Title
MicroRNA target site identification in helper T cell differentiation

Author
Kageyama, Robin

Publication Date
2017

Peer reviewed|Thesis/dissertation

MicroRNA target site identification in helper T cell differentiation

by

Robin Kageyama

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Acknowledgements

I would first like to thank my mentor and advisor, Mark Ansel for everything he has done for me during my time at UCSF. Thank you for allowing me to explore outside of my comfort zone and pursue bioinformatics even though I came into your lab knowing nothing. Your mentorship has helped me to find confidence and clarity in both scientific and personal goals. I honestly could not imagine a better mentor and will continue to recommend you to other trainees long after I have left your lab. I would also like to thank the rest of the Ansel lab for creating a wonderful, fun, and supportive environment to work in. I would also like to thank Dirk Baumjohann for his mentorship early on, providing an example to me of what could be accomplished with hard and well thought out work. A special thanks to Adam Litterman for endless discussions about every facet of my project. My productivity tripled when you joined the lab, helping me to clarify thoughts on algorithms and experimental design, while providing constant can-do motivational pushes.

I would like to thank my thesis committee, Prescott Woodruff, Robert Blelloch, and Richard Locksley. You provided incredibly useful personal and professional guidance that will continue to aid me in the future to come. Thank you to the BMS program, especially Lisa Magargal, Monique Piazza and Demian Sainz, for all the essential help you have given me over the past few years. Also thank you to Shane Crotty for taking a chance on me and giving me the opportunity that led me down this path in the first place.

Thank you to all of my friends who have helped me stay sane, provided scientific insight and guidance, and provided inspiring examples of what comes next after grad school. You have all made San Francisco feel like a place where I belong.

Finally, I want to thank all of my wonderful family for all of their support and faith in my success. To my parents Beck and Ben, I literally could not have done any of this without you. Your constant belief that I could do this is what kept me strong. To my sister Maya, you have always been someone I look up to despite being my younger sister and I always hope to make you proud. To my new family, the Simpsons, for being so welcoming and wonderful, I am so grateful to be let into your lives. And to my wife Laura, you are beyond the best thing that has ever happened to me. You are a brilliant and inspiring scientist, and I will forever be awed by your drive to make the world a better place. You are my best friend and role model. I love you and I like you.

Contributions of Co-authors to Presented Work

Chapter 2 of this thesis is as yet unpublished. The authors on this manuscript are Robin Kageyama, Adam Litterman, Misty Montoya and K. Mark Ansel. R.K. designed, performed and analyzed the experiments under the supervision of K.M.A., with A.L. contributing to the design and analysis of some of the experiments. M.M. generated the miR-18 gene expression data. Special thanks to the UCSF Institute for Human Genetics for their help in sequencing.

Abstract

MicroRNAs are important mediators in the control of helper T cell differentiation, where small perturbations in responses to extracellular signals leads to early polarization and specialization. Identifying gene targets of highly expressed helper T cell miRNAs can lead to identification of novel players in these networks. High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) is a powerful tool for identifying these gene targets by pulling down the miRNA binding protein argonaute (Ago) and sequencing the co-immunoprecipitated miRNA and targeted cognate mRNA fragment.

We performed HITS-CLIP on mouse CD4+ helper T cells and identified over X unique miRNA binding sites. While a majority of these sites were in well annotated 3' untranslated regions (UTRs), we also identified a number of miRNA sites in coding regions, downstream of annotated 3'UTRs and in difficult to annotate regions of the genome. We identified hundreds of binding sites that were dependent of the presence of miR-29a, a miRNA highly expressed in CD4 T cells and implicated for its roles in T cell differentiation. We also observed a role for miR-29 in IL-17 production, which led us to identify a number of miR-29 targeted genes with roles in Th17 differentiation, one of which, ICOS, was a new target that we found to be directly regulated by miR-29. Our study identifies Ago dependent miRNA binding sites important in Th17 biology and identifies a role for miR-29 in IL17 production.

Additionally, our lab has generated a tool that aims to create an easy to use interface for labs working with CLIP-Seq data to create high quality graphics that are easily edited in a vector based graphics program. ClipPlot is a webapp that can be run

on a local server and used by all members of a lab simultaneously, requiring only the technical expertise of a single user. While this is not a replacement for sophisticated visualization tools like IGV, ClipPlot provides an easy to use platform for researchers working with CLIP-Seq or RNA-Seq data to quickly create presentable and easily manipulated graphics, visualizing the shape of sequencing data to defined regions.

Table of contents

List of Figures

Chapter 2

Chapter 3

Chapter 1

Introduction

MicroRNAs: biogenesis and function

MicroRNAs are important evolutionarily conserved post-transcriptional regulators of gene expression. Initially transcribed as a non-coding RNA primary transcript (pri-miR), their hairpin sequences are cleaved from the pri-miR by the RNAse containing Drosha and DGCR8 (Y. Lee et al., 2003). A single pri-miR may have multiple hairpins, which can lead to a single pri-miR transcript "cluster" giving rise to more than one mature miRNA. After Drosha/DGCR8 processing, the ~70nt hairpin (pre-miRNA) is exported from the nucleus where the nuclease Dicer removes the hairpin, creating a double stranded mRNA composed of two mature miRNA sequences ~22nt in length (Hutvágner et al., 2001). One of these strands is then bound by the protein Argonaute (Ago), which is associated with the RNA-induced silencing complex (RISC). Either of these strands (referred to as 3p and 5p) may be loaded in to the RISC, however one of the two is usually dominant, with the less used strand commonly referred to as the "star" strand (Khvorova et al., 2003). Ago and the mRNA are targeted to mRNA transcripts with sequence complementarity to the seed sequence of the mature mRNA. The seed sequence consists of nucleotides 2-8, and is the major determinant of miRNA gene targeting, though it is not the only determinant, as will be discussed later on.

Since the first identified miRNA *lin-4* was identified in *C.Elegans* over 20 years ago, we have identified thousands of unique miRNAs, hundreds of which are highly conserved across a multitude of species(Lagos-Quintana et al., 2001; R. C. Lee et al., 1993; Wightman et al., 1993). The target sites of these miRNAs often lie in the 3' untranslated region (UTR) of the regulated mRNAs, and some of the most potent miRNA binding sites are highly conserved (Friedman et al., 2008). However, there are

also many non-conserved miRNA binding sites, as well as non-conserved miRNAs which have also been demonstrated to be able to repress mRNAs (Agarwal et al., 2015). There are also miRNA-mRNA interactions that have been described outside of the 3'UTR including in coding regions, and 5'UTRs (Hafner et al., 2010; Lytle et al., 2007; Ray M Marín, 2013).

MicroRNA target identification

Because miRNA binding is determined by such a small region 7nt long, this can lead to tens of thousands of potential binding sites for any single miRNA. There is also the fact that as few as 6nt of seed complementary can be necessary for miRNA mediated repression of a target transcript, and that 3'UTRs can be kilobases in length. This means that a single miRNA can bind a large number of mRNAs, and that mRNAs can in turn be bound by a number of different miRNAs. The advantage of this, is that a single miRNA can have quite a large effect on a number of different genes simultaneously, creating a large network that can together have a much larger phenotypic effect (Ebert and Sharp, 2012).

However, there are huge challenges to overcome in the study of miRNAs when it comes to identifying these target sites. With a binding site as small as 6nt, one could expect to see a binding site every 4^6 (4096) nt, and with potentially thousands of miRNAs, a 3'UTR can be littered with potential binding sites. Even putting aside the fact that the 3' end of a miRNA can be sufficient for binding, along with other non-canonical

interactions, prediction of miRNA gene targeting quickly becomes overwhelming (Grimson et al., 2007).

There have been a number of quality studies that have aimed to annotate and identify functional miRNA binding sites from many different angles (N D Mendes, 2009). The two major routes either involve *in silico* prediction, or functional identification. These two methods are not diametrically opposed, as the two often complement each other, with *in silico* approaches incorporating more and more of the observations from experimental data sources. One of the more popular computation miRNA target prediction algorithms is Targetscan, which predicts functional miRNA targets by seed sequence, position within the 3'UTR, evolutionary conservation of the target site, sequence context of the surrounding site, accessibility of the mRNA due to secondary structures as well as many other factors (Agarwal et al., 2015). There have also been a number of approaches to identify functional miRNA targets through experimental observations. Potential sites can be identified by prediction algorithms, and these sites can be tested through reporter assays, cloning the predicted site into the 3'UTR of a reporter gene. Differentially gene expression analysis can also be used as a tool to identify miRNA regulated genes after manipulating expression levels of individual miRNAs. This is a useful in conjunction with target prediction, however this can be challenging due to the moderate effects of miRNAs on gene expression.

In recent years, new techniques have emerged to approach target site identification by crosslinking immunoprecipitation of the Ago protein and the bound mRNA fragment and miRNA (Jaskiewicz et al., 2012; Zhang and Darnell, 2011). These tools have been incredibly useful at identifying sites of Ago interaction, suggesting

miRNA binding at these sites. Crosslinking immunoprecipitation (CLIP) involves crosslinking RNA binding proteins to their bound RNA, allowing pulldown of the RNA for further analysis (Ule et al., 2005). One of these techniques, high throughput sequencing and crosslinking immunoprecipitation (HITS-CLIP) involves high throughput sequencing of the RNA, aligning it to the genome and revealing read-dense genomic regions or "peaks" that are bound by protein (Licatalosi et al., 2008). HITS-CLIP of Ago is able to reveal potential miRNA binding sites, but it can't be known exactly which miRNA is binding to any given peak, or that the Ago binding is dependent on miRNAs in the first place. To get around this problem, one group used miR-155-/- T cells to try to map miR-155 dependent Ago binding sites (Loeb et al., 2012). Differential HITS-CLIP (dCLIP) has been with miRNA knockouts to identify Ago peaks that disappear in the absence of a single miRNA (Bracken et al., 2014; Loeb et al., 2012).

There are a few caveats regarding HITS-CLIP that need to be addressed. First of all, there are many non-canonical binding sites that are mapped by the dCLIP datasets that do not contain seed sequences(Licatalosi et al., 2008; Loeb et al., 2012). There is the question if these sites are actively regulating mRNA expression levels, if they are secondary effects of a miRNA deletion, or spurious false positives. It is controversial to what extent these non-canonical sites are capable of inducing post-transcriptional silencing or degradation of mRNA targets (Agarwal et al., 2015; Loeb et al., 2012). Additionally, a major limitation of the technique is that only genes that are expressed will show miRNA-mRNA interaction sites, taking some of the air out of the sails of the claim of global mapping. This limitation is also one of the greatest strengths of HITS-CLIP. Instead of providing a schematic of all possible targets of a miRNA, instead, a view of

miRNA regulation is seen in a snapshot moment. HITS-CLIP isn't an experiment to be done once, but a technique to probe deeper into the role of a miRNAs in a specific instance of a cellular state, a tissue response to a pathogen, or in fate decisions of differentiation.

MiRNAs in helper T cell differentiation and function

The immune system is a complex and delicately maintained system consisting of innate defense mechanisms and adaptive systems with antigenic memory. These systems are controlled by the expression of tightly regulated soluble cytokines which create an environment that orchestrates a coordinated response to a large variety of pathogenic insults.

CD4+ helper T cells respond to these environmental signals, either from circulating cytokines or cognate interactions with antigen presenting cells that provide co-stimulation through the T cell receptor. In response, CD4 T cells undergo rapid proliferation and differentiation. These differentiation lineages differ depending on the response needed, and include T helper 1 (Th1), T helper 2 (Th2), T helper 17 (Th17) and regulatory T (Treg) cells(Zhu et al., 2010). Differentiation into one of these subsets requires specific cytokine signals, which induce a specific transcriptional program, in part through expression of lineage specific transcription factors. These helper T cell subsets secrete unique cytokine profiles which not only aid in a coordinated immune response, but reinforce T cell differentiation in positive feedback loops.

This thesis in particular focus on the miRNA-29 family, a highly conserved miRNA expressed in CD4 T cells (Kuchen et al., 2010). There are 3 mature variants of

miR-29: miR-29a, miR-29b and miR-29c. There are two miR-29 clusters in the mouse and human genomes, miR-29ab1 which contains miR-29a and miR-29b-1, and the second cluster miR29b2c, which contains miR-29b-2 and miR-29c. The mature 3p strand miRNAs are identical, but the pre-miRNAs differ in sequence as do the 5p star strands. There have been numerous studies implicating miR-29 in cancer (Y. Wang et al., 2013) and fibrotic disease (He et al., 2013). The role of miR-29 in T cells however is still under investigation. Complete knockouts of miR-29 have shown that T cell maturation seems to be unimpaired in a cell intrinsic fashion, however loss of miR-29 leads to a loss of repression of Ifnar1 in thymic epithelial cells. This leads to early thymic involution and cell extrinsic inhibition of thymic cellularity(Papadopoulou et al., 2012). Additionally, previous studies have shown a role for miR-29 in directing T cell differentiation to Th1 cells. DGCR8-/- CD4 T cells lacking most miRNAs, demonstrated a strong Th1 differentiation phenotype. The addition of miR-29 back into the cells, was able to partially restore wildtype levels of Th1 polarization (Steiner et al., 2011). A second study was able to show increases in Th1 cell differentiation after transgenic expression of a miR-29 sponge in CD4 T cells (Ma et al., 2011). These studies identified the Th1 transcription factors T-bet and Eomes (Steiner et al., 2011) as well as the cytokine IFNg (Ma et al., 2011) as direct targets of miR-29. It has also been suggested that IFNg induces miR-29 expression, creating a negative feedback loop (Schmitt and Philippidou, 2012; Smith et al., 2012). These effects on transcription factors demonstrate one way in which moderate effects of a miRNA on a target can have much larger downstream effects. This may also be the case with epigenetic regulators as miR-29 has been shown to be highly downregulated in T-cell acute lymphoblastic

leukemia (T-LL). One study demonstrated that restoration of miR-29 led to increases in demethylation of T-ALL associated genes, which was suggested to be through the miR-29 targeted DMNT3 DNA methyltransferase and TET family members(Oliveira et al., 2015). This is one of a few instances describing a role for miR-29 in T cells in the context of disease, which include IL-21 dependent miR-29 inhibition of HIV infection, and dysregulated miR-29 expression in T cells from MS patients (Adoro et al., 2015; Smith et al., 2012).

Overview of Thesis

Gaining a greater understanding of the targets of miR-29 during CD4 T cell activation can provide not just therapeutic targets, but a greater understanding of the pathways that manipulation of miR-29 perturbs. This is true for all expressed miRNAs, and discovering what those targets are, when they are bound, and when they are having an effect can provide insight into the deep regulatory web of miRNA interactions, the nodes or genes that that web touches, and in turn, potentially reveal new elements and structures of those pathways. With so many potential targets, how can we hope to cut through the noise and identify miRNA targets that are important to a phenotypic question? How do we search for miRNA targets beyond the 3'UTR, when scanning the genome could mean hundreds of thousands of potential hits? Chapter 2 of this thesis discusses our effort to identify miRNA binding sites using HITS-CLIP, with a particular focus on miR-29. We identify miRNA binding sites capable of regulating gene expression on a global scale, including sites outside of 3'UTRs. Combining HITS-CLIP peaks with prediction algorithms and seed sequence complementarity, we identify gene

targets that enrich for genes differentially expressed in the presence or absence of individual miRNAs.

In our characterization of miR-29 deficient CD4 T cells, we found that the miR-29 deficient T cells overexpressed IL-17 as compared to wildtype controls. To identify genes that may be contributing to this phenotype, we performed HITS-CLIP on both wildtype and miR-29 deficient T cells, and identified peaks with differential expression between the two groups. We identified a number of genes with a differentially expressed Ago binding peak that have known roles in Th17 differentiation and IL-17 expression including the co-stimulatory surface molecule ICOS (Hutloff et al., 1999). We were able to show that the identified Ago binding peak in ICOS was sufficient to induce RNA instability when cloned into the GFP 3'UTR, and that this instability was dependent on the presence of miR-29.

Chapter 3 discusses a program I developed called ClipPlot. This program was created as a visualization tool for the generation of Clip-seq and HITS-CLIP data. Creation of this tool came out of a need to generate figure or presentation quality images of HITS-CLIP data, that was both customizable, created easily editable images, and was accessible to a lay user. ClipPlot can be set up by a single user in the lab on a local server, and accessed by others in the lab over the network through an intuitive web interface. Users can create vector images, that can be edited easily to suit ones needs. The program was generated in python and uses scipy, Jupyter and iPython to create the GUI and graphics. ClipPlot also contains a number of my custom python modules that can be used to aid in analysis of HITS-CLIP data.

I started working in the Ansel lab working on miR-17~92 and the cluster's role in the function of follicular helper T cells (Tfh). Appendix 1 contains my first publication with the Ansel lab, working with Dirk Baumnjohann, where we found that the miR-17~92 cluster was essential for the differentiation of Tfh cells, due in part to inhibition of *Rora*. My work in the project included contribution to a number of T cell cultures, adoptive transfers, flow cytometry analysis and single cell sorting. Major contributions are seen in figure 5, where I performed and analyzed microarray data of Tfh cells deficient in miR-17~92, and in figure 6, where I created luciferase reporter constructs. Appendix 2 contains another publication I contributed towards in the lab that uses Global Cross-linking Protein Purification (GCLiPP) to globally identify RNA binding protein sites, and found patterns of rapid evolution, high GC content and greater folding likelihood among highly bound regions. I contributed to the data analysis, software pipeline construction, and biochemical assay development.

Chapter 2

High throughput identification of microRNA binding sites in CD4 T cell differentiation

Abstract

Background: MicroRNAs are important mediators in the control of helper T cell differentiation, where small perturbations in responses to extracellular signals leads to early polarization and specialization. Identifying gene targets of highly expressed helper T cell miRNAs can lead to identification of novel players in these networks. High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) is a powerful tool for identifying these gene targets by pulling down the miRNA binding protein argonaute (Ago) and sequencing the co-immunoprecipitated miRNA and targeted cognate mRNA fragment.

