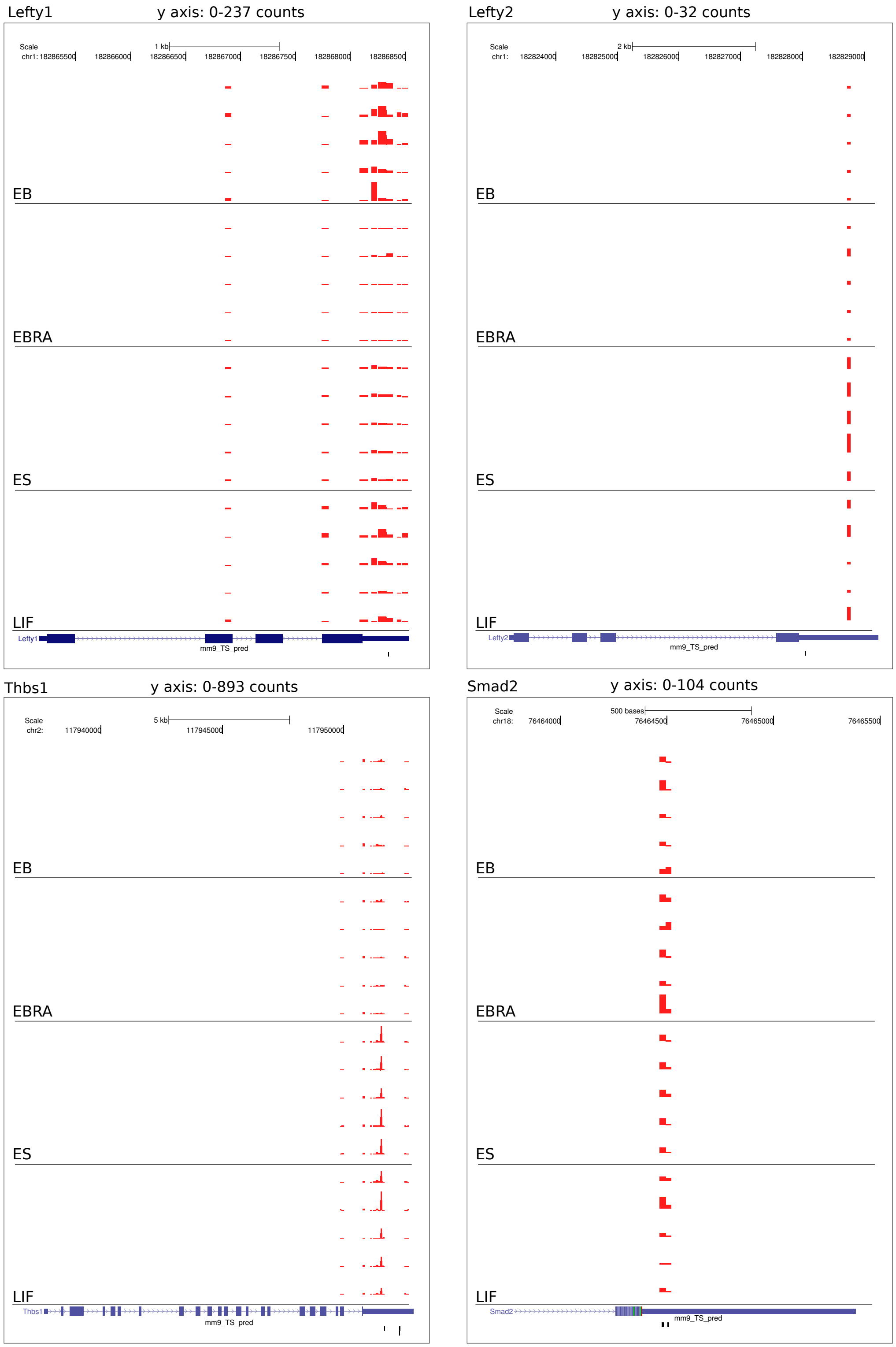


Figure 32. Genes of TGF-beta Signaling Pathway with their RISC binding sites - part 2.
Figure shows overview of the RISC binding portions of the genes, along with the TargetScan prediction track. RISC binding sites show bars proportional to the numbers of reads registered with text above each image identifies maximum value for the highest bar in all replicates. CLIP peaks from Watson strand appear in red, from Crick strand in blue.

approximation through seed site matching or some heuristic algorithms. This problem may be solved in the future with help of the CLASH and related methods of direct target site identity determination; however, at this time there is only a single study of this kind.