Results: We performed HITS-CLIP on mouse Cd4+ helper T cells and identified miRNA binding sites. While a majority of these sites were in well annotated 3' untranslated regions (UTRs), we also identified a number of miRNA sites in coding regions, downstream of annotated 3'UTRs and in difficult to annotate regions of the genome. We identified hundreds of binding sites that were dependent of the presence of miR-29a, a miRNA highly expressed in CD4 T cells and implicated for its roles in T cell differentiation. We also observed a role for miR-29 in IL-17 production, which led us to identify a number of miR-29 targeted genes with roles in Th17 differentiation, one of which, ICOS, was a new target that we found to be directly regulated by miR-29.

Conclusions: Our study identifies Ago dependent miRNA binding sites important in Th17 biology and identifies a role for miR-29 in IL17 production.

Introduction

Helper T cell development and function are complex processes that require proper expression of transcription factors and effector cytokines in specific amounts at specific times (Zhu et al., 2010). Dysfunction in these processes can lead to allergy, autoimmunity and chronic inflammatory disease (Zhu and Paul, 2008). MicroRNAs (miRNAs) are small non-coding RNAs that post-transcriptionally inhibit messenger RNAs (mRNA) through interaction with Argonaute (AGO) proteins in the RISC complex (Ansel, 2013) . While it is clear that miRNAs are important mediators of helper T cell differentiation, the effect of a single miRNA-mRNA target interaction can have moderate effects.

While miRNAs may have relatively modest effects on a single gene transcript, by targeting hundreds of genes, whole networks can be manipulated by a single miRNA. Uncovering these networks can lead us not only to a greater understanding of the role of miRNAs in cellular function, but also reveal new unappreciated gene functions. Determining the targets of miRNAs can be done on a gene by gene basis, however transcriptome wide scale mapping of miRNA targeting has largely been limited to a computational *in silico* approaches. A major determinant of miRNA mRNA interactivity is determined by the seed region at positions 2-7 of the 5' end of a mature miRNA. To move beyond the limitations of target determination by sequence alone, groups have used a technique called high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) (Chi et al., 2009; Loeb et al., 2012). This technique sequences RNA crosslinked to immunoprecipitated AGO to identify miRNAs and the mRNA fragments they are bound to. This provides more than an unbiased approach to

identifying regions of AGO binding, it provides information about targets that are likely to be actively regulated in the context of whatever system the assay is being performed in. It is well known that miRNAs have a large effect controlling helper T cell differentiation. Teasing apart the specific effects of miRNAs in this process has been a challenge for the field.

To understand the role of miRNAs in T cells, we have to address the problem with both molecular and phenotypic strategies, finding specific and relevant gene targets as well as studying their dysregulation in disease.

In this study, we perform transcriptome wide Ago HITS-CLIP on differentiating CD4 T cells. Combining our analysis of Ago binding sites with computationally derived predicted miRNA target sites, we were able to identify 289 miR-29 target sites. We found many of these targets to be in 3'UTRs, however we were also able to identify miRNA binding sites in coding regions as well as poorly annotated regions of the genome. We focus on miR-29, a highly expressed miRNA in CD4 T cells that been described to regulate CD4 T cell differentiation, including Th1 differentiation (Papadopoulou et al., 2012; Smith et al., 2012; Steiner et al., 2011). We found that mice deficient in miR-29 revealed a suppressive role for mIR-29 in IL-17 production. In combination with our HITS-CLIP target sites, we were able to identify miR-29 targets connected with IL17 production and TH17 differentiation, including ICOS, a previously undescribed target of miR-29.

Results

Argonaute HITS-CLIP identifies miRNA targets actively bound during T cell differentiation.

To identify argonaute dependent miRNA binding sites in CD4 T cells, we performed Ago-HITS-CLIP on mouse CD4+ T cells isolated from spleen and lymph nodes, activated for 3 days with αCD3/αCD28 followed by 2 days rest in IL2. We modified the published HITS-CLIP protocol to increase efficiency of pulldown and recovery of material (Fig 1A, Methods)(Loeb et al., 2012). Cells were UV irradiated to crosslink mRNA and miRNAs to RNA binding proteins, followed by cell lysis, single stranded RNAse digestion and argonaute immunoprecipitation. Libraries were generated independently, sequenced and aligned to the mouse genome. We used 10 independent biological replicates to generate our datasets and to identify peaks of argonaute binding.

We would expect close to a 1:1 ratio of miRNA:mRNA molecules to be bound to each pulled down argonaute molecule, and we found 55% of aligned reads mapped to miRNAs (data not shown). We were able to detect over 600 unique miRNAs as annotated by miRBase. After subtracting annotated miRNA reads from our dataset, ~50% of aligned reads mapped to 3'UTRs (Fig 1B). Many reads were also found closely downstream or upstream of the 5' and 3' untranslated regions (UTRs) as well as coding regions and in introns.

To validate the quality of our data, we started by looking at binding peaks within highly expressed genes, and particularly those that have been previously described to have important miRNA binding sites within annotated 3'UTRs. We had previously

described that miR-27 was a potent inhibitor of GATA3 and Th2 differentiation (Pua et al., 2016). We were able to observe robust Ago binding at the described miR-27 binding site in the 3'UTR of *Gata3*, as well as a number of other peaks at predicted miRNA binding sites (Fig 1C). We could also see a strong peak at a previously described mIR-19 target site in the *Socs1* 3'UTR (Fig 1D)(Simpson et al., 2014). We also observed peaks at sites without predicted miRNA target sites, some of which had seed complementarity to highly expressed miRNAs, and others which did not.

Argonaute binding sites improve target predictions of gene regulation

Our goal is the identification of the network of gene targets that any given miRNA regulates to discover important players in a phenotype. There are a number of resources available that have used predictive algorithms to try to identify likely miRNA binding sites within 3' UTR regions, including Targetscan (Agarwal et al., 2015). While these algorithms are very useful in the identification of where miRNAs can bind, we wanted to narrow down our lists to find genes that are actually bound by a miRNA of interest, within the context of T cell differentiation. First we needed to see how well our CLIP identified Ago binding regions compared to Targetscan predictions. We used custom peak calling scripts to identify 50nt regions closely matching a standard distribution curve (T. Wang et al., 2014). Due to the sheer number of potential miRNA binding sites in any given sequence when we take into account all of thousands of reported miRNAs, it is impossible to determine the miRNA likely bound with Ago at any given site. We can however, use information about the profiles of miRNAs bound to Ago

that were sequenced along with the mRNA reads. We decided to limit our analysis to only the top 50 miRNA families (determined by seed sequence, Fig 2b).

Another tool we have at our disposal is a number of different gene expression data sets from miRNA deficient mice. We referred back to data from published DGCR8-/- CD4+ T cells transfected with mimic control or miR-29a mimic (Steiner et al., 2011). Selecting for genes with predicted miR-29 target sequences in their 3'UTR reveals a predicable enrichment for genes that have been downregulated in the presence of the miR-29 mimic as opposed to the control. While this list enriched for genes that could be bound by miR-29 (809 genes), when we narrowed down the list further to genes that had an ago binding peak at the predicted site (183 genes) we saw a much greater enrichment for genes downregulated by miR29 mimic (Fig 2C). Negative control miRNA target gene lists showed no change compared to the list of all miRNA targeted genes (Supplemental figure 2a). We observed similar patterns for other miRNAs. When comparing miR18-/- CD4 T cells to wildtype controls, we found that genes with miR-18 target site predictions enriched for genes upregulated in the knockouts and that this enrichment was improved by selecting for target sites with Ago binding peaks in the HITS CLIP data (Figure 2D, Montoya 2017). This pattern also held true for miR-19 target containing genes in miR-17~92 deficient CD4 T cells compared to wildtype controls (Fig 2E) (Baumjohann et al., 2013). This demonstrates that we can use our binding peaks in coordination with prediction algorithms to narrow down our list of genes that are functionally regulated by miRNAs within the context of a system of interest.

Ago HITS-CLIP reveals easily missed context specific binding sites outside of annotated 3'UTRs

One of the major limitations of relying on prediction algorithms, is that we are limited by the annotations for transcript locations. CLIP data from a cell type or even different stimulation conditions can be limited by the genes that are expressed, or isoforms that may drastically change the length or sequence of the 3'UTR. For example, many genes have 3'UTRs that go beyond Refseq annotations, such as Ago2 (Fig 3a). There are a large number of Ago binding sites downstream of the annotated 3'UTR, many of which have seed complementarity to highly expressed miRNAs (Fig 3b) and there are other regions of dense Ago binding peaks that are poorly annotated and are easily missed when limiting analysis to well defined 3'UTRs. This is especially true for regions that are tremendously difficult to annotate such as the TCR. We see one of these dense ago peak regions in TCRb, a 5kb stretch that is rich for Ago binding and has predicted seed sequences (Fig 3E).

Some attention has been payed to regions outside the 3'UTR including the 5'UTR and coding sequences. These areas have been shown to have limited regulatory activity even in the presence of Ago binding. We saw a number of sites in coding regions that had seed complementarity (Figure 3B). We also found that identifying genes with these sites was not devoid of predictive ability for gene regulation, as genes with miR-29 seed sequences in coding regions were more downregulated in the gene expression array after miR-29 mimic transfection (Figure 3C).

A recent study identified a novel miRNA binding site for IL4 that had been missed by Targetscan as well as previously published HITS-CLIP datasets. They found a miR-

24 binding site at the 5' edge of the 3'UTR that was a potent regulator of transcript stability. It was missed by predictions due to the overlap of the mature miRNA with the coding sequence. It was also not a clear argonaute binding site in available HITS CLIP experiments, most likely because of the lack of expression of IL4 in those datasets. Our Th2 cells had detectable levels of IL4, allowing us to visualize an Ago peak at the miR-24 site (Fig 3D).

These examples illustrate the importance and utility of using Ago-CLIP in cells specific to the system being studied. Gene expression difference, alternative splicing and variable regions can differ from system to system, and having matched ago binding can greatly improve upon the picture of the influence of miRNAs.

Comparative HITS-CLIP with miR29 deficient CD4 T cells reveals miR29 dependent AGO binding sites

While a major advantage of AGO-CLIP is the ability to infer miRNA binding, one limitation is the inability to determine which miRNA is responsible for Ago targeting to an individual peak. While we can use seed sequence to infer the likelihood of miRNA binding, this doesn't allow the discovery of noncanonical sites, or the ability to distinguish between tightly packed predicted miRNA sites. To get around these limitations, we turned to differential HITS-CLIP. Our lab has been interested in the miRNA miR-29, due to its high expression in CD4 T cells, as well as multiple known functions for the miRNA in the function of CD4 T cell subsets. We were especially interested to identify miR-29 binding sites that were actively bound by miR-29 during CD4 T cell differentiation.

MiR-29 knockout mice develop severe defects in their T cell compartment due to thymic involution. To get around this problem, we generated miR-29 flx/flx mice that were bred to CD4-Cre, to obtain T cell specific miR-29 deficient cells. 6 independent HITS CLIP experiments with a total of 20 samples were performed for both WT and KO CD4 T cells after 5 days of in vitro stimulation. Differentially expressed peaks between both WT and KO samples were identified using the published dCLIP analysis pipeline toolkit on the combined datasets (Fig 4A and B) (T. Wang et al., 2014).

If we separated out genes that contained differentially expressed peaks, we saw that those peaks that overlapped with 8mer complementary target sites were significantly enriched for genes downregulated in DGCR8-/- cells transfected with miR-29 compared to control. This was true as well for genes containing any type of canonical miR-29 site (8mer, 7merm8,7mer1a,6mer) (Fig 4C). However, when we looked at non-canonical sites, we saw that this difference was greatly reduced, suggesting that non-canonical miR-29 dependent argonaute binding sites, may not contribute as strongly to downregulation of their target transcripts (data not shown).

miR-29ab-1 conditional knockout CD4 T cells produce higher levels of IL17 after simulation

Because we were describing a new miRNA knockout, and using that system to define novel miR29 binding sites, we wanted to see if there was a phenotype in these mice that would help inform which gene targets we might be interested in. It has been previously described that miR29 is important for Th1 differentiation in the absence of other miRNAs, and in the total miR29 knockout mice, there are demonstrated

reductions in Th1 cells and Th17 cells. We found that consistent with the total knockouts, the miR29 conditional knockout CD4 T cells produced significantly less IL17 in vitro after polarization to IL17 cells in vitro (Fig 5A and B). This reduction in IL17 was not limited to Th17 polarized cells. If we polarized cells under different conditions, Th1, Th2 or iTreg, we still were able to see reduction in IL17 production, though often expression was very low in the non Th17 conditions. We also saw the expected increase in IFNg in the miR29 knockouts under non-polarizing conditions, however we no difference could be seen in the Th1 conditions (Fig 5C). This could be due to the strength of the Th1 polarization conditions, and the relatively modest effect of miRNAs.

We also noticed that there were no significant differences in the amount of RORgt expression between the wildtype and knockout T cells (Fig 5D). This was surprising due to the increase in IL17 observed, but this, along with the fact that Il17 was up in the KO under multiple polarizing conditions, suggest that miR29 may dampen IL17 expression independently of RORgt.

To further investigate the role of miR-29 on IL-17 production, we referred to our list of ago binding peaks with miR-29 seed sequences that were significantly less expressed in the mIR-29 deficient CD4s. We were in fact able to see multiple genes that could be contributing the observed phenotype. One of the clearest hits was in the 3'UTR of ICOS where we saw a highly bound ago binding peak that was reduced 5 fold in the absence of miR-29 (Fig 6A). ICOS has been suggested to be important for TH17 differentiation through regulation of cMaf and subsequently IL23R expression. We also saw a highly differentially expressed peak in the 3'UTRs of Cflar and Sat1b, both of which have been previously described to have a role in IL17 regulation (Fig 6B and C).

ICOS and Cflar are useful examples of how we can use target prediction in harmony with differential peak binding, but there may also be functional targets without canonical sites, that can warrant further study to determine functional relevance.

ICOS is directly regulated by miR-29

While we are interested in many of the genes regulated by miR29 to inhibit IL17, we decided to start with a single example to start to piece together the network of genes. We first turned our attention to ICOS, which has yet to be described as a target of miR-29, but which has a clear and differentially expressed peak with an 8mer seed in its 3'UTR. We wanted to see if we could identify whether ICOS was directly regulated by miR-29. We cloned the predicted miR-29 binding site into the 3'UTR of GFP in a GFP expression plasmid. To test the stability of the RNA in the presence of absence of miR-29, we in vitro transcribed GFP from the plasmid and transfected the RNA directly into stimulated miR29 $^{flx/flx}$ CD4-Cre^{+} or wildtype control CD4 T cells. We used a sequence cloned from the identified miR-29 dependent Ago binding ICOS site, as well as a target site that had perfect complementarity to the mature miR-29. We also included a scrambled version of the ICOS site as a scrambled negative control. All 3 RNA templates were combined and transfected together to control for differences in transfection efficiency and RNA extraction. After transfection with the RNA, we took cells at multiple timepoints and purified out the RNA. After reverse transcription, we performed qPCR specific for our transfected mRNAs. After normalizing to the scrambled sequence at each timepoint, we saw that there was a clear reduction in mRNA containing either the ICOS site, or the miR29 complementary site very early on at 30

minutes after transfection in the WT cells, but not in the miR29 knockout cells. This shows that the stability of the GFP RNA containing the ICOS miR-29 binding site was dependent of the expression of miR29.

Discussion

MicroRNAs continue to be an extraordinary resource for the mapping and discovery of new pathways and interactions, as well as providing a means for subtle manipulation of these systems. Hundreds of thousands of miRNA-mRNA interactions have been described, predicted and studied with a wide variance of detail. The amount of information is very nearly overwhelming and much like the sequencing of the human genome, this information does not give us immediate answers. It is important that we begin to narrow down this information with careful consideration given to what system or cell type we wish to survey for miRNA mRNA interactions.

HITS-CLIP is a powerful tool for identifying miRNA binding sites actively being bound by argonaute with a specific context (Zhang and Darnell, 2011). Indeed, a number of other studies have started to do similar work in different contexts. Initial HITS-CLIP studies were conducted on mouse brain tissue, focusing on miR-124 (Boudreau et al., 2014). Follow up studies included the use of miRNA knockout mice, including a study that used activated mouse CD4 T cells to identify hundreds of non-canonical miR-155 dependent Ago binding sites (Loeb et al., 2012). Other studies have focused on cardiac tissue and embryonic stem cells, as well as on schistosomes and viral miRNAs (Spengler et al., 2016; Yang Eric Guo, 2015; Zhao et al., 2015).

In this study, we are looking specifically at miRNA target interactions within the context of differentiating CD4 T cells. Our data adds to the growing accumulation of miRNA dependent Ago binding sites.

We also demonstrate a continued need to look beyond the 3'UTR of mRNA transcripts when considering important sites for miRNA interactions. There have been a

number of improvements considering miRNA sites beyond the 3'UTR. Targetscan used to consider only annotated transcripts, but recently was updated to include 3pSeq data to extended 3'UTR sequence and allow for rare but potentially meaningful isoforms to be considered when searching for miRNA targets. Numerous Ago-Clip-Seq studies have identified Ago binding sites outside the 3'UTR, including the 5'UTR and coding regions, however searchable databases from these studies rarely include these regions. Our datasets do not limit analysis to 3'UTRs, but instead to regions of minimum expression.

We can combine a generic search for miRNA binding sites using only a seed sequence as our guide, Targetscan predictions using their sophisticated and robust models, with our HITS-CLIP binding data. This allows us to do two things. First, we can identify novel binding sites outside of regions commonly surveyed for seed sequences. Second, we can pick out miRNA binding sites that have already been predicted to be excellent target candidates but also are actively bound by Ago in our specific system. We have shown that combining predictions with actively bound sites can greatly enrich for targets that are regulated in the presence or absence of a given miRNA.

Our study contains a dataset that goes beyond CD4 T cell activation. We further look at Th17 cells differentiated in vitro, as well as the role of miR-29 in repression of IL-17 expression. MiR-29 is a highly conserved miRNA family highly expressed in CD4 T cells, and has been associated with acute myeloid leukemia, fibrosis, and Th1 differentiation (Liston et al., 2012). Previous studies focusing on the role of miR-29 in T cells have used a knockout mouse, which is complicated by *Ifnar* dependent thymic involution, leading to decreased thymic cellularity (Papadopoulou et al., 2012). To get

around this, we generated T cell specific miR-29ab1 floxed mice. This allowed us to identify increased expression of IL-17 by stimulated CD4 T cells and ask more specific questions with our HITS-CLIP and utilize a strength of the technique. Performing HITS-CLIP on Th17 polarized wildtype and miR-29 deficient CD4 T cells allowed us to identify a number of Th17 specific gene targets of miR-29 that only showed up in the context of Th17 cells because expression is required for HITS-CLIP capture. This revealed that miR-29 hits different pathways at once. While T-bet may be an important miR-29 target in Th1 differentiation, targets like ICOS can have effects on Th17 differentiation. It is unsurprising that such a highly expressed miRNA can have effects on a number of different pathways, but our experiments provide the basis for developing schematics of miRNA interactions that change based on the state of the cell.