5.9 Closing Remarks

While encouraged by the initial success with experimental derivation of target sites and measuring the level of their association with RISC, we are aware that this study is only a first step for functional characterisation of the processes involved in neuronal differentiation. However, we have validated that this differentiation model, as well as experimental techniques and methods of analysis that we have developed, do produce a set of target sites which can be quantitatively compared between the differentiation conditions. Furthermore, the genes determined to be differentially bound by RISC in the terminal differentiation condition of EBRA are enriched for those involved in neuronal fate specification, as determined by two independent methods. Moreover, we demonstrate that our approach is capable of shortlisting signalling pathways supported by literature as being involved in the process. We are also capable of demonstrating the consequence of microRNA level change on RISC binding in corresponding target sites, athough of note is the approximate definition of such miR-specific set. One notable shortcoming of our study comes from the fact that our method measures directly just the transcriptome binding of RISC. We have yet insufficient data to infer to what extent such changes in RISC affinity are consequential on corresponding protein production or mRNA levels. Our data, however, helps to identify interesting examples of microRNA-mediated regulation within smaller functional groups, making it accessible to study using low throughput methods.

6 Appendix

Supplementary Digital Data Files are in the human readable tab separated format with first line serving as a column header if not stated otherwise. They are attached on an accompanying digital media carrier.

1. Table of RAW target site counts with corresponding RNAseq reads in reference dataset of target sites with corresponding annotation, Intersection with Leung et al. dataset and TargetsScan predictions

   exportCLIPNSRAnnot

2. Table of variance stabilized and library size normalized target site counts in reference dataset

   exportVSDCLIP

3. Table of differentially bound target sites according to DESeq

   files named export*****dq, where ***** is one of the five comparisons

4. Table of differentially bound target sites according to generalized linear model

   files named export*****glm, where ***** is one of the five comparisons

5. Code for generalized linear model

   runningGLM.R contains example code and DESeq_extra.R code necessary to run glm, clip_analysis.pdf contains description of related dataset

6. Table of primer sequences for forward and reverse pool of Not-So-Random transcriptome sequencing method

   primersHiSeqClipNsr.xlsx

References

[1] V Ambros. micrornas: tiny regulators with great potential. *Cell*, 107(7):823–6, Dec 2001.

[2] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

[3] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from rna-seq data. *Genome Res*, 22(10):2008–17, Oct 2012.

[4] Christopher D Armour, John C Castle, Ronghua Chen, Tomas Babak, Patrick Loerch, Stuart Jackson, Jyoti K Shah, John Dey, Carol A Rohl, Jason M Johnson, and Christopher K Raymond. Digital transcriptome profiling using selective hexamer priming for cdna synthesis. *Nat Methods*, 6(9):647–9, Sep 2009.

[5] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, May 2000.

[6] Timothy L Bailey. Dreme: motif discovery in transcription factor chip-seq data. *Bioinformatics*, 27(12):1653–9, Jun 2011.

[7] E Bernstein, A A Caudy, S M Hammond, and G J Hannon. Role for a bidentate ribonuclease in the initiation step of rna interference. *Nature*, 409(6818):363–6, Jan 2001.

[8] Gabriela Bindea, Jerome Galon, and Bernhard Mlecnik. Cluepedia cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics*, 29(5):661–3, Mar 2013.

[9] Gabriela Bindea, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf-Herman Fridman, Franck Pages, Zlatko Trajanoski, and Jerome Galon. Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, 25(8):1091–3, Apr 2009.

[10] Julius Brennecke, Alexander Stark, Robert B Russell, and Stephen M Cohen. Principles of microrna-target recognition. *PLoS Biol*, 3(3):e85, Mar 2005.

[11] Sung Wook Chi, Gregory J Hannon, and Robert B Darnell. An alternative mode of microrna target recognition. *Nat Struct Mol Biol*, 19(3):321–327, 2012.

[12] Sung Wook Chi, Julie B Zang, Aldo Mele, and Robert B Darnell. Argonaute hits-clip decodes microrna-mrna interaction maps. *Nature*, 460(7254):479–86, Jul 2009.

[13] Robert B Darnell. Hits-clip: panoramic views of protein-rna regulation in living cells. *Wiley Interdiscip Rev RNA*, 1(2):266–86, Sep-Oct 2010.

[14] Elad Elkayam, Claus-D Kuhn, Ante Tocilj, Astrid D Haase, Emily M Greene, Gregory J Hannon, and Leemor Joshua-Tor. The structure of human argonaute-2 in complex with mir-20a. *Cell*, 150(1):100–10, Jul 2012.

[15] Anthony P Fejes, Gordon Robertson, Mikhail Bilenky, Richard Varhol, Matthew Bainbridge, and Steven J M Jones. Findpeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–30, Aug 2008.

[16] A Fire, S Xu, M K Montgomery, S A Kostas, S E Driver, and C C Mello. Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *Nature*, 391(6669):806–11, Feb 1998.

[17] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mrnas are conserved targets of micrornas. *Genome Res*, 19(1):92–105, Jan 2009.

[18] Mathieu Gabut, Payman Samavarchi-Tehrani, Xinchen Wang, Valentina Slobodeniuc, Dave O'Hanlon, Hoon-Ki Sung, Manuel Alvarez, Shaheynoor Talukder, Qun Pan, Esteban O Mazzoni, Stephane Nedelec, Hynek Wichterle, Knut Woltjen, Timothy R Hughes, Peter W Zandstra, Andras Nagy, Jeffrey L Wrana, and Benjamin J Blencowe. An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming. *Cell*, 147(1):132–46, Sep 2011.

[19] Jesse Gillis and Paul Pavlidis. The role of indirect connections in gene networks in predicting function. *Bioinformatics*, 27(13):1860–6, Jul 2011.

[20] Sander Granneman, Grzegorz Kudla, Elisabeth Petfalski, and David Tollervey. Identification of protein binding sites on u3 snorna and pre-rrna by uv cross-linking and high-throughput analysis of cdnas. *Proc Natl Acad Sci U S A*, 106(24):9613–8, Jun 2009.

[21] Brenton R Graveley. Splicing up pluripotency. *Cell*, 147(1):22–4, Sep 2011.

[22] Sam Griffiths-Jones. mirbase: the microrna sequence database. *Methods Mol Biol*, 342:129–38, 2006.

[23] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Par-clip–a method to identify transcriptome-wide the binding sites of rna binding proteins. *J Vis Exp*, (41), 2010.

[24] S M Hammond, E Bernstein, D Beach, and G J Hannon. An rna-directed nuclease mediates post-transcriptional gene silencing in drosophila cells. *Nature*, 404(6775):293–6, Mar 2000.

[25] S M Hammond, S Boettcher, A A Caudy, R Kobayashi, and G J Hannon. Argonaute2, a link between genetic and biochemical analyses of rnai. *Science*, 293(5532):1146–50, Aug 2001.

[26] K A Heidenreich. Insulin and igf-i receptor signaling in cultured neurons. *Ann N Y Acad Sci*, 692:72–88, Aug 1993.

[27] Aleksandra Helwak, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. Mapping the human mirna interactome by clash reveals frequent noncanonical binding. *Cell*, 153(3):654–65, Apr 2013.

[28] T Herget, H Specht, C Esdar, S A Oehrlein, and A Maelicke. Retinoic acid induces apoptosis-associated neural differentiation of a murine teratocarcinoma cell line. *J Neurochem*, 70(1):47–58, Jan 1998.

[29] Y Hochberg and Y Benjamini. More powerful procedures for multiple significance testing. *Stat Med*, 9(7):811–8, Jul 1990.

[30] Hristo B Houbaviy, Michael F Murray, and Phillip A Sharp. Embryonic stem cell-specific micrornas. *Dev Cell*, 5(2):351–8, Aug 2003.

[31] Nicholas T Ingolia, Liana F Lareau, and Jonathan S Weissman. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802, Nov 2011.

[32] F JACOB and J MONOD. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318–56, Jun 1961.

[33] Ana Jovicic, Reema Roshan, Nicoleta Moisoi, Sylvain Pradervand, Roger Moser, Beena Pillai, and Ruth Luthi-Carter. Comprehensive expression analyses of neural cell-type-specific mirnas identify new determinants of the specification and maintenance of neuronal phenotypes. *J Neurosci*, 33(12):5127–37, Mar 2013.

[34] Chryssa Kanellopoulou, Stefan A Muljo, Andrew L Kung, Shridar Ganesan, Ronny Drapkin, Thomas Jenuwein, David M Livingston, and Klaus Rajewsky. Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev*, 19(4):489–501, Feb 2005.