Figure 1: Argonaute HITS-CLIP identifies miRNA target sites actively bound during T cell differentiation

A) In vitro activated CD4 T cells were crosslinked by UV radiation. Ago2 was immunoprecipiated from cell lysates and treated with RNAse. After linker ligation and gel electrophoresis size selection for argonaute, co-precipitated mRNAs and miRNAs were extracted for library prep and sequencing. B) Distribution of reads aligning to annotated gene features after miRNA reads were removed based on miRBase annotations. C-D) Example of the HITS-CLIP RNAseq. Reads aligning to the GATA3 3'UTR and the Socs1 3'UTR. Tick marks represent predicted miRNA target sites.

Argonaute HITS-CLIP identifies miRNA target sites actively bound during T cell differentiation

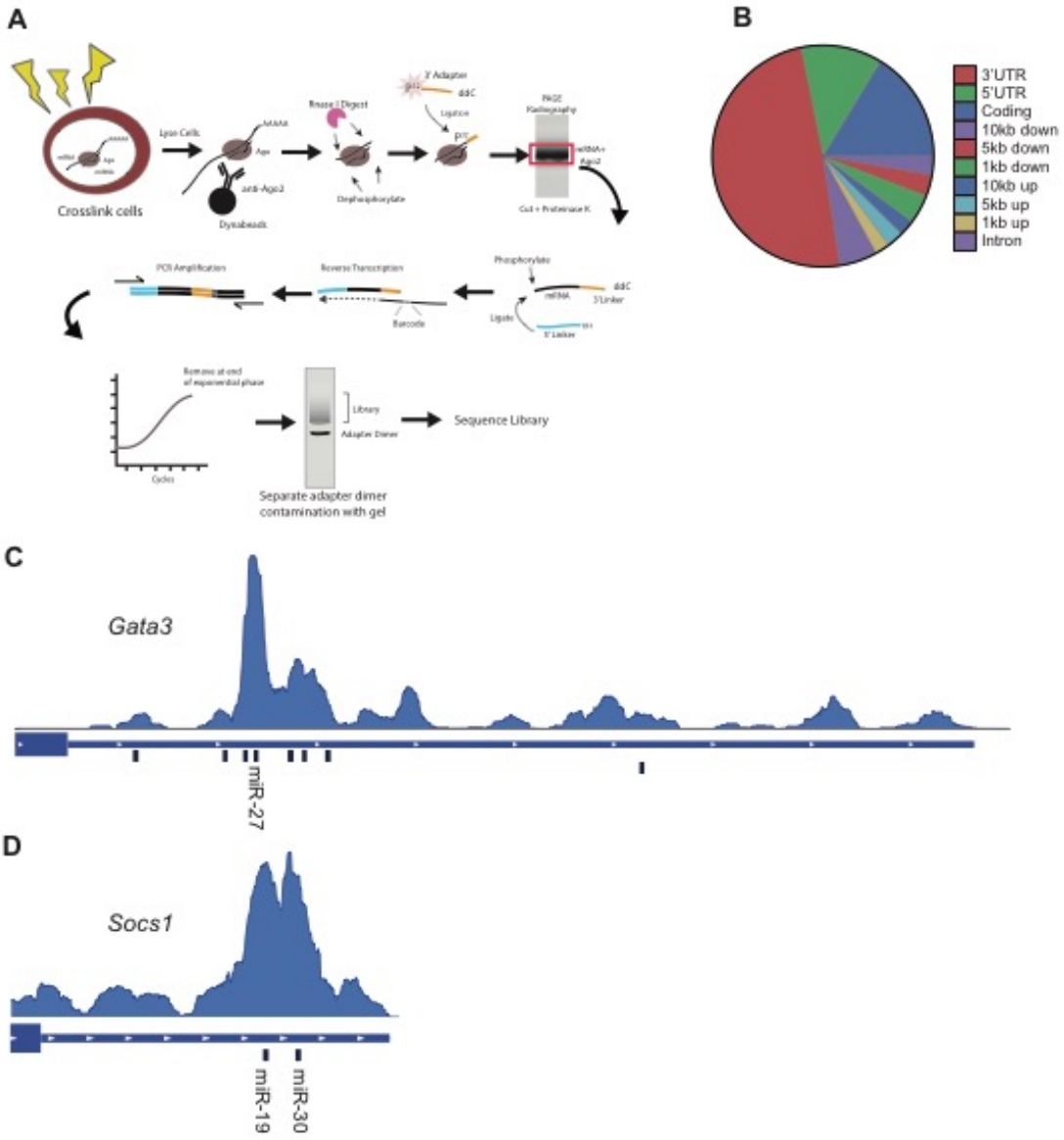
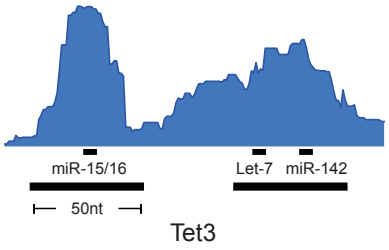


Figure 2: Ago HITS-CLIP predicts miRNA regulated target genes in synergy with target prediction algorithms

A) Example of 50nt peak called in 3'UTR of Tet3 with miRNA seed matches highlighted. B) Top miRNA families sequenced by seed sequence and percent of total miRNA reads. Family names have been shorthanded. C-E) Cumulative distribution of gene expression after transfection of DGCR8-/- CD4 T cells with miR29 mimic vs cells transfected with control mimic (C), RNAseq of miR18a-/- CD4 T cells vs wildtype controls (D) or RNAseq of miR17~92-/- sorted follicular helper T cells compared to wildtype control (E). Black lines represent the set of genes that have at least one predicted miRNA binding site in their 3'UTR. Blue lines represent the set of genes containing Targetscan predicted miR29, miR18, and miR19 sites respectively. Red lines represent genes containing both the predicted miRNA site, but also contain a peak at the same site in the HITS-CLIP data. Pvalues were calculated with a two-sided KS test compared to the set of all targeted genes.

Ago HITS-CLIP predicts miRNA regulated target genes in synergy with target prediction algorithms

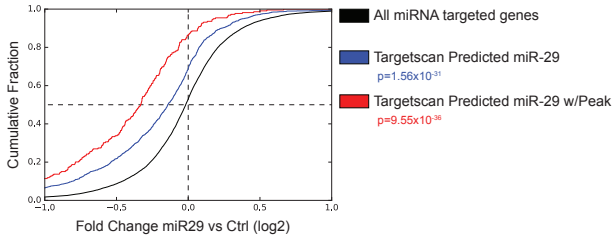
A



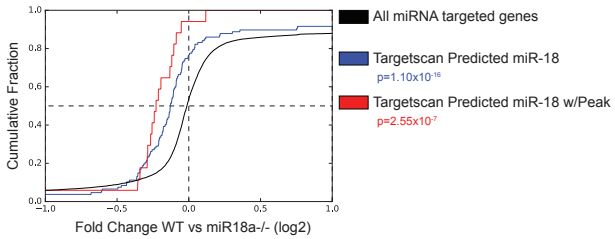
B

seed	miRNA Family	Percent
ACACTAC	miR-142-3p	34.21%
ATAAGCT	miR-21	23.93%
TGCTGCT	miR-15/16/195	9.97%
CTACCTC	let-7	9.02%
TGTTTAC	miR-30	4.25%
GCACTTT	miR-17/20/93/106-5p	2.13%
TTGGGAG	miR-150	1.68%
TGGTGCT	miR-29	1.67%
ACTGTGA	miR-27	1.30%
TACTTGA	miR-26	1.25%
ACTTTAT	miR-142-5p	1.14%
AATGTGA	miR-23	0.86%
TTTGCAC	miR-19	0.69%
TTGCACT	miR-130/301	0.46%
AGCATTA	miR-155	0.44%
GTGCAAT	miR-25/32/92	0.43%
TTCCGTT	miR-191	0.39%
GTGTGAG	miR-342	0.34%
ACTGCAG	miR-17/20/93/106-3p	0.31%
ATTCTTT	miR-186	0.30%
GTCTTCC	miR-7	0.29%
TCTTGCC	miR-31	0.27%
GGGATGC	miR-324	0.27%
CTGAGCC	miR-24	0.23%
CACTGCC	miR-34	0.23%
TGAATGT	miR-181	0.22%
TACTGTA	miR-101	0.22%
GCACCTT	miR-18	0.20%
GTGTCAT	miR-425/489	0.16%
ACTGTAG	miR-139	0.16%
CAGTGCG	miR-106	0.14%

C



D



E

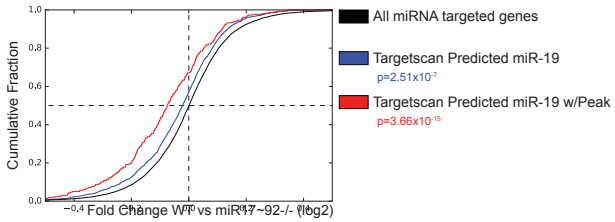


Figure 3: Ago HITS-CLIP reveals easily missed context specific binding sites outside of annotated 3'UTRs.

Examples of HITS-CLIP peaks identified outside of annotated 3'UTRS.

A) Ago binding peaks downstream of the annotated Ago2 3'UTR. B) Ago binding peak in exon 15 of smarca5 with a miR-29 8mer target site. C) Tcrb 5kb region in chromosome 6 with Ago binding peaks. D) CDF plot showing the set of genes with a called binding peak in a coding region that also contain a seed match to miR-29. DE gene expression is the same as in Fig 2C E) Peak in IL4 with a miR24 binding site overlapping the 3'UTR and the end of the coding region.

Ago HITS-CLIP reveals easily missed context specific binding sites outside of annotated 3'UTRs

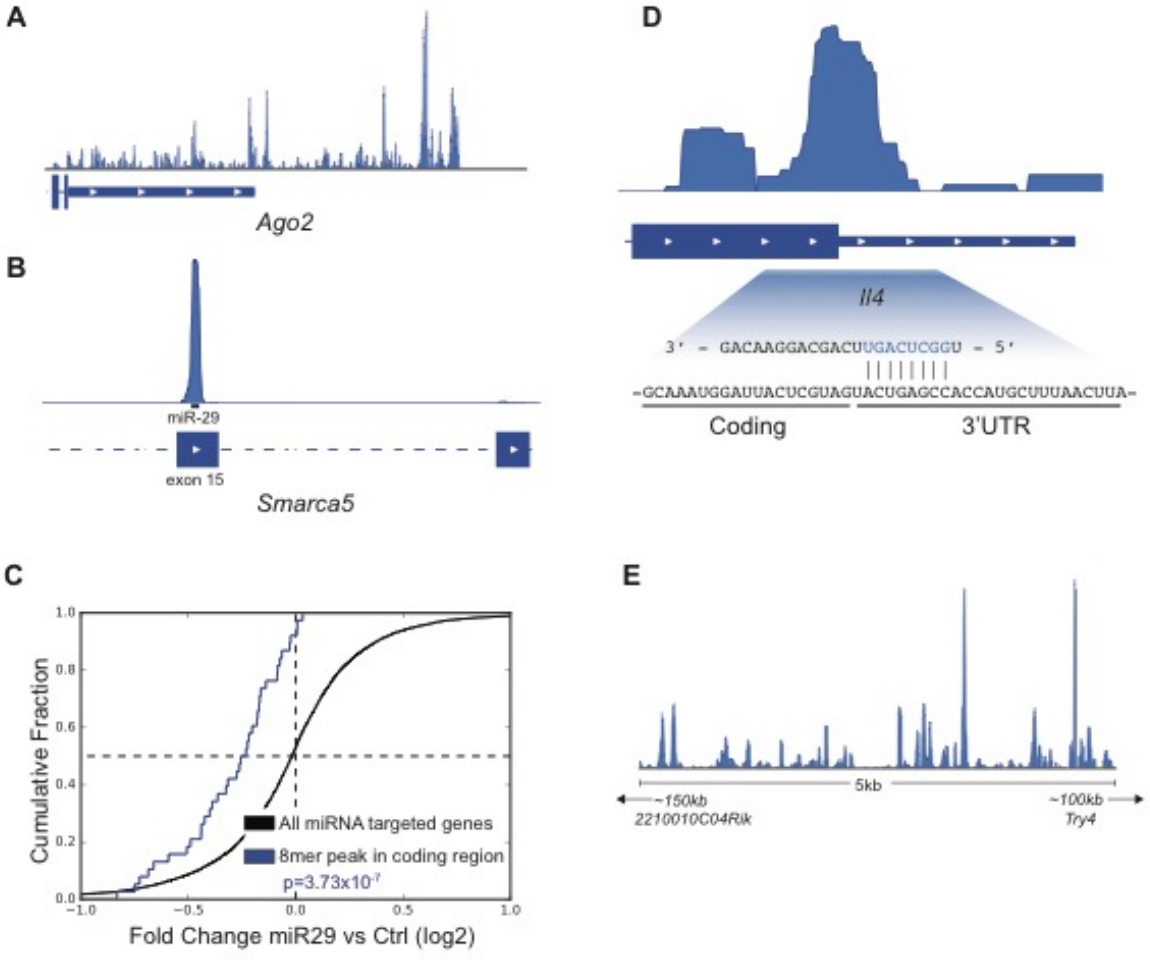


Figure 4: Comparative HITS-CLIP with miR29 deficient CD4 T cells reveals miR29 dependent AGO binding sites

(A-B) Examples of miR-29 dependent AGO binding peaks in 3'UTRs. (A) 7mer-1a miR-29 seed site in *Zfp36l1* from miR-29ab$^{flx/flx}$ CD4-Cre^{-}(blue) and miR-29ab$^{flx/flx}$ CD4-Cre^{+} (red). (B) 8mer miR-29 seed site in the *Tet3* 3'UTR.

Cumulative distribution of gene expression after transfection of DGCR8-/- CD4 T cells with miR29 mimic vs cells transfected with control mimic. The lines represent the set of genes that have at least one predicted miRNA binding site in their 3'UTR(black), the set of genes containing differentially expressed peaks downregulated in the absence of miR-29 either with no seed sequence (red), with a seed sequence (green) or with an 8mer sequence (blue). Pvalues were calculated with a two-sided KS test compared to the set of all targeted genes.

Comparative HITS-CLIP with miR29 deficient CD4 T cells reveals miR29 dependent AGO binding sites

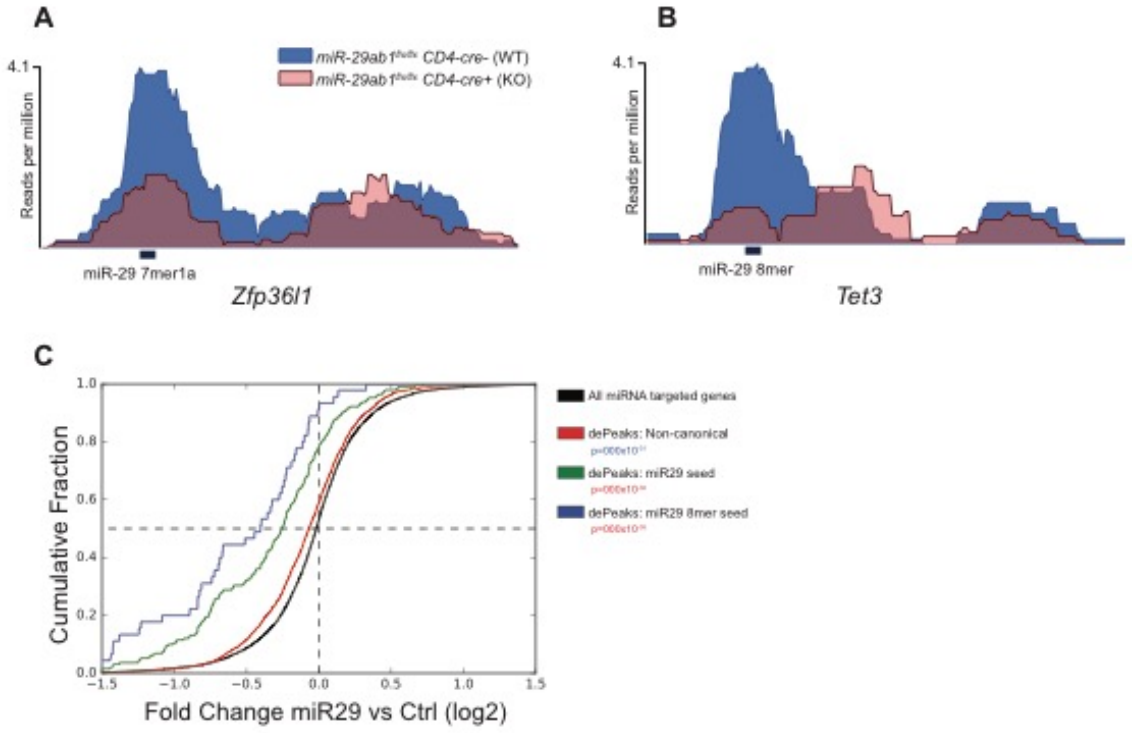


Figure 5: In vitro differentiation profile of miR-29ab-1 conditional knockout CD4 T cells

CD4 T cells stimulated in vitro for 3 days with anti CD3/28, 2 days rest in IL2 and restimulated in PMA/Ionomycin for 4 hours before fixation. A-B) Il-17 expression in cells cultured in Th17 polarizing conditioned media. C-D) Expression of IFNg C) and RORgt (D) in Th0, Th1, Th2 and Th17 in both miR29flxflx CD4-Cre+ and CD4-Cre- T cells, by intracellular cytokine stains.

miR-29ab-1 conditional knockout CD4 T cells produce higher levels of IL17 after simulation