[35] Fedor V Karginov, Cecilia Conaco, Zhenyu Xuan, Bryan H Schmidt, Joel S Parker, Gail Mandel, and Gregory J Hannon. A biochemical approach to identifying microrna targets. *Proc Natl Acad Sci U S A*, 104(49):19291–6, Dec 2007.

[36] Fedor V Karginov and Gregory J Hannon. The crispr system: small rna-guided defense in bacteria and archaea. *Mol Cell*, 37(1):7–19, Jan 2010.

[37] Julian Konig, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. iclip–transcriptome-wide mapping of protein-rna interactions with individual nucleotide resolution. *J Vis Exp*, (50), 2011.

[38] Ashish Lal, Francisco Navarro, Christopher A Maher, Laura E Maliszewski, Nan Yan, Elizabeth O'Day, Dipanjan Chowdhury, Derek M Dykxhoorn, Perry Tsai, Oliver Hofmann, Kevin G Becker, Myriam Gorospe, Winston Hide, and Judy Lieberman. mir-24 inhibits cell proliferation by targeting e2f2, myc, and other cell-cycle genes via binding to "seedless" 3'utr microrna recognition elements. *Mol Cell*, 35(5):610–25, Sep 2009.

[39] R C Lee, R L Feinbaum, and V Ambros. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, 75(5):843–54, Dec 1993.

[40] Rosalind Lee, Rhonda Feinbaum, and Victor Ambros. A short history of a short rna. *Cell*, 116(2 Suppl):S89–92, 1 p following S96, Jan 2004.

[41] Anthony K L Leung, Amanda G Young, Arjun Bhutkar, Grace X Zheng, Andrew D Bosson, Cydney B Nielsen, and Phillip A Sharp. Genome-wide identification of ago2 binding sites from mouse embryonic stem cells with and without mature micrornas. *Nat Struct Mol Biol*, 18(2):237–44, Feb 2011.

[42] Lee P Lim, Nelson C Lau, Earl G Weinstein, Aliaa Abdelhakim, Soraya Yekta, Matthew W Rhoades, Christopher B Burge, and David P Bartel. The micrornas of caenorhabditis elegans. *Genes Dev*, 17(8):991–1008, Apr 2003.

[43] Inna Lipchina, Yechiel Elkabetz, Markus Hafner, Robert Sheridan, Aleksandra Mihailovic, Thomas Tuschl, Chris Sander, Lorenz Studer, and Doron Betel. Genome-wide identification of microrna targets in human es cells reveals a role for mir-302 in modulating bmp response. *Genes Dev*, 25(20):2173–86, Oct 2011.

[44] Jidong Liu, Michelle A Carmell, Fabiola V Rivas, Carolyn G Marsden, J Michael Thomson, Ji-Joon Song, Scott M Hammond, Leemor Joshua-Tor, and Gregory J Hannon. Argonaute2 is the catalytic engine of mammalian rnai. *Science*, 305(5689):1437–41, Sep 2004.

[45] Sara Macias, Mireya Plass, Agata Stajuda, Gracjan Michlewski, Eduardo Eyras, and Javier F Caceres. Dgcr8 hits-clip reveals novel functions for the microprocessor. *Nat Struct Mol Biol*, 19(8):760–6, Aug 2012.

[46] Collin Melton, Robert L Judson, and Robert Blelloch. Opposing microrna families regulate self-renewal in mouse embryonic stem cells. *Nature*, 463(7281):621–6, Feb 2010.

[47] J A Nelder. Generalized linear models for enzyme-kinetic data. *Biometrics*, 47(4):1605–10; discussion 1610–5, Dec 1991.

[48] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, 20(1):110–21, Jan 2010.

[49] C Pourcel, G Salvignol, and G Vergnaud. Crispr elements in yersinia pestis acquire new repeats by preferential uptake of bacteriophage dna, and provide additional tools for evolutionary studies. *Microbiology*, 151(Pt 3):653–63, Mar 2005.

[50] Judith Reeks, James H Naismith, and Malcolm F White. Crispr interference: a structural perspective. *Biochem J*, 453(2):155–66, Jul 2013.

[51] B J Reinhart, F J Slack, M Basson, A E Pasquinelli, J C Bettinger, A E Rougvie, H R Horvitz, and G Ruvkun. The 21-nucleotide let-7 rna regulates developmental timing in caenorhabditis elegans. *Nature*, 403(6772):901–6, Feb 2000.