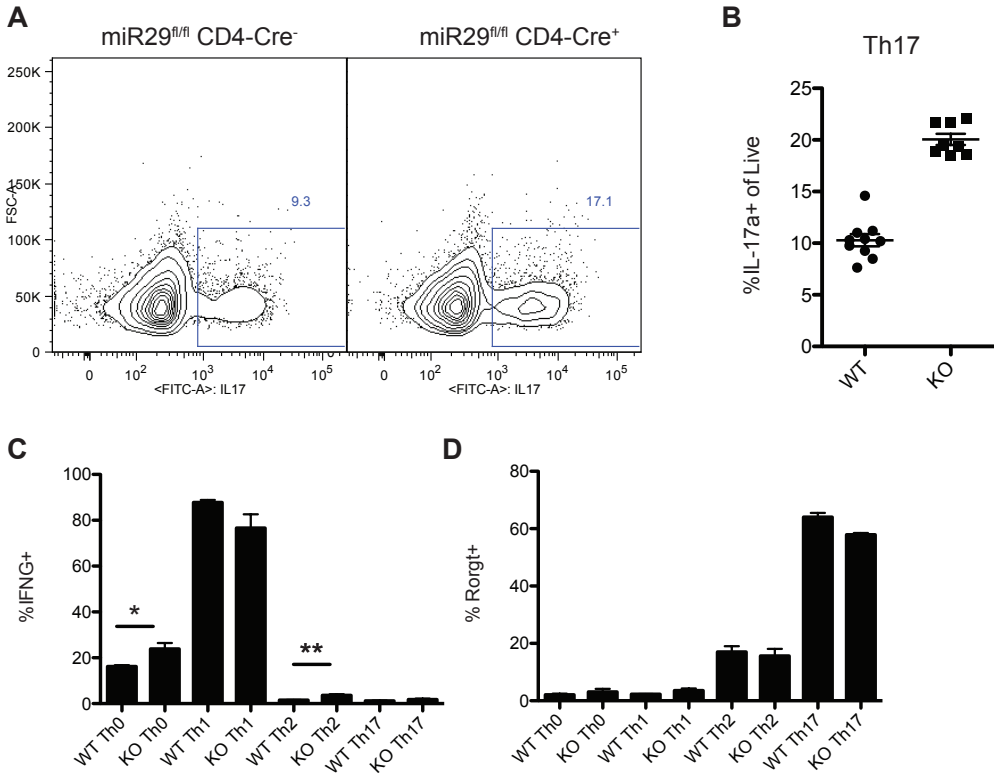


Figure 6 Differential HITS-CLIP targets with known associations with Th17 differentiation

(A-C) HITS-CLIP examples of 3'UTR segments from Th17 associated genes with differentially expressed peaks between the wildtype and miR-29 deficient CD4 T cells. (A) miR-29 8mer site in the 3'UTR of Icos (B) miR-29 8mer site in the 3'UTR of Cflar. (C) Non-canonical differentially expressed peak with no miR-29 seed in the 3'UTR of Sat1b.

Differential HITS-CLIP identifies targets with known associations with Th17 differentiation

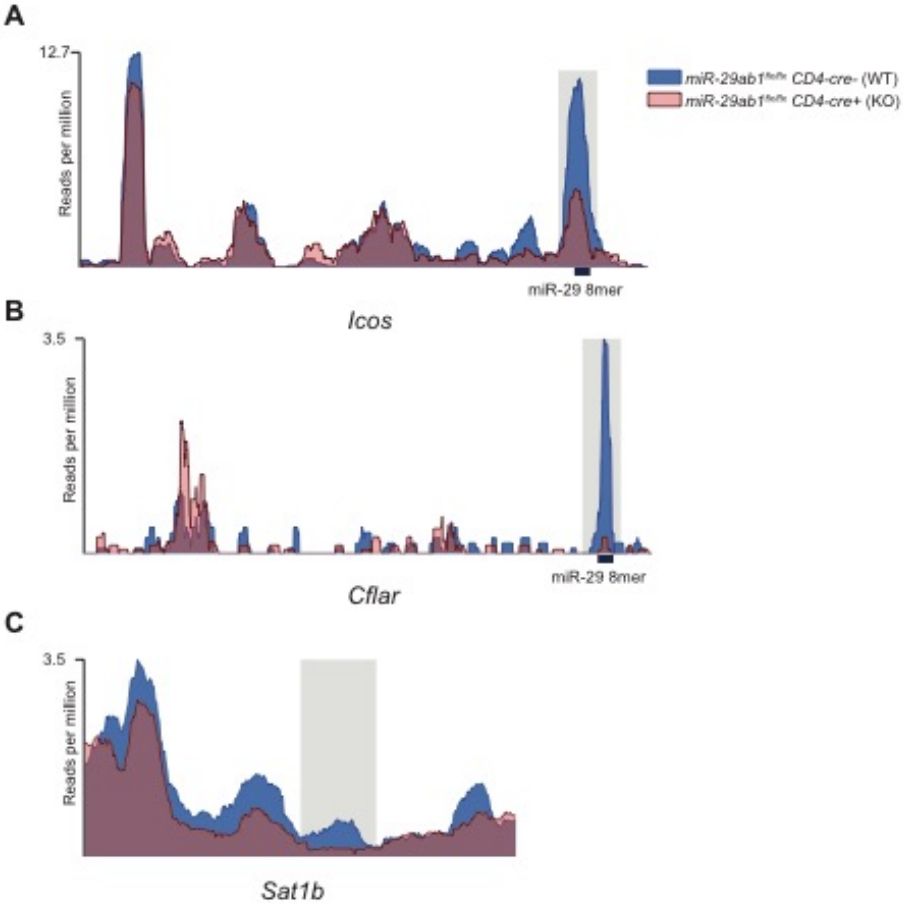
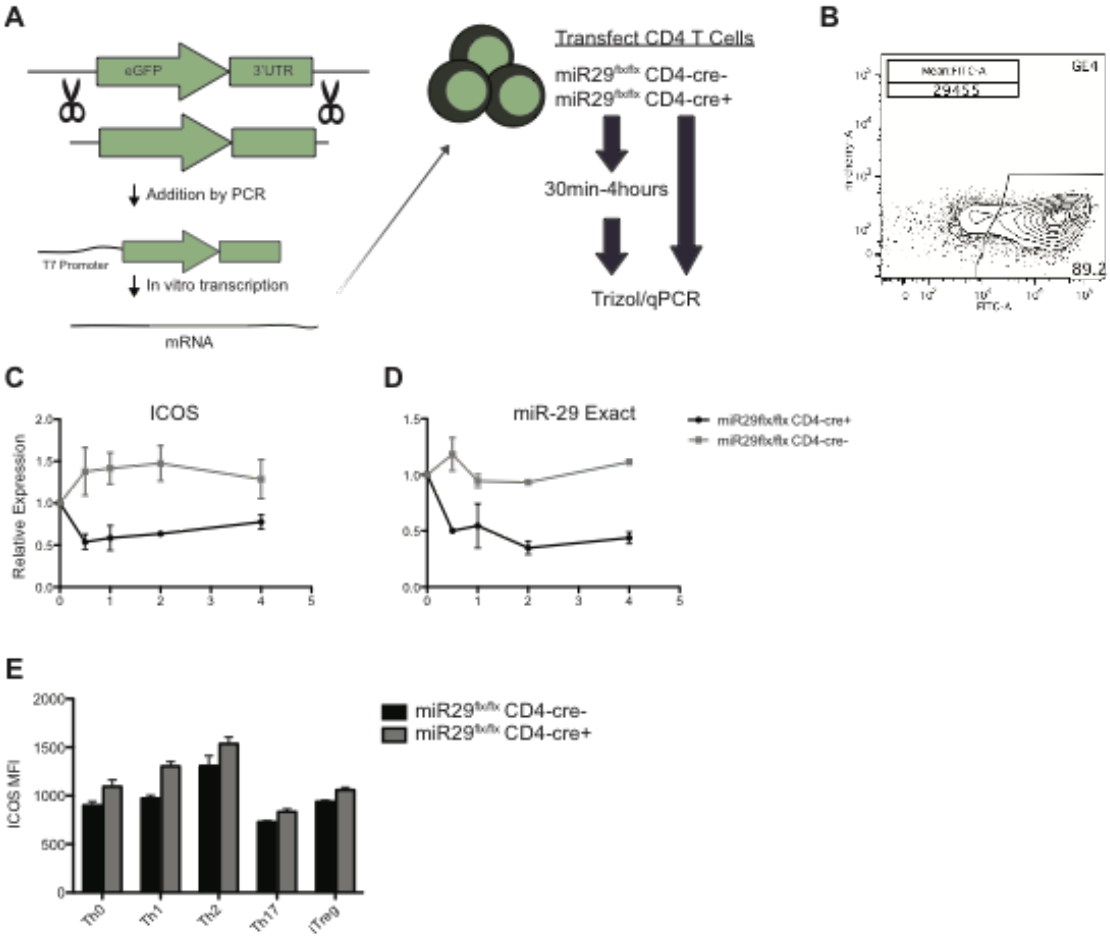


Figure 7: ICOS is directly regulated by miR-29

(A) Generation of mIR-29 reporter constructs was done by cloning either the miR-29 8mer ICOS target site (see fig6), a scrambled control or miR-29-Exact, a sequence with exact miR-29 sequence complementarity, into the 3'UTR of GFP. After isolation of the restriction digested fragment DNA, a T7 promoter sequence was added to the 5' end by PCR and RNA was made by in vitro transcription with T7 polymerase. All three RNA constructs were pooled and transfected into miR-29ab$^{flx/flx}$ CD4-Cre^{-} and miR-29ab$^{flx/flx}$ CD4-Cre^{+} isolated stimulated CD4 T cells. After transfection, cells were washed repeatedly and lysed in TRIZOL or rested in vitro for 0.5-4 hours before TRIZOL lysis. RNA was exracted for qPCR analysis. (B) GFP expression of cells 4 hours after transfection. (C-D) ICOS peak sequence and miR-29-Exact complementary sequence RNA timecourse after transfection. Measured by qPCR. Each sample was normalized to co-tranfected levels of the scrambled control RNA. (E) Mean fluorescence intensity of ICOS measured in day 5 restimulated CD4+ Th0, Th1, Th2, Th17 and iTreg cultures by flow cytometry.

ICOS is directly regulated by miR-29



Chapter 3

ClipPlot: User friendly Clip-Seq graphing tool for presentation quality figures

Introduction

In recent years, the generation of data has massively increased, testing our facilities for data analysis. With the increase in data, comes the challenges of organizing the data, visualizing and summarizing the data, and presenting it in a clear and understandable format.

Sequencing data comes in many forms, including RNAseq, whole genome sequencing and many of these formats can be distilled into simplified lists of reads per gene, as the relative position of the reads in a gene is not nearly as important as the presence of the read in the gene at all.

Crosslinking immunoprecipitation RNA-sequencing (CLIP-Seq) however, greatly depends on the position of the reads in a transcript for the analysis. RNA binding proteins (RBPs) are post-transcriptional regulators of mRNA, binding to cis-regulatory elements often in 3' and 5' untranslated regions (UTRs) and controlling everything from capping and splicing to relocation within cellular compartments and inducing degradation. CLIP-Seq involves the crosslinking of these proteins to their bound mRNAs, digesting unbound RNA, isolating the RPBs of choice, and sequencing the bound RNA. A major advantage of this technique is that it allows the identification of specific regulatory sequences in the RNA transcript by looking for peaks of sequenced and aligned RNA fragments.

There are a number of tools available for the identification, annotation and statistical analysis of these CLIP-Seq RNA peaks, and they are becoming more accessible to a lay researcher. Even visualization of this data has been simplified into an easy to use package called the integrated genomics viewer (IGV). IGV and the

Santa Cruz Genome Browser have been incredibly powerful tools for CLIP-Seq for their ability to scan genome wide datasets quickly and efficiently, while adding annotations and various types of sequencing data in parallel. However a major limitation of these tools, is that they lack easy to use export features for generating figure quality plots. Currently, CLIP-Seq figures are generated by either screenshots from IGV, heavily edited outputs from IGV or the UCSC Genome Browser, or by using custom scripts that either go unpublished, or have a high technical barrier to entry.

Our lab has generated a tool that aims to create an easy to use tool for labs working with CLIP-Seq data to create high quality graphics that are easily edited in a vector based graphics program. ClipPlot is an ipython notebook based webapp that can be run on a local server and used by all members of a lab simultaneously, requiring only the technical expertise of a single user. ClipPlot is designed with ease of use as a priority. Track files are placed in a central location, can be in multiple supported formats ,and can include annotations such as bed files to denote regions of interest. While this is not a replacement for sophisticated visualization tools like IGV, ClipPlot provides an easy to use platform for researchers working with CLIP-Seq or RNA-Seq data to quickly create presentable and easily manipulated graphics, visualizing the shape of sequencing data to defined regions.

Installation

The simplicity of CLIP-Plot installation is that it exists within an IPython Jupyter notebook. There are a number of prerequisites that are required for its usage. These include a current version of IPython as well as the Jupyter notebook package. SciPy is

required, specifically, stats, numpy and matplotlib. The software also relies heavily on the usage of Samtools for processing of indexed and compressed Bam files.

Files can be placed in the ClipPlot default folder, which can be changed in the defaults.py file. This can also be changed on a per-user basis under the “Defaults” accordion button within the notebook itself. File formats supported are Bam (.bam), Bed (.bed) and Bigwig (.bw). Bam files must be sorted and include a matched .idx index (See pre-processing data section). Users must also supply genome annotations to use the gene and transcript identification, which should be placed in the folder Annotations. We have included the annotations for the mouse genome, including refseq annotations (refseq_names.txt), which can be used in gene selection, as well as for drawing the gene annotation track.

In the current early version of this software, the server can be setup manually by placing the notebook in an easily accessible location on a networked computer with access to the alignment files. The user can access the host computer by mapping the host localhost and port to their own with:

```
ssh -L localhost:8889:localhost:8889 User@hostaddress
```

The host computer can start the process in jupyter, using the same port as the user. This will create an instance of the notebook that can be accessed remotely, and will shut down when the user logs off:

```
jupyter notebook --port=8889
```

The user can then access the notebook by navigating to:

http://localhost:8889/tree

and using the folder navigation to access the notebook ClipPlotNotebook.ipynb. Running the first cell in the notebook will run the ipywidgets based GUI.

Usage

Select Region

ClipPlot is designed to be used in conjunction with another genomics viewer such as IGV or the UCSC Genome Browser, as ClipPlot does not have a way to easily scan the genome quickly for areas of interest. The easiest way to frame an area to plot, is to identify genomic coordinates of interest in another viewing program and enter those coordiantes in the Location box. After entering the Strand, the box GeneID will autopopulate. If there is more than one gene at this location, you can choose which one you wish to display along with your sequencing data. The RefseqID field will update with a list of annotations for that GeneID, which can include splice variants. Only one of these can be displayed at a time currently. If you don't have location data and just want to display a gene, you can select Lookup: By GeneID, which will display the 3'UTR of your selected gene. You still have the option to pick isoforms in the RefseqID box.

Files

The files displayed are those that were placed in the ClipPlot Default folder and include those ending in .bam or .bw. A single file can be selected from the Alignments to Use, or multiple files can be selected using ⌘(Osx) or Ctrl(Win) click. The order in which the files are selected are the order in which the tracks will appear when plotted. Selected files will also appear in the box Antisense. Files generated in the antisense direction can be selected here to insure proper strand specific plotting.

For those who are working with AGO-CLIP data, it is often useful to have miRNA binding site data. For this reason, we have included the ability to add Targetscan annotations to the plots. Selecting Targetscan from the bedfile list, will open up a selection of miRNA family from Targetscan. The default file is the included Targetscan.bed, and was generated from Targetscan 7.1 mouse, which includes regions outside Refseq annotated 3'UTRs. Users can select a single miRNA family from this list, or select *"All"* to show all predicted mIRNA binding sites within the framed region.

Formatting

A number of different options are available for formatting and can be toggled on and off. They are listed and described below:

- **BedLabels**: Adds labels to each element of the chosen bed file. If multiple bed files are chosen, it labels the last one chosen. Future versions will have options to change the look of the labels, however currently they are set.

- **Left To Right**: Displays all genes as 5'->3' regardless of the strand. When this is unchecked, graphics will be relative to the genome.
- **Legend**: Adds a legend for each track, with the name of the file being shown (without path).
- **Stagger Bedtracks**: If multiple bedfile regions overlap, instead of being combined, they will be offset from each other.
- **Show RefSeq:** Toggle to display the Refseq annotation of coding, intronic and untranslated regions of the gene selected in the RefseqID box.
- **Shade Bed**: Highlight bedtrack regions with a shaded box.
- **Autoscale**: Scale each track individually. If this is not selected, all tracks will be scaled to the track with the highest maximum.
- **ShowBoundingBox**: Toggles display of a box around each of the plots.
- **Scale to RPM**: Scales each of the tracks to reads per million. Uses the number of reads from each bam file. Utilizes samtools idxstats.

Color Picker

Colors can be selected for the tracks. Currently up to 4 colors can be selected, and they will be applied to the tracks in the order that they are generated. If more than 4 tracks are selected, the colors will cycle through the 4 colors selected.

Output Format

The main function of ClipPlot is to create simple high quality graphics for presentations and figures. With that in mind, a number of different options are provided for the figure output.

- **Xscale**: Defines in inches how wide the figure should be. If set to zero, the X axis will scale with the size of the region graphed. Users can set this to zero for consistency in scale, or specify a number for consistent image size.
- **Font Size**: Defines font size used. Default is 12
- **DPI**: Dots per inch of output. Default is 300.
- **File Format**: Specifies file format. Default is pdf, which is recommended for further vector image manipulation in programs like Adobe Illustrator.
- **File Suffix**: Files are saved in the format geneid.pdf. Changing the suffix allows additions. Example: Setting File Suffix to "_version1" will save the file as geneid_version1.pdf.
- **Output Folder**: Specifies the folder where the graphic will be saved. The green check or red X after the box indicate whether the path provided is recognized by the system.

When all settings are chosen, users simply hit the button "Graph" and the plot will simultaneously be saved to the output directory and displayed in the browser. If the image is not full size, inspection of the image will lead to a full resolution image depending on browser limitations (Tested in Chrome on OSX).

Limitations and Future Directions

There are a number of features that are still missing from a program like this, some of which are inherent to the design, and others that will be added in a future update of the program. The program makes a number of assumptions regarding the structure of the input data. Greater flexibility with input formats would be useful in deployment to a wider audience. Another feature that is needed, is the ability to create various user profiles. Currently each time a user creates an instance of ClipPlot, all settings are reset to default and must be manually entered to get the type of output plot desired. This can easily be solved by user detection or selection, and the ability to update default settings including file locations for each user individually.

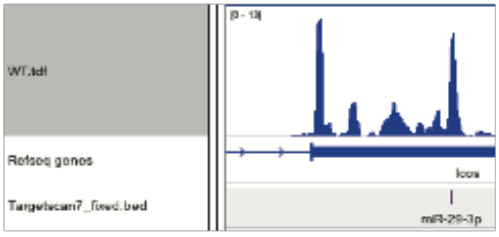
Reliance on a jupyter notebook is not ideal due to a lack of flexibility, and a reliance of running an active instance of the notebook for each individual user. I am currently working on adapting this program to a class based method that can be more easily adapter to other web based display platforms such as flask and bokeh. Ideally this would provide the flexibility to run ClipPlot without the GUI if preferred.

Finally, the most useful feature still lacking is the ability to easily scan a region, and reframe as needed. This may be outside of the program, as primarily ClipPlot is useful for figure generation and not data analysis. Choosing a frame to graph is best done in IGV currently. It may be nice to include this feature in a future iteration of the program, but a more immediate solution may be to simply add +10, +1kb, +10kb buttons to allow easy regeneration of plots.

Figure 1: ClipPlot Figure Generation

(A) Representation of plot from the Integrated Genomics Viewer of the ICOS 3'UTR with HITS-CLIP data from CD4 T cells and an annotated miR-29 target site. (B) ClipPlot representation of the same region.

A


WT.tdf
Refseq genes
Targetscan7_fixed.bed
Icos
miR-29-3p

B

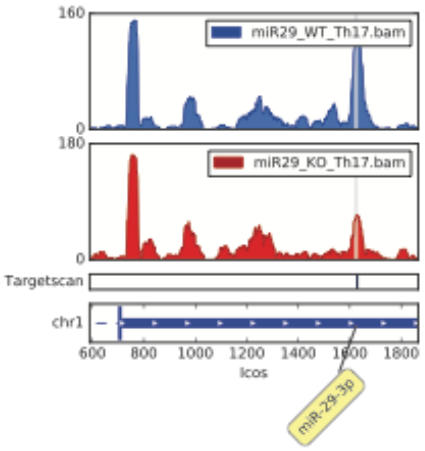

160
miR29_WT_Th17.bam
0
180
miR29_KO_Th17.bam
0
Targetscan
chr1
600
800
1000
1200
1400
1600
1800
Icos
miR-29-3p

Figure 2: Region and file selection

Interface for selecting region to graph in ClipPlot by geneID or location. Dropdown menus for GeneID and RefseqID allow selection of desired isoform. Bedfile selection can either accommodate any formatted bedfile, or users can select miRNA binding sites from a list. Multiple alignments may be selected in the desired order, and the option to graph the antisense strand (antisense to the strand selected at the top) depending on sequencing platform used.

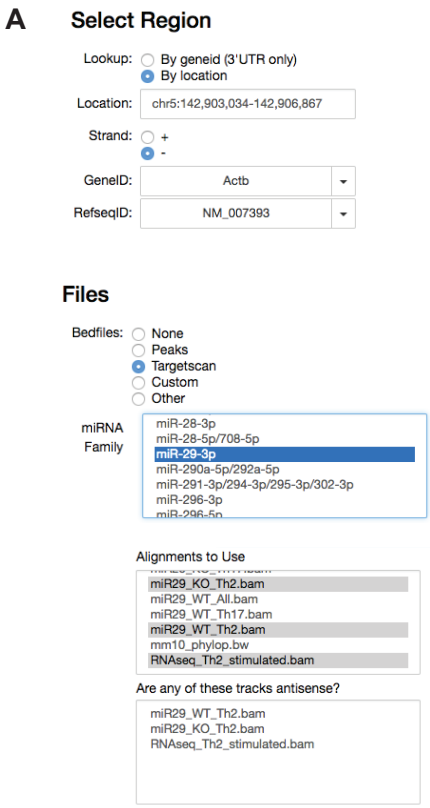

A
Select Region
Lookup:
By geneid (3'UTR only)
By location
Location:
chr5:142,903,034-142,906,867
Strand:
+
-
GeneID:
Actb
RefseqID:
NM_007393
Files
Bedfiles:
None
Peaks
Targetscan
Custom
Other
miRNA Family
miR-28-3p
miR-28-5p/708-5p
miR-29-3p
miR-290a-5p/292a-5p
miR-291-3p/294-3p/295-3p/302-3p
miR-296-3p
Alignments to Use
miR29_KO_Th2.bam
miR29_WT_All.bam
miR29_WT_Th17.bam
miR29_WT_Th2.bam
mm10_phylop.bw
RNAseq_Th2_stimulated.bam
Are any of these tracks antisense?
miR29_WT_Th2.bam
miR29_KO_Th2.bam
RNAseq_Th2_stimulated.bam

Figure 3: Display options and output format selection

Interface for selecting various display options including color of tracks, font size and DPI. Users can also change the output format to matplotlib allowed formats.

A

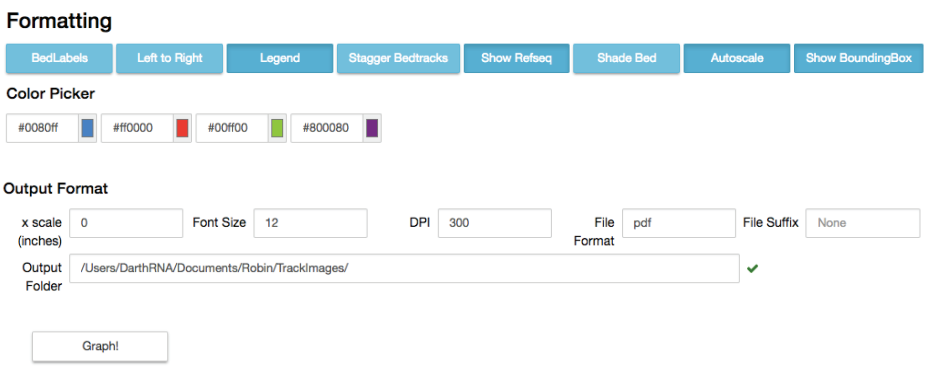

Formatting
BedLabels
Left to Right
Legend
Stagger Bedtracks
Show Refseq
Shade Bed
Autoscale
Show BoundingBox
Color Picker
#0080ff
#ff0000
#00ff00
#800080
Output Format
x scale (inches)
0
Font Size
12
DPI
300
File Format
pdf
File Suffix
None
Output Folder
/Users/DarthRNA/Documents/Robin/TrackImages/
Graph!

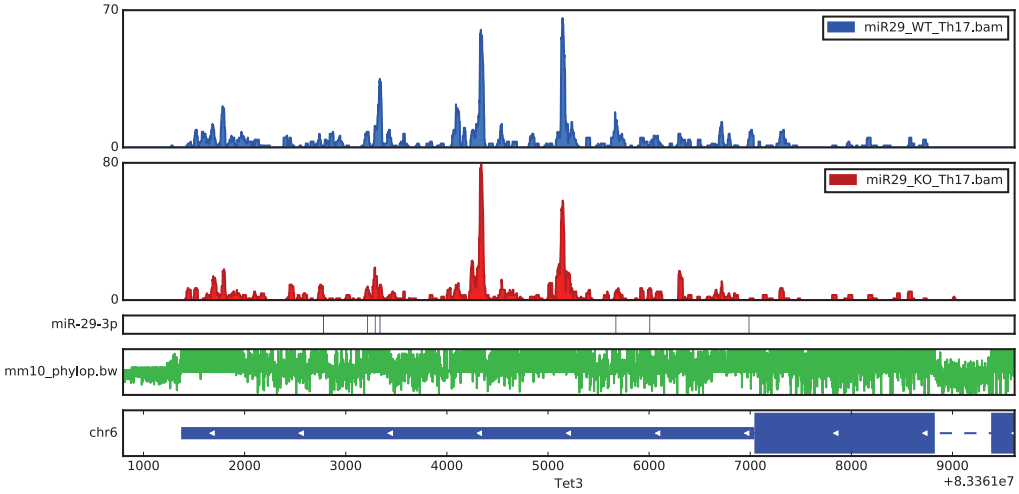
Figure 4: ClipPlot Output

(A) If a bed file is chosen to be displayed, regions in the bed within the selected range will be displayed to assist in site identification if individual bed region labeling is turned off. Display name and genomic start and stop positions. (B) Representative ClipPlot graph of Tet3 with options selected in Fig3. Contains two HITS-Clip bam tracks from wildtype and mIR-29 deficient CD4 T cells. Mir-29 predicted target sites are highlighted below.

A

```
Figure will be saved as: /Users/DarthRNA/Documents/Robin/TrackImages/Tet3.pdf
======
Regions
======
miR-29-3p [83363776, 83363784]
miR-29-3p [83364211, 83364218]
miR-29-3p [83364288, 83364295]
miR-29-3p [83364335, 83364343]
miR-29-3p [83366668, 83366675]
miR-29-3p [83367003, 83367010]
miR-29-3p [83367984, 83367991]
```

B



Chapter 4

Discussion, conclusions and future directions

Conclusion

Our work contained in this thesis focuses on the identification and characterization of miRNA binding sites in the context of CD4 helper T cell differentiation. A large fraction of my work was dedicated to the task of optimizing the use of HITS-CLIP for use in our lab and the generation of tools and pipelines for the analysis of that data. Using HITS-CLIP, we were able to map thousands of Ago-mRNA sites of interactions as well as quantify the miRNAs that were bound by Ago. Dense genomic regions of sequenced mRNA allowed us to identify potential miRNA sites of binding (peaks) that we combined with target predictions and gene expression analysis to further increase confidence in site identification. Analysis of the data was accomplished through the use of custom built pipelines and bioinformatic tools. Peak visualization was accomplished partially through the use of ClipPlot software. This was used to generate figures for the thesis, and allowed other members of the lab network access to all HITS-CLIP data, to help them find targets for other miRNAs they were studying as well as generation of visualizations for presentations and publications. The use of miR-29 CD4 conditional knockout mice provided the opportunity to perform HITS-CLIP in the presence or absence of the single miRNA miR-29. This revealed differential Ago peaks that were dependent on the expression of miR-29, suggesting sites of miR-29 interaction and possible regulation of the transcript mRNA. The miR-29 deficient T cells also revealed a defect in the production of Il-17, which was observed under a variety of CD4+T cell subset polarizing conditions, but was most strongly seen in Il-17 conditions. We were able to identify a number of miR-29 target genes with the potential to contribute to the observed phenotype. One of these genes, ICOS, contained a highly

expressed miR-29 dependent Ago binding peak in the 3'UTR. We were able to use RNA transfection to observe miR-29 dependent regulation of the isolated target site, confirming ICOS as a direct target of miR-29.

There are continuing challenges to overcome in the pursuit to create a map of all miRNA-mRNA target interactions. Some of this comes from the degenerate sequence binding of miRNAs to their target. Less appreciated, is the context the rest of the cellular state contributes to the ability of any given miRNA to bind or regulate a target. There are many factors that can contribute to the ability of miRNAs to bind. First of all, gene expression, while simple, can be overlooked when searching for miRNA binding sites. High expression of a gene with multiple copies of a single miRNA binding site, could act to sponge miRNA expression, leading to reduce binding elsewhere. Changes in splicing can lead to variable levels of different isoforms of the same gene, all with different untranslated regions and miRNA binding sites. It is also important to consider that changes in expression of RNA binding proteins can change the secondary structure of mRNA, reducing or enhancing accessibility to Ago binding and miRNA regulation.

A major advantage of HITS-CLIP is the ability to gain experimental evidence of Ago binding to mRNA, but there are also limitations. On the flip side of being able to see miRNAs that are actively regulating highly expressed genes, it means we are also unable to detect miRNA binding sites on lowly expressed or transiently expressed transcripts. This can be overcome by utilizing HITS-CLIP in a variety of systems, tissue types and conditions. There is also the concern of false positive hits. Ago binding peaks may not always mean that miRNAs are binding there, or that it will lead to posttranscriptional regulation. By combining HITS-CLIP with bioinformatic predictions,

gene expression analysis, and genetic manipulation like knocking out miRNAs, confidence in these sites increases, and they can become a powerful tool for deciding on targets to investigate. One of the biggest advantages of working with miRNAs, is that finding the macroscopic effects of a miRNA is only one level of what is being uncovered. Discovering the targets of that miRNA, allows the discovery of new pathways. While the list of possible targets may seem immense, HITS-CLIP can trim that down to a list of most likely candidates, as well as identify a few genes that weren't predicted in the first place.

References

Adoro, S., Cubillos-Ruiz, J.R., Chen, X., Deruaz, M., Vrbanac, V.D., Song, M., Park, S., Murooka, T.T., Dudek, T.E., Luster, A.D., Tager, A.M., Streeck, H., Bowman, B., Walker, B.D., Kwon, D.S., Lazarevic, V., Glimcher, L.H., 2015. IL-21 induces antiviral microRNA-29 in CD4 T cells to limit HIV-1 infection. Nat Comms 6, 7562. doi:10.1038/ncomms8562

Agarwal, V., Bell, G.W., Nam, J.-W., Bartel, D.P., 2015. Predicting effective microRNA target sites in mammalian mRNAs. Elife 4. doi:10.7554/eLife.05005

Ansel, K.M., 2013. RNA regulation of the immune system. Immunol Rev 253, 5–11. doi:10.1111/imr.12062

Baumjohann, D., (null), Clingan, J.M., Morar, M.M., Patel, S., de Kouchkovsky, D., Bannard, O., Bluestone, J.A., Matloubian, M., Ansel, K.M., Jeker, L.T., 2013. The microRNA cluster miR-17[sim]92 promotes TFH cell differentiation and represses subset-inappropriate gene expression. Nat Immunol 14, 840–848. doi:doi:10.1038/ni.2642

Boudreau, R.L., Jiang, P., Gilmore, B.L., Spengler, R.M., Tirabassi, R., Nelson, J.A., Ross, C.A., Xing, Y., Davidson, B.L., 2014. Transcriptome-wide Discovery of microRNA Binding Sites in Human Brain. Neuron 81, 294–305. doi:10.1016/j.neuron.2013.10.062

Bracken, C.P., Li, X., Wright, J.A., Lawrence, D., Pillman, K.A., Salmanidis, M., Anderson, M.A., Dredge, B.K., Gregory, P.A., Tsykin, A., Neilsen, C., Thomson, D.W., Bert, A.G., Leerberg, J.M., Yap, A.S., Jensen, K.B., Khew-Goodall, Y., Goodall, G.J., 2014. Genome-wide identification of miR-200 targets reveals a regulatory network controlling cell invasion. EMBO J. doi:10.15252/embj.201488641

Chi, S.W., Zang, J.B., Mele, A., Darnell, R.B., 2009. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature 460, 479–486. doi:10.1038/nature08170

Ebert, M.S., Sharp, P.A., 2012. Roles for microRNAs in conferring robustness to biological processes. Cell 149, 515–524. doi:10.1016/j.cell.2012.04.005

Friedman, R.C., Farh, K.K.H., Burge, C.B., Bartel, D.P., 2008. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 19, 92–105. doi:10.1101/gr.082701.108

Grimson, A., Farh, K.K.-H., Johnston, W.K., Garrett-Engele, P., Lim, L.P., Bartel, D.P., 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 27, 91–105. doi:10.1016/j.molcel.2007.06.017

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G.S., Dewell, S., Zavolan, M., Tuschl, T., 2010. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. Cell 141, 129–141. doi:10.1016/j.cell.2010.03.009

He, Y., Huang, C., Lin, X., Li, J., 2013. MicroRNA-29 family, a crucial therapeutic target for. Biochimie 95, 1355–1359. doi:10.1016/j.biochi.2013.03.010

Hutloff, A., Dittrich, A.M., Beier, K.C., Eljaschewitsch, B., Kraft, R., Anagnostopoulos, I., Kroczek, R.A., 1999. ICOS is an inducible T-cell co-stimulator structurally and

functionally related to CD28. Nature 397, 263–266. doi:10.1038/16717
Hutvágner, G., McLachlan, J., Pasquinelli, A.E., Bálint, E., Tuschl, T., Zamore, P.D., 2001. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science 293, 834–838. doi:10.1126/science.1062961
Jaskiewicz, L., Bilen, B., Hausser, J., Zavolan, M., 2012. Argonaute CLIP - A method to identify in vivo targets of miRNAs. Methods 58, 106–112. doi:10.1016/j.ymeth.2012.09.006
Khvorova, A., Reynolds, A., Jayasena, S.D., 2003. Functional siRNAs and miRNAs exhibit strand bias. Cell 115, 209–216.
Kuchen, S., Resch, W., Yamane, A., Kuo, N., Li, Z., Chakraborty, T., Wei, L., Laurence, A., Yasuda, T., Peng, S., Hu-Li, J., Lu, K., Dubois, W., Kitamura, Y., Charles, N., Sun, H.-W., Muljo, S., Schwartzberg, P.L., Paul, W.E., O'Shea, J., Rajewsky, K., Casellas, R., 2010. Regulation of MicroRNA Expression and Abundance during Lymphopoiesis. Immunity 32, 828–839. doi:10.1016/j.immuni.2010.05.009
Lagos-Quintana, M., Rauhut, R., Lendeckel, W., Tuschl, T., 2001. Identification of novel genes coding for small expressed RNAs. Science 294, 853–858. doi:10.1126/science.1064921
Lee, R.C., Feinbaum, R.L., Ambros, V., 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75, 843–854.
Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., Kim, V.N., 2003. The nuclear RNase III Drosha initiates microRNA processing. Nature 425, 415–419. doi:10.1038/nature01957
Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., Darnell, J.C., Darnell, R.B., 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 456, 464–469. doi:doi:10.1038/nature07488
Liston, A., Papadopoulou, A.S., Danso-Abeam, D., Dooley, J., 2012. MicroRNA-29 in the adaptive immune system: setting the threshold. Cell. Mol. Life Sci. 69, 3533–3541. doi:10.1007/s00018-012-1124-0
Loeb, G.B., Khan, A.A., Canner, D., Hiatt, J.B., Shendure, J., Darnell, R.B., Leslie, C.S., Rudensky, A.Y., 2012. Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. Mol Cell 48, 760–770. doi:10.1016/j.molcel.2012.10.002
Lytle, J.R., Yario, T.A., Steitz, J.A., 2007. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5“ UTR as in the 3” UTR. Proc Natl Acad Sci USA 104, 9667–9672. doi:10.1261/rna.5241404
Ma, F., Xu, S., Liu, X., Zhang, Q., Xu, X., Liu, M., Hua, M., Li, N., Yao, H., Cao, X., 2011. The microRNA miR-29 controls innate and adaptive immune responses to intracellular bacterial infection by targeting interferon-γ. Nat Immunol 12, 861–869. doi:10.1038/ni.2073
N D Mendes, A.T.F.M.F.S., 2009. Current tools for the identification of miRNA genes and their targets. Nucleic Acids Res. 37, 2419. doi:10.1093/nar/gkp145
Oliveira, L.H., Schiavinato, J.L., Fr guas, M.S., Lucena-Araujo, A.R., Haddad, R., Ara jo, A.L.G., Dalmazzo, L.F., Rego, E.M., Covas, D.T., Zago, M.A., Panepucci, R.A., 2015. Potential roles of microRNA-29a in the molecular pathophysiology of T-cell acute lymphoblastic leukemia. Cancer Sci 106, 1264–1277. doi:10.1111/cas.12766