[52] Nicole T Schirle and Ian J MacRae. The crystal structure of human argonaute2. *Science*, 336(6084):1037–40, May 2012.

[53] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–504, Nov 2003.

[54] Chanseok Shin, Jin-Wu Nam, Kyle Kai-How Farh, H Rosaria Chiang, Alena Shkumatava, and David P Bartel. Expanding the microrna targeting code: functional sites with centered pairing. *Mol Cell*, 38(6):789–802, Jun 2010.

[55] Zoltan Simandi, Balint Laszlo Balint, Szilard Poliska, Ralph Ruhl, and Laszlo Nagy. Activation of retinoic acid receptor signaling coordinates lineage commitment of spontaneously differentiating mouse embryonic stem cells in embryoid bodies. *FEBS Lett*, 584(14):3123–30, Jul 2010.

[56] Ji-Joon Song, Jidong Liu, Niraj H Tolia, Jonathan Schneiderman, Stephanie K Smith, Robert A Martienssen, Gregory J Hannon, and Leemor Joshua-Tor. The crystal structure of the argonaute2 paz domain reveals an rna binding motif in rnai effector complexes. *Nat Struct Biol*, 10(12):1026–32, Dec 2003.

[57] Ji-Joon Song, Stephanie K Smith, Gregory J Hannon, and Leemor Joshua-Tor. Crystal structure of argonaute and its implications for risc slicer activity. *Science*, 305(5689):1434–7, Sep 2004.

[58] H Tabara, M Sarkissian, W G Kelly, J Fleenor, A Grishok, L Timmons, A Fire, and C C Mello. The rde-1 gene, rna interference, and transposon silencing in c. elegans. *Cell*, 99(2):123–32, Oct 1999.

[59] Jean-Leon Thomas, Kasey Baker, Jinah Han, Charles Calvo, Harri Nurmi, Anne C Eichmann, and Kari Alitalo. Interactions between vegfr and notch signaling pathways in endothelial and neural cells. *Cell Mol Life Sci*, 70(10):1779–92, May 2013.

[60] Daniel W Thomson, Cameron P Bracken, and Gregory J Goodall. Experimental strategies for microrna target identification. *Nucleic Acids Res*, 39(16):6845–53, Sep 2011.

[61] Jernej Ule, Kirk Jensen, Aldo Mele, and Robert B Darnell. Clip: a method for identifying protein-rna interaction sites in living cells. *Methods*, 37(4):376–86, Dec 2005.

[62] Jernej Ule, Kirk B Jensen, Matteo Ruggiu, Aldo Mele, Aljaz Ule, and Robert B Darnell. Clip identifies nova-regulated rna networks in the brain. *Science*, 302(5648):1212–5, Nov 2003.

[63] K Unsicker, C Meier, K Krieglstein, B M Sartor, and K C Flanders. Expression, localization, and function of transforming growth factor-beta s in embryonic chick spinal cord, hindbrain, and dorsal root ganglia. *J Neurobiol*, 29(2):262–76, Feb 1996.

[64] Shobha Vasudevan, Yingchun Tong, and Joan A Steitz. Switching from repression to activation: micrornas can up-regulate translation. *Science*, 318(5858):1931–4, Dec 2007.

[65] Yangming Wang, Rostislav Medvid, Collin Melton, Rudolf Jaenisch, and Robert Blelloch. Dgcr8 is essential for microrna biogenesis and silencing of embryonic stem cell self-renewal. *Nat Genet*, 39(3):380–5, Mar 2007.

[66] Yanli Wang, Gang Sheng, Stefan Juranek, Thomas Tuschl, and Dinshaw J Patel. Structure of the guide-strand-containing argonaute silencing complex. *Nature*, 456(7219):209–13, Nov 2008.

[67] Hynek Wichterle, Ivo Lieberam, Jeffery A Porter, and Thomas M Jessell. Directed differentiation of embryonic stem cells into motor neurons. *Cell*, 110(3):385–97, Aug 2002.

[68] Ho-Keun Yi, Sun-Young Kim, Pyoung-Han Hwang, Chan-Young Kim, Doo-Hyun Yang, Youngman Oh, and Dae-Yeol Lee. Impact of pten on the expression of insulin-like growth factors (igfs)

and igf-binding proteins in human gastric adenocarcinoma cells. *Biochem Biophys Res Commun*, 330(3):760–7, May 2005.