Papadopoulou, A.S., Dooley, J., Linterman, M.A., Pierson, W., Ucar, O., Kyewski, B., Zuklys, S., Hollander, G.A., Matthys, P., Gray, D.H.D., De Strooper, B., Liston, A., 2012. The thymic epithelial microRNA network elevates the threshold for infection-associated thymic involution via miR-29a mediated suppression of the IFN-α receptor. Nat Immunol 13, 181–187. doi:10.1038/ni.2193

Pua, H.H., Steiner, D.F., Patel, S., Gonzalez, J.R., Ortiz-Carpena, J.F., (null), Chiou, N.-T., Gallman, A., de Kouchkovsky, D., Jeker, L.T., McManus, M.T., Erle, D.J., Ansel, K.M., 2016. MicroRNAs 24 and 27 Suppress Allergic Inflammation and Target a Network of Regulators of T Helper 2 Cell-Associated Cytokine Production. Immunity 1–13. doi:10.1016/j.immuni.2016.01.003

Ray M Marín, M.Š.J.V., 2013. Searching the coding region for microRNA targets. RNA 19, 467–474. doi:10.1261/rna.035634.112

Schmitt, M.J., Philippidou, D., 2012. Interferon-γ-induced activation of Signal Transducer and Activator of Transcription 1 (STAT1) up-regulates the tumor suppressing microRNA-29 family in …. Cell.

Simpson, L.J., Patel, S., Bhakta, N.R., Choy, D.F., Brightbill, H.D., Ren, X., Wang, Y., Pua, H.H., Baumjohann, D., Montoya, M.M., Panduro, M., Remedios, K.A., Huang, X., Fahy, J.V., Arron, J.R., Woodruff, P.G., Ansel, K.M., 2014. A microRNA upregulated in asthma airway T cells promotes TH2 cytokine production. Nat Immunol 15, 1162–1170. doi:10.1038/ni.3026

Smith, K.M., Guerau-de-Arellano, M., Costinean, S., Williams, J.L., Bottoni, A., Cox, G.M., Satoskar, A.R., Croce, C.M., Racke, M.K., Lovett-Racke, A.E., Whitacre, C.C., 2012. miR-29ab1 Deficiency Identifies a Negative Feedback Loop Controlling Th1 Bias That Is Dysregulated in Multiple Sclerosis. J Immunol 189, 1567–1576. doi:10.4049/jimmunol.1103171

Spengler, R.M., Zhang, X., Cheng, C., McLendon, J.M., Skeie, J.M., Johnson, F.L., Davidson, B.L., Boudreau, R.L., 2016. Elucidation of transcriptome-wide microRNA binding sites in human cardiac tissues by Ago2 HITS-CLIP. Nucleic Acids Res. 44, 7120–7131. doi:10.1093/nar/gkw640

Steiner, D.F., Thomas, M.F., Hu, J.K., Yang, Z., Babiarz, J.E., Allen, C.D.C., Matloubian, M., Blelloch, R., Ansel, K.M., 2011. MicroRNA-29 Regulates T-Box Transcription Factors and Interferon-γ Production in Helper T Cells. Immunity 35, 169–181. doi:10.1016/j.immuni.2011.07.009

Ule, J., Jensen, K., Mele, A., Darnell, R.B., 2005. CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods 37, 376–386. doi:10.1016/j.ymeth.2005.07.018

Wang, T., Xie, Y., Xiao, G., 2014. Genome Biology | Full text | dCLIP: a computational approach for comparative CLIP-seq analyses. Genome Biol.

Wang, Y., Zhang, X., Li, H., Yu, J., Ren, X., 2013. European Journal of Cell Biology. European Journal of Cell Biology 92, 123–128. doi:10.1016/j.ejcb.2012.11.004

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

Yang Eric Guo, T.O.J.A.S., 2015. Herpesvirus saimiri MicroRNAs Preferentially Target Host Cell Cycle Regulators. Journal of Virology 89, 10901. doi:10.1128/JVI.01884-15

Zhang, C., Darnell, R.B., 2011. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat. Biotechnol. 29, 607–614. doi:10.1038/nbt.1873
Zhao, J., Luo, R., Xu, X., Zou, Y., 2015. High-throughput sequencing of RNAs isolated by cross-linking immunoprecipitation (HITS-CLIP) reveals Argonaute-associated microRNAs and targets in …. Parasites & ….
Zhu, J., Paul, W.E., 2008. CD4 T cells: fates, functions, and faults. Blood 112, 1557–1569. doi:10.1182/blood-2008-05-078154
Zhu, J., Yamane, H., Paul, W.E., 2010. Differentiation of Effector CD4 T Cell Populations *. Annu. Rev. Immunol. 28, 445–489. doi:10.1146/annurev-immunol-030409-101212

nature immunology

The microRNA cluster miR-17~92 promotes T_{FH} cell differentiation and represses subset-inappropriate gene expression

Dirk Baumjohann[1], Robin Kageyama[1], Jonathan M Clingan[2,3], Malika M Morar[4], Sana Patel[1], Dimitri de Kouchkovsky[4], Oliver Bannard[5], Jeffrey A Bluestone[4,6], Mehrdad Matloubian[2], K Mark Ansel[1,7] & Lukas T Jeker[4,6,7]

Follicular helper T cells (T_{FH} cells) are the prototypic helper T cell subset specialized to enable B cells to form germinal centers (GCs) and produce high-affinity antibodies. We found that expression of microRNAs (miRNAs) by T cells was essential for T_{FH} cell differentiation. More specifically, we show that after immunization of mice with protein, the miRNA cluster miR-17~92 was critical for robust differentiation and function of T_{FH} cells in a cell-intrinsic manner that occurred regardless of changes in proliferation. In a viral infection model, miR-17~92 restrained the expression of genes 'inappropriate' to the T_{FH} cell subset, including the direct miR-17~92 target *Rora*. Removal of one *Rora* allele partially 'rescued' the inappropriate gene signature in miR-17~92-deficient T_{FH} cells. Our results identify the miR-17~92 cluster as a critical regulator of T cell–dependent antibody responses, T_{FH} cell differentiation and the fidelity of the T_{FH} cell gene-expression program.

T cell–dependent antibody responses are a pillar of adaptive immunity; they constitute protective responses to a wide variety of pathogens, form the basis of the immunological memory induced by the vast majority of effective vaccines and underlie the pathogenesis of many autoimmune and allergic disorders[1,2]. Follicular helper T cells (T_{FH} cells) are a subset of $CD4^+$ T cells specialized to provide signals that induce the growth, differentiation, immunoglobulin isotype switching, affinity maturation and antibody secretion of B cells[1]. They are defined by Bcl-6, a transcriptional repressor that is necessary and sufficient to direct T_{FH} cell differentiation[3–5], and by abundant expression of the chemokine receptor CXCR5 and the immunoregulatory receptor PD-1 (ref. 1). T_{FH} cell differentiation begins very early in the immune response, coincident with rapid proliferation that expands the pool of responding cells. Expression of Bcl-6 is induced very early during T cell activation and is further upregulated in developing T_{FH} cells[6] in conjunction with upregulation of CXCR5 expression and downregulation of CCR7 expression[7]. Such changes in the expression of homing receptors allow developing T_{FH} cells to migrate to the boundary between the T cell zone and B cell follicles of secondary lymphoid organs, where they encounter antigen-specific B cells[1]. Continued cognate interactions with antigen-presenting B cells in the germinal centers (GCs) of lymphoid follicles further polarize T_{FH} cells[8] and help to maintain the T_{FH} cell phenotype[9]. Beyond their established role in orchestrating humoral immunity, T_{FH} cells and transient T_{FH}-like transition states of activated $CD4^+$ T cells have been linked to T helper type 1 (T_H1) differentiation[10,11] and the generation of memory helper T cells[12,13].

MicroRNAs (miRNAs) have emerged as important regulators of many aspects of the differentiation and function of cells of the immune system[14]. The fates of activated helper T cells are very sensitive to precise 'dosing' of regulatory factors[10] and are therefore subject to regulation by the fine-tuning activity of miRNAs. There is some evidence that miRNAs regulate the T_{FH} cell gene-expression program[5] and the plasticity of T_{FH} cells[15]. However, the contribution of miRNAs to T_{FH} cell differentiation and function remains largely unknown.

Here we show that global miRNA expression in $CD4^+$ T cells was absolutely required for the differentiation of T_{FH} cells *in vivo*, independently of any proliferative defects associated with miRNA deficiency. Furthermore, we found that the miR-17~92 cluster was particularly important for robust T_{FH} cell responses. In a protein-immunization model, miR-17~92 contributed to the differentiation of an early $CXCR5^{hi}Bcl\text{-}6^{hi}$ T_{FH} cell population, in part by targeting the mRNA that encodes the tumor suppressor PTEN. In a viral infection model, miR-17~92 repressed the expression of genes encoding molecules associated with the function of other helper T cell subsets but 'inappropriate' and not normally expressed in T_{FH} cells. We identified

[1]Department of Microbiology & Immunology, Sandler Asthma Basic Research Center, University of California, San Francisco, San Francisco, California, USA. [2]Division of Rheumatology, Department of Medicine, and Rosalind Russell Medical Research Center for Arthritis, University of California, San Francisco, San Francisco, California, USA. [3]Graduate Program in Biomedical Sciences, University of California, San Francisco, California, USA. [4]Diabetes Center and Department of Medicine, University of California, San Francisco, San Francisco, California, USA. [5]Department of Microbiology & Immunology, Howard Hughes Medical Institute, University of California, San Francisco, California, USA. [6]Department of Pathology, University of California, San Francisco, San Francisco, California, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to L.T.J. (ljeker@diabetes.ucsf.edu) or K.M.A. (mark.ansel@ucsf.edu).

Received 27 March; accepted 15 May; published online 30 June 2013; doi:10.1038/ni.2642

and confirmed *Rora* (which encodes the transcription factor RORα) as a direct target of miR-17~92 that contributed to the substantial phenotypic changes observed. We conclude that miRNAs are important regulators of the differentiation and function of T_{FH} cells.

RESULTS

The differentiation and function of T_{FH} cells requires miRNAs

To investigate the global role of miRNAs in the differentiation and function of T_{FH} cells, we obtained naive, congenically marked ($CD45.2^+$) $CD4^+$ T cells from mice that are deficient in the miRNA-biogenesis factor DGCR8 and are thus deficient in miRNAs ($CD4\text{-}Cre^+Dgcr8^{fl/fl}$; called '$Dgcr8^{\Delta/\Delta}$' here), with transgenic expression of the OT-II ovalbumin (OVA)-specific T cell antigen receptor (TCR), and from their control, miRNA-sufficient ($CD4\text{-}Cre^+Dgcr8^{+/fl}$; called '$Dgcr8^{+/\Delta}$' here) OT-II counterparts. We transferred those cells into $CD45.1^+$ wild-type recipient mice and subsequently immunized the recipients with OVA. Considerably fewer miRNA-deficient OT-II cells than control OT-II cells accumulated in the draining lymph nodes 4.5 d after immunization (**Fig. 1a**). Among the remaining $Dgcr8^{\Delta/\Delta}$ OT-II cells, the frequency of $PD\text{-}1^{hi}CXCR5^{hi}$ T_{FH} cells was much lower than that among the transferred control cells (**Fig. 1a**), whereas the frequency of endogenous T_{FH} cells was very similar in both sets of recipients (**Supplementary Fig. 1a**). The diminished generation of T_{FH} cells resulted in significantly lower relative and absolute numbers of $Fas^+Bcl\text{-}6^+$ GC B cells (**Fig. 1b**). Thus, T cell–intrinsic miRNAs were critical for T_{FH} cell responses and GC formation.

To distinguish between impaired proliferation and a potential intrinsic defect in T_{FH} cell differentiation, we tracked the generation of T_{FH} cells according to the number of cell divisions in the adoptive-transfer model[6]. The miRNA-deficient OT-II cells proliferated less than did their control, miRNA-sufficient ($CD4\text{-}Cre^+Dgcr8^{+/+}$; called '$Dgcr8^{+/+}$' here) OT-II counterparts (**Fig. 1c** and **Supplementary Fig. 1b**). Early induction of Bcl-6 expression was similar in miRNA-deficient and control cells, as all proliferating cells upregulated Bcl-6 expression (**Fig. 1c** and **Supplementary Fig. 1b**, $Dgcr8^{\Delta/\Delta}$ versus unimmunized). However, miRNAs were critical for further upregulation of Bcl-6 expression in developing T_{FH} cells as they continued to proliferate (**Fig. 1c**, $Dgcr8^{\Delta/\Delta}$ versus $Dgcr8^{+/+}$). In addition, miRNA-deficient T cells completely failed to upregulate CXCR5 expression, sustained abnormally high CCR7 expression, failed to accumulate in proximity to B cells at the boundary between the T cell and B cell zones and did not enter B cell follicles (**Fig. 1c** and **Supplementary Fig. 1b–e**). Thus, miRNAs were essential for the differentiation and function of T_{FH} cells.

Regulation of T_{FH} cell and GC responses by miR-17~92

Very little is known about the functions of specific miRNAs in T_{FH} cells. A published report has proposed that Bcl-6 inhibits miR-17~92 expression to prevent it from directly repressing CXCR5 expression[5], which would interfere with T cell migration and inhibit the generation and function of T_{FH} cells. However, T cell activation induces miR-17~92 expression[16,17], and overexpression of miR-17~92 in lymphocytes leads to a lupus-like autoimmune syndrome with high antibody titers, which suggests enhanced T_{FH} cell function[18]. To directly determine whether miR-17~92 inhibits or promotes the generation of T_{FH} cells, we infected mice lacking miR-17~92 only in T cells (with *lox*P-flanked alleles encoding miR-17~92 ($Mirc1^{fl/fl}$) deleted by Cre recombinase expressed from the T cell–specific *Cd4* promoter ($CD4\text{-}Cre^+Mirc1^{fl/fl}$); called 'T17~92$^{\Delta/\Delta}$' here) or their control counterparts with sufficient miR-17~92 in T cells ($CD4\text{-}Cre^+Mirc1^{+/+}$ or $CD4\text{-}Cre^-Mirc1^{fl/fl}$; collectively called 'T17~92$^{+/+}$' here) with lymphocytic choriomeningitis virus (LCMV), Armstrong strain. Infected T17~92$^{\Delta/\Delta}$ mice had considerably fewer splenic T_{FH} cells than did T17~92$^{+/+}$ mice and had severe impairment in the generation of GC B cells (**Fig. 2a**). Infected T17~92$^{\Delta/\Delta}$ mice also had overall lower spleen cellularity and frequency of activated ($CD44^{hi}$) T cells than

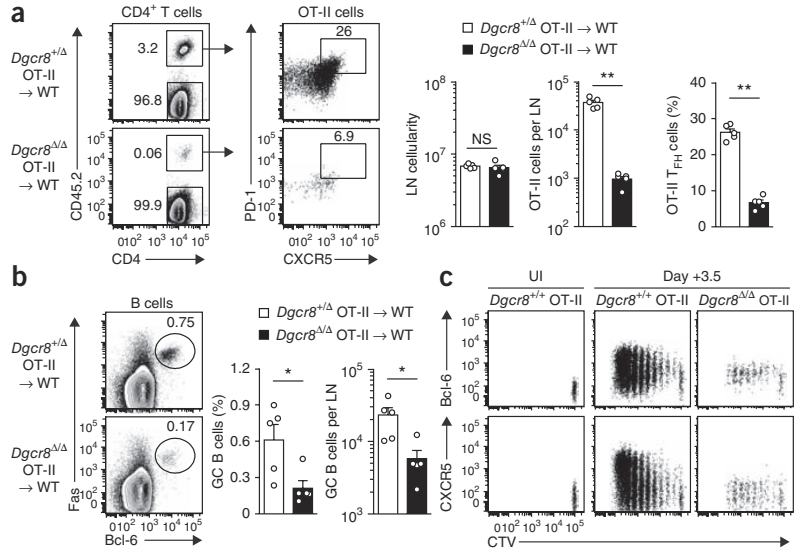


Figure 1 T cell–expressed miRNAs are essential for the differentiation of T_{FH} cells and induction of GC B cells. (**a**) Flow cytometry of cells from draining lymph nodes of wild-type ($CD45.1^+$) recipient mice (WT) given adoptive transfer of naive $Dgcr8^{+/\Delta}$ or $Dgcr8^{\Delta/\Delta}$ OT-II ($CD45.2^+$) cells (left margin: donor→recipient), followed by subcutaneous immunization of recipients in the hind footpads with NP-OVA in the adjuvant alum and analysis 4.5 d after immunization (left); gated on live $CD4^+B220^-$ lymphocytes. Numbers adjacent to outlined areas indicate percent $CD45.2^+$ (OT-II) cells (top right) or $CD45.2^-$ (host) cells (bottom right) among total $CD4^+$ T cells (far left plots) or percent $PD\text{-}1^{hi}CXCR5^{hi}$ T_{FH} cells among transferred OT-II cells (middle left plots). Right, quantification of the results at left; each symbol represents an individual mouse ($n = 5$). LN, lymph node. (**b**) Flow cytometry and quantification of GC B cells in the draining lymph nodes of mice as in **a**; B cells were gated as live $CD19^+B220^+$ lymphocytes. Numbers above outlined areas indicate percent $Fas^+Bcl\text{-}6^+$ cells. (**c**) Flow cytometry of cells from draining popliteal lymph nodes of wild-type recipient mice given adoptive transfer of naive $Dgcr8^{+/+}$ or $Dgcr8^{\Delta/\Delta}$ OT-II ($CD45.2^+$) cells labeled with the fluorescent dye CellTrace Violet (CTV), followed by no immunization (UI; left) or immunization of recipients in the hind footpads with NP-OVA in alum (middle and right) and analysis of the expression of Bcl-6 and CXCR5 at 3.5 d after immunization; OT-II cells were gated as live $CD45.2^+CD4^+B220^-$ lymphocytes. NS, not significant; $^*P < 0.05$ and $^{**}P < 0.01$ (two-tailed nonparametric Mann-Whitney test). Data are representative of three (**a**,**b**) or five (**c**) independent experiments (mean and s.e.m. in **a**,**b**).

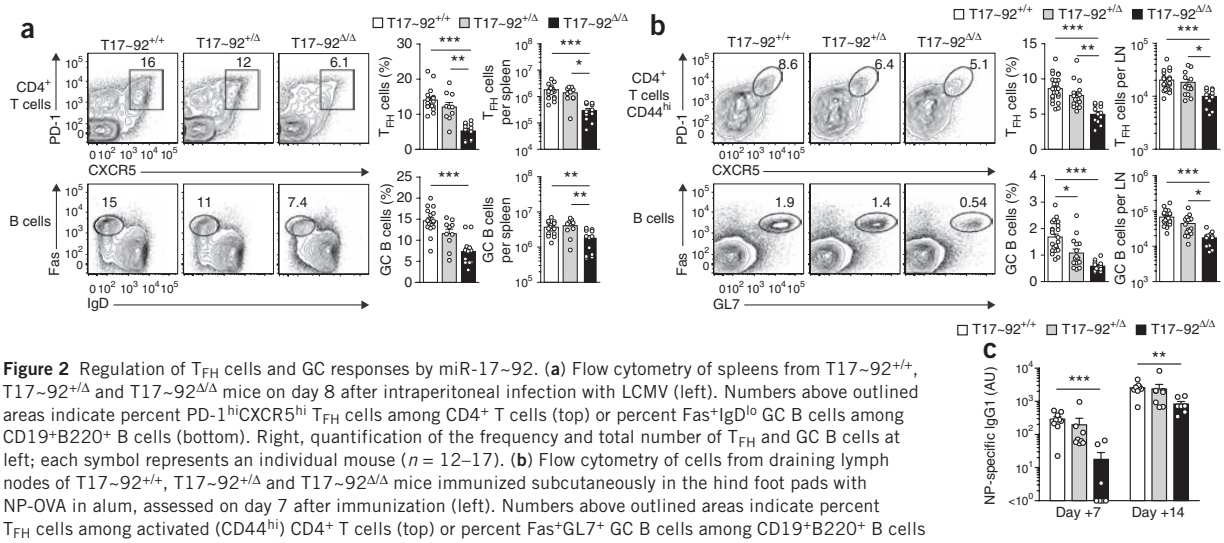


Figure 2 Regulation of T_{FH} cells and GC responses by miR-17~92. (**a**) Flow cytometry of spleens from T17~92$^{+/+}$, T17~92$^{+/\Delta}$ and T17~92$^{\Delta/\Delta}$ mice on day 8 after intraperitoneal infection with LCMV (left). Numbers above outlined areas indicate percent PD-1^{hi}CXCR5^{hi} T_{FH} cells among CD4$^+$ T cells (top) or percent Fas$^+$IgD^{lo} GC B cells among CD19$^+$B220$^+$ B cells (bottom). Right, quantification of the frequency and total number of T_{FH} and GC B cells at left; each symbol represents an individual mouse (n = 12–17). (**b**) Flow cytometry of cells from draining lymph nodes of T17~92$^{+/+}$, T17~92$^{+/\Delta}$ and T17~92$^{\Delta/\Delta}$ mice immunized subcutaneously in the hind foot pads with NP-OVA in alum, assessed on day 7 after immunization (left). Numbers above outlined areas indicate percent T_{FH} cells among activated (CD44^{hi}) CD4$^+$ T cells (top) or percent Fas$^+$GL7$^+$ GC B cells among CD19$^+$B220$^+$ B cells (bottom). Right, quantification of the results at left (as in **a**; n = 12–24 mice). (**c**) Concentration of NP-specific immunoglobulin G1 (IgG1) in serum from T17~92$^{+/+}$, T17~92$^{+/\Delta}$ and T17~92$^{\Delta/\Delta}$ mice immunized with NP-OVA in alum, presented as arbitrary units (AU); each symbol represents an individual mouse (n = 7–9). $^*P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$ (nonparametric Kruskal-Wallis test with Dunn's *post-hoc* test). Data are pooled from three independent experiments (**a**,**b**; mean and s.e.m.) or are representative of two independent experiments (**c**; mean and s.e.m.).

did T17~92$^{+/+}$ mice (**Supplementary Fig. 2a**,**b**). Although T17~92$^{\Delta/\Delta}$ mice also lacked miR-17~92 in cytotoxic CD8$^+$ T cells, which mediate LCMV clearance at this stage of disease, viral clearance was similar at day 8 after infection in T17~92$^{\Delta/\Delta}$ and T17~92$^{+/+}$ mice (data not shown). Thus, the impaired antiviral T_{FH} response was not an indirect consequence of diminished clearance of the virus. Of note, deletion of one copy of the miR-17~92 cluster in mice (CD4-Cre$^+$*Mirc1*$^{+/fl}$; called 'T17~92$^{+/\Delta}$' here) resulted in an intermediate phenotype (**Fig. 2a** and **Supplementary Fig. 2a**,**b**).

By using immunization with nitrophenyl (NP)-OVA protein as a second, noninfectious model, we confirmed that miR-17~92 expressed by T cells was required for T_{FH} cell differentiation and was indirectly required for the formation of GC B cells (**Fig. 2b**). Again, the deletion of one copy of the miR-17~92 cluster in T17~92$^{+/\Delta}$ mice resulted in an intermediate phenotype (**Fig. 2b**). In contrast to results obtained with the LCMV model, the cellularity of draining lymph nodes and the frequency of activated T cells were similar in T17~92$^{\Delta/\Delta}$ and T17~92$^{+/+}$ mice (**Supplementary Fig. 2c**,**d**), which indicated a specific defect in the generation of T_{FH} cells. That defect also resulted in delayed and significantly lower titers of NP-specific antibody (**Fig. 2c**), and we observed a similar trend for antibody responses to LCMV (**Supplementary Fig. 2e**). In summary, T cell–intrinsic miR-17~92 was required for optimal T_{FH} cell and GC responses, including the production of antigen-specific antibodies.

Robust T_{FH} cell differentiation depends on miR-17~92

Although overexpression of miRNAs of the miR-17~92 cluster promotes T cell proliferation[17–19], adoptively transferred naive OT-II cells derived from T17~92$^{\Delta/\Delta}$ donor mice (called '17~92$^{\Delta/\Delta}$ OT-II' here) or from T17~92$^{+/\Delta}$ donor mice (called '17~92$^{+/\Delta}$ OT-II' here) had only slightly less proliferation (17~92$^{\Delta/\Delta}$ OT-II) or unchanged proliferation (17~92$^{+/\Delta}$ OT-II) relative to that of their miR-17~92-sufficient control counterparts (OT-II cells derived from T17~92$^{+/+}$ donor mice; called '17~92$^{+/+}$ OT-II' here; **Fig. 3a** and **Supplementary Fig. 3a**). Polyclonal CD4$^+$ T cells from T17~92$^{\Delta/\Delta}$ mice also proliferated slightly less than control cells from T17~92$^{+/+}$ mice did when activated by costimulation with small amounts of antibody to CD28 (anti-CD28) *in vitro*. However, that defect was overcome by an increase in the amount of costimulation with anti-CD28 (**Supplementary Fig. 3b**). Thus, the miR-17~92 cluster was largely dispensable for T cell proliferation under these conditions, possibly due to partial compensation by the closely related miR-106a~363 and miR-106b~25 clusters.

In contrast, transferred miR-17~92-deficient (17~92$^{\Delta/\Delta}$ OT-II) cell populations had a much lower frequency and total number of Bcl-6^{hi}CXCR5$^+$ developing T_{FH} cells (**Fig. 3b**). Tracking the early generation of T_{FH} cells showed a differentiation defect that was independent of cell division. Upregulation of the expression of both Bcl-6 and CXCR5 was impaired in 17~92$^{\Delta/\Delta}$ OT-II cells, which resulted in a much smaller proportion of Bcl-6$^+$CXCR5$^+$ developing T_{FH} cells at each cell division than among their miR-17~92-sufficient (17~92$^{+/+}$ OT-II) control counterparts (**Fig. 3c**,**d** and **Supplementary Fig. 3c**). The defective T_{FH} differentiation of 17~92$^{\Delta/\Delta}$ OT-II cell populations was also reflected by their lower frequency of interleukin 21 (IL-21)-producing cells (**Fig. 3e**) and much greater frequency of dividing CXCR5$^-$ cells expressing the high affinity IL-2 receptor α-chain CD25 (**Fig. 3f**), which inhibits T_{FH} differentiation[20,21]. The generation of T_{FH} cells was also lower among 17~92$^{+/\Delta}$ OT-II cells (**Supplementary Fig. 3d**), which indicated that miRNAs of the miR-17~92 cluster were limiting factors for T_{FH} cell differentiation. In summary, 17~92$^{\Delta/\Delta}$ CD4$^+$ T cells had a T_{FH} cell–differentiation defect similar to that of cells that lack all miRNAs, which emphasized the prominent functional importance of this miRNA cluster.

Overexpression of miR-17~92 promotes T_{FH} cell generation

Consistent with the idea that T_{FH} cell differentiation depends on miR-17~92, adoptively transferred OT-II cells overexpressing the miR-17~92 cluster in the form of a human transgene (CD4-Cre$^+$*Rosa26-Mirc1*$^{tg/tg}$; called '17~92$^{tg/tg}$' here) showed enhanced generation of T_{FH} cells relative to that of OT-II cells with normal expression of miR-17~92, without substantially greater proliferation after immunization with NP-OVA (**Fig. 4a–d**). In unimmunized mice

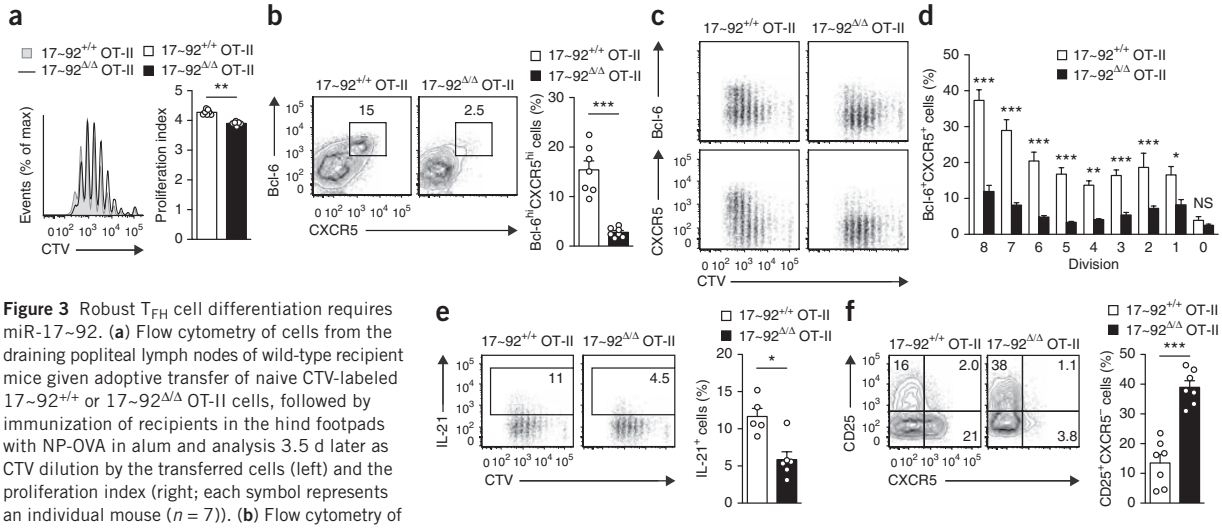


Figure 3 Robust T_{FH} cell differentiation requires miR-17~92. (**a**) Flow cytometry of cells from the draining popliteal lymph nodes of wild-type recipient mice given adoptive transfer of naive CTV-labeled 17~92$^{+/+}$ or 17~92$^{\Delta/\Delta}$ OT-II cells, followed by immunization of recipients in the hind footpads with NP-OVA in alum and analysis 3.5 d later as CTV dilution by the transferred cells (left) and the proliferation index (right; each symbol represents an individual mouse ($n = 7$)). (**b**) Flow cytometry of cells from recipient mice as in **a**; numbers above outlined areas indicate percent Bcl-6^{hi}CXCR5^{hi} T_{FH} cells among the transferred cells (left). Right, quantification of the results at left; each symbol represents an individual mouse ($n = 7$). (**c**) Flow cytometry of dividing OT-II cells from recipient mice as in **a**, showing the expression kinetics of Bcl-6 and CXCR5 at 3.5 d after immunization. (**d**) Frequency of Bcl-6^{+}CXCR5^{+} cells at each division, among OT-II cells from recipient mice as in **a** ($n = 7$). (**e**) Flow cytometry of cells from recipient mice from an experiment similar to that in **a** (left); numbers in outlined areas indicate percent IL-21-producing OT-II cells. Right, quantification of the results at left; each symbol represents an individual mouse ($n = 5$–6). (**f**) Flow cytometry of cells from recipient mice as in **a** (left); numbers in quadrants indicate percent CD25^{+}CXCR5^{-} OT-II cells (top left), CD25^{+}CXCR5^{+} OT-II cells (top right) or CD25^{-}CXCR5^{+} OT-II cells (bottom right). Right, quantification of CD25^{+}CXCR5^{-}cells; each symbol represents an individual mouse ($n = 7$). $^{*}P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$ (two-tailed nonparametric Mann-Whitney test (**a**–**c**,**e**,**f**) or two-way analysis of variance with Bonferroni's *post-hoc* test (**d**)). Data are representative of five (**a**–**d**,**f**) or two (**e**) independent experiments (mean and s.e.m. in **a**,**b**,**d**–**f**).

heterozygous for the miR-17~92-overexpressing transgene only in T cells (CD4-Cre^{+}*Mirc1*$^{tg/+}$; called 'T17~92$^{tg/+}$' here), the number of endogenous polyclonal T_{FH} cells was also much greater in Peyer's patches, with a correspondingly greater abundance of GC B cells, relative to that of their T17~92$^{+/+}$ counterparts with normal expression of miR-17~92 (**Fig. 4e**). Although T17~92$^{tg/+}$ mice had a generally greater abundance of total B cell and CD4^{+} T cell numbers, GC B cells and T_{FH} cell populations were 'preferentially' expanded. Finally, the abundance of CXCR5^{hi}PD-1^{hi}Foxp3^{+} follicular regulatory T cells (T_{reg} cells) correlated with the miR-17~92 'dose', but the abundance of polyclonal T_{reg} cells did not (**Supplementary Fig. 4**), which suggested that among all T_{reg} cells, the subset of T_{reg} cells located in GCs (follicular T_{reg} cells) were particularly sensitive to regulation by miR-17~92. Thus, artificially increasing the amount of miR-17~92 enhanced T_{FH} cell differentiation, and constitutive overexpression of miR-17~92 led to the accumulation of T_{FH} cells.

Repression of *Pten* by miR-17~92 early in T_{FH} cell differentiation

Pten is reported to be an important target of miR-17~92 that contributes to miR-17~92-overexpressing disease models of autoimmunity and lymphomagenesis[18,22,23]. We found higher PTEN expression in all responding 17~92$^{\Delta/\Delta}$ OT-II cells than in their 17~92$^{+/+}$ OT-II counterparts at 48 h after immunization of wild-type recipient mice with NP-OVA (**Supplementary Fig. 5a**), and especially in the first few cell divisions at later time points (**Supplementary Fig. 5b**). Conversely, 17~92$^{tg/tg}$ OT-II cells had lower PTEN expression than did their 17~92$^{+/+}$ OT-II counterparts (**Supplementary Fig. 5c**). To assess the functional relevance of miR-17~92-mediated repression of PTEN, we limited *Pten* to one allele. Deletion of one allele of *Pten* resulted in lower PTEN expression (**Supplementary Fig. 5d**) and partially restored induction of the expression of Bcl-6 and CXCR5 in early cell divisions of 17~92$^{\Delta/\Delta}$*Pten*$^{+/\Delta}$ OT-II cells (**Supplementary Fig. 5e**). However, 17~92$^{\Delta/\Delta}$*Pten*$^{+/\Delta}$ and 17~92$^{\Delta/\Delta}$*Pten*$^{+/+}$ OT-II cell populations had a similar frequency of Bcl-6^{+}CXCR5^{+} cells among those cells that had proliferated the most (**Supplementary Fig. 5e**), which suggested substantial contributions from additional targets.

Repression of T_{FH} cell–inappropriate genes by miR-17~92

Although the repression of individual miRNA target genes is generally modest, the aggregate biological effect can be large[24,25]. To obtain a sufficient number of T_{FH} cells for genome-wide transcript analysis, we transferred SMARTA CD4^{+} T cells, which have transgenic expression of a LCMV-specific TCR, derived from T17~92$^{+/+}$ donor mice (called '17~92$^{+/+}$ SM' here) or T17~92$^{\Delta/\Delta}$ donor mice (called '17~92$^{\Delta/\Delta}$ SM' here), into wild-type recipient mice, infected the recipients with LCMV and then purified donor-derived T_{FH} cells for microarray analysis (**Fig. 5a**). We recovered fewer 17~92$^{\Delta/\Delta}$ SM T_{FH} cells than 17~92$^{+/+}$ SM T_{FH} cells from infected recipient mice in this system, which was due to a proportional overall lower abundance of 17~92$^{\Delta/\Delta}$ SM cells (**Supplementary Fig. 6**). Genome-wide transcript analysis showed that as a group, expression of predicted mRNA targets[26] of each miRNA family in the miR-17~92 cluster was derepressed in 17~92$^{\Delta/\Delta}$ SM T_{FH} cells (**Fig. 5b**). In contrast, expression of predicted miR-29 targets shown before to be actively repressed by miR-29 in T cells[19] was unaffected in 17~92$^{\Delta/\Delta}$ SM T_{FH} cells. The group of mRNAs with moderately upregulated expression showed enrichment for predicted targets of miR-17~92 (**Supplementary Tables 1** and **2**). In addition, 17~92$^{\Delta/\Delta}$ SM T_{FH} cells expressed a recognizable set of T_{FH} cell–inappropriate genes, including *Ccr6*, *Il1r2*, *Il1r1*, *Rora*, and *Il22* (**Fig. 5c** and **Supplementary Table 1**). We confirmed higher expression of CCR6 and IL-1R2 protein by flow cytometry. Each had high expression in many 17~92$^{\Delta/\Delta}$ SM T_{FH} cells

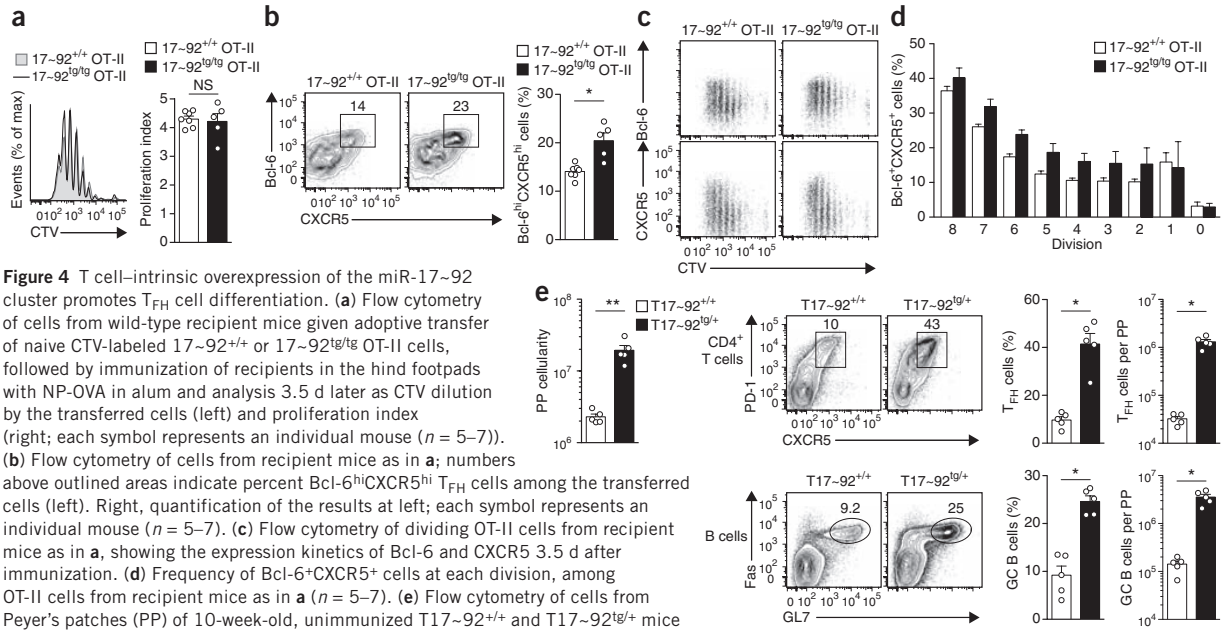


Figure 4 T cell–intrinsic overexpression of the miR-17~92 cluster promotes T_{FH} cell differentiation. (**a**) Flow cytometry of cells from wild-type recipient mice given adoptive transfer of naive CTV-labeled 17~92$^{+/+}$ or 17~92$^{tg/tg}$ OT-II cells, followed by immunization of recipients in the hind footpads with NP-OVA in alum and analysis 3.5 d later as CTV dilution by the transferred cells (left) and proliferation index (right; each symbol represents an individual mouse (n = 5–7)). (**b**) Flow cytometry of cells from recipient mice as in **a**; numbers above outlined areas indicate percent Bcl-6^{hi}CXCR5^{hi} T_{FH} cells among the transferred cells (left). Right, quantification of the results at left; each symbol represents an individual mouse (n = 5–7). (**c**) Flow cytometry of dividing OT-II cells from recipient mice as in **a**, showing the expression kinetics of Bcl-6 and CXCR5 3.5 d after immunization. (**d**) Frequency of Bcl-6^{+}CXCR5^{+} cells at each division, among OT-II cells from recipient mice as in **a** (n = 5–7). (**e**) Flow cytometry of cells from Peyer's patches (PP) of 10-week-old, unimmunized T17~92$^{+/+}$ and T17~92$^{tg/+}$ mice (with an equal number of Peyer's patches from each mouse, to allow comparison of total cellularity (far left)). Numbers above outlined areas (contour plots; middle) indicate PD-1^{hi}CXCR5^{hi} T_{FH} cells among CD19^{-}CD4^{+} T cells (top) or percent Fas^{+}GL7^{+} GC B cells among CD19^{+} B cells (bottom). Right, quantification of the results at left; each symbol represents an individual mouse (n = 5). $^{*}P < 0.05$ and $^{**}P < 0.01$ (two-tailed nonparametric Mann-Whitney test). Data are representative of four independent experiments (**a**–**d**) or two independent experiments (**e**; mean and s.e.m. in **a**,**b**,**d**,**e**).

but in only a few CXCR5^{-} 17~92$^{\Delta/\Delta}$ SM non-T_{FH} cells (**Fig. 5a**,**d**). Most of those non-T_{FH} cells were T-bet^{hi} T_H1 cells (data not shown). We confirmed additional gene dysregulation in T_{FH} cells by quantitative PCR. *Il1r1* and *Rora* were derepressed in 17~92$^{\Delta/\Delta}$ SM T_{FH} cells (**Fig. 5e**). *Ex vivo* restimulation of 17~92$^{+/+}$ SM and 17~92$^{\Delta/\Delta}$ SM cells also resulted in a greater proportion of IL-22^{+}IL-17A^{-} cells and, to a lesser extent, IL-22^{+}IL-17A^{+} cells, but not a greater proportion of cells producing only IL-17A in 17~92$^{\Delta/\Delta}$ SM than in their 17~92$^{+/+}$ SM counterparts (**Fig. 5f**). Thus, miR-17-92 repressed *Ccr6*, *Il1r2*, *Il1r1*, *Rora* and *Il22* during T_{FH} cell differentiation in infection with LCMV. However, it remained unclear if those genes were directly targeted by miR-17~92 or whether the observed dysregulation was an indirect effect.

Rora is a functionally relevant miR-17~92 target

Because the transcription factor RORα (encoded by *Rora*) is sufficient to induce IL-1R1 expression[27] and CCR6 expression[28], and IL-1R1 expression partially depends on RORα[27], we considered the possibility that unrestrained RORα expression may have accounted for part of the observed subset-inappropriate gene expression in 17~92$^{\Delta/\Delta}$ SM T_{FH} cells. The 3′ untranslated region of *Rora* has two clusters of predicted miRNA-binding sites, each with four conserved miR-17~92-binding sites (**Supplementary Fig. 7**). Transfection of 17~92$^{+/+}$ OT-II and 17~92$^{\Delta/\Delta}$ OT-II cells with luciferase reporter constructs showed that endogenous miR-17~92 repressed both clusters, whereas miR-17~92-overexpressing (17~92$^{tg/tg}$) OT-II cells had enhanced repression (**Fig. 6a**). We isolated the effect of each miRNA by transfecting polyclonal miRNA-deficient (*Dgcr8*$^{\Delta/\Delta}$) CD4^{+} T cells with reporter constructs and individual miRNA mimics. This analysis showed perfect correlation between target-site predictions and repressive activity of the corresponding miRNAs (**Fig. 6b**). We concluded that all four miRNA families of the miR-17~92 cluster contributed to robust inhibition of *Rora* expression.

To assess the functional relevance of miR-17~92-mediated repression of *Rora in vivo*, we limited *Rora* to one functional allele by intercrossing T17~92$^{\Delta/\Delta}$ SMARTA mice and staggerer mice, which have a spontaneously mutated allele (*Rora*^{sg}) that does not encode a functional RORα protein. *Rora* heterozygosity (*Rora*$^{+/sg}$) in 17~92$^{\Delta/\Delta}$ SM T_{FH} cells restored *Rora* mRNA to the abundance observed in 17~92$^{+/+}$ SM T_{FH} cells (**Supplementary Fig. 8**). Adoptive transfer of 17~92$^{\Delta/\Delta}$*Rora*$^{+/+}$ SM cells into wild-type mice, followed by infection of the recipients with LCMV, led to higher expression of CCR6 and IL-1R2 than that of their 17~92$^{+/+}$*Rora*$^{+/+}$ SM counterparts mainly in T_{FH} cells (**Fig. 6c**), which confirmed our results reported above (**Fig. 5d**). In contrast, many fewer 17~92$^{\Delta/\Delta}$*Rora*$^{+/sg}$ SM cells than 17~92$^{\Delta/\Delta}$*Rora*$^{+/+}$ SM cells had higher CCR6 expression than that of their 17~92$^{+/+}$*Rora*$^{+/+}$ SM counterparts (**Fig. 6c**). Thus, limiting *Rora* to one functional allele partially restored proper regulation of CCR6 despite the absence of miR-17~92. Notably, 17~92$^{+/+}$*Rora*$^{+/+}$ SM, 17~92$^{\Delta/\Delta}$*Rora*$^{+/+}$ SM and 17~92$^{\Delta/\Delta}$*Rora*$^{+/sg}$ SM T_{FH} cells showed no difference in expression of the closely related transcription factor RORγt, which can also induce CCR6 expression (**Fig. 6d**). Microarray experiments also indicated no difference between 17~92$^{+/+}$ SM and 17~92$^{\Delta/\Delta}$ SM T_{FH} cells in RORγt expression (data not shown). Thus, it is unlikely that RORγt was the driving force behind the dysregulated gene signature of 17~92$^{\Delta/\Delta}$ SM T_{FH} cells. IL-1R2 expression was not affected by limiting *Rora* to one functional allele (**Fig. 6c**). In contrast, the frequency of IL-22-producing cells among 17~92$^{\Delta/\Delta}$*Rora*$^{+/sg}$ SM cells was about half that of 17~92$^{\Delta/\Delta}$*Rora*$^{+/+}$ SM cells (**Fig. 6e**). We concluded that miR-17~92 was needed to directly repress *Rora* during T_{FH} cell differentiation to prevent subset-inappropriate gene expression.

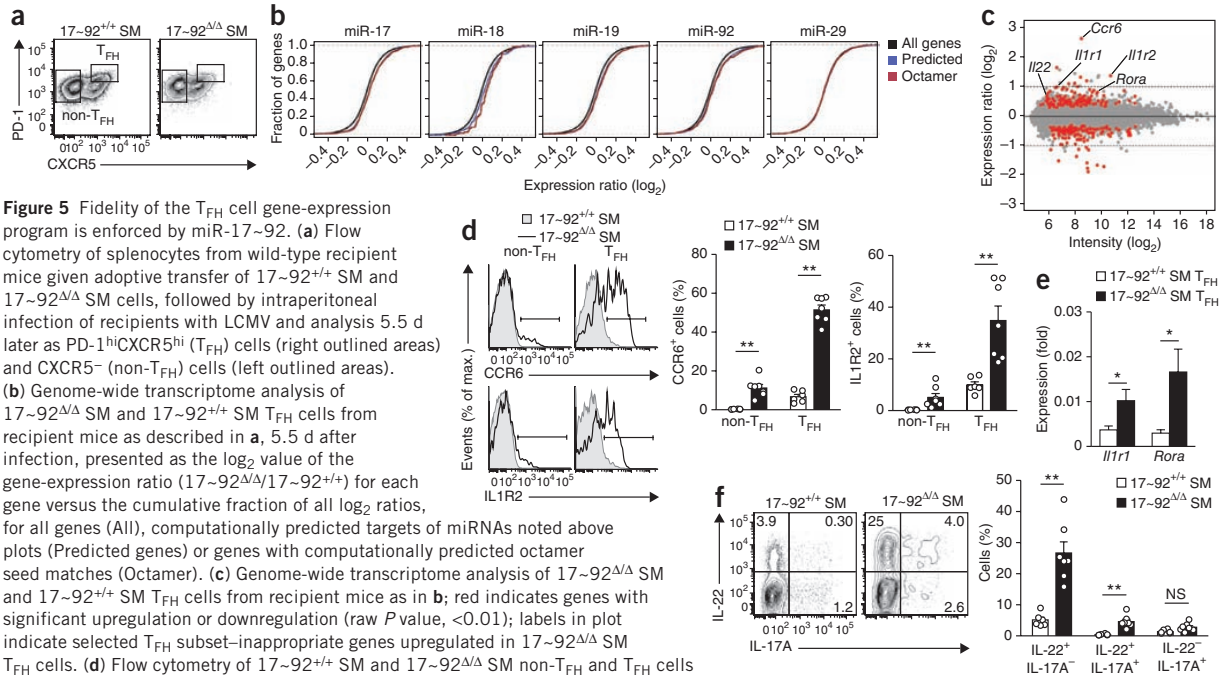


Figure 5 Fidelity of the T_{FH} cell gene-expression program is enforced by miR-17~92. (**a**) Flow cytometry of splenocytes from wild-type recipient mice given adoptive transfer of 17~92$^{+/+}$ SM and 17~92$^{\Delta/\Delta}$ SM cells, followed by intraperitoneal infection of recipients with LCMV and analysis 5.5 d later as PD-1^{hi}CXCR5^{hi} (T_{FH}) cells (right outlined areas) and CXCR5^{-} (non-T_{FH}) cells (left outlined areas). (**b**) Genome-wide transcriptome analysis of 17~92$^{\Delta/\Delta}$ SM and 17~92$^{+/+}$ SM T_{FH} cells from recipient mice as described in **a**, 5.5 d after infection, presented as the $\log_2$ value of the gene-expression ratio (17~92$^{\Delta/\Delta}$/17~92$^{+/+}$) for each gene versus the cumulative fraction of all $\log_2$ ratios, for all genes (All), computationally predicted targets of miRNAs noted above plots (Predicted genes) or genes with computationally predicted octamer seed matches (Octamer). (**c**) Genome-wide transcriptome analysis of 17~92$^{\Delta/\Delta}$ SM and 17~92$^{+/+}$ SM T_{FH} cells from recipient mice as in **b**; red indicates genes with significant upregulation or downregulation (raw *P* value, <0.01); labels in plot indicate selected T_{FH} subset–inappropriate genes upregulated in 17~92$^{\Delta/\Delta}$ SM T_{FH} cells. (**d**) Flow cytometry of 17~92$^{+/+}$ SM and 17~92$^{\Delta/\Delta}$ SM non-T_{FH} and T_{FH} cells (gated as in **a**), assessed 5.5 d after infection of mice as in **a** (left); bracketed lines indicate CCR6^{+} cells (top row) or IL-1R2^{+} cells (bottom row). Right, quantification of the results at left; each symbol represents an individual mouse (n = 6–7 per genotype). (**e**) Expression of *Il1r1* and *Rora* in 17~92$^{+/+}$ SM and 17~92$^{\Delta/\Delta}$ SM T_{FH} cells on day 5.5 in an experiment similar to that in **a** (n = 5 mice per genotype); results are presented relative to expression of the housekeeping gene *Hprt*. (**f**) Flow cytometry of 17~92$^{+/+}$ SM and 17~92$^{\Delta/\Delta}$ SM cells obtained from mice (n = 6–7) 5.5 d after infection as in **a**, then restimulated with the phorbol ester PMA and ionomycin (left); number in quadrants indicate percent cells in each. Right, quantification of IL-22- and/or IL-17A-producing cells. $^{*}P < 0.05$ and $^{**}P < 0.01$ (two-tailed nonparametric Mann-Whitney test). Data are representative of at least six (**a**) or two (**d**–**f**) independent experiments (mean and s.e.m. in **d**–**f**) or are pooled from four independent experiments (**b**,**c**).

DISCUSSION

Better understanding of the genetic programs that regulate the differentiation and plasticity of T_{FH} cells might lead to new strategies for rational vaccine design and suppression of antibody-mediated autoimmune diseases. Major advances delineating important roles for Bcl-6 and other proteins have been achieved[1]. In contrast, very little is known about the roles of miRNAs in T_{FH} cell differentiation. We found that miRNAs were absolutely critical for the differentiation and function of T_{FH} cells and that the miR-17~92 cluster in particular was required for robust T_{FH} cell responses in a T cell–intrinsic manner. Those SMARTA T_{FH} cells that did develop in the absence of miR-17~92 failed to suppress the direct target *Rora* and a 'suite' of other T_{FH} cell–inappropriate genes normally expressed by the T_H17 and T_H22 subsets of helper T cells. We conclude that miR-17~92 promotes T_{FH} cell differentiation and maintains the fidelity of T_{FH} cell identity by repressing non-T_{FH} cell genes both directly and indirectly. Together with published studies showing that miR-17~92 regulates the proliferation and survival of T cells[17–19], our findings indicate that miR-17~92 constitutes a central coordinator of the fate of activated T cells.

The global roles of miRNAs in T_{FH} cell responses have been difficult to study because of the substantial defects in the survival and proliferation of miRNA-deficient T cells. We overcame that roadblock by adoptive transfer of OVA-specific OT-II T cells and by using dilution of intravital dye to analyze the early stages of the T_{FH} differentiation of miRNA-sufficient and miRNA-deficient cells that had survived and divided the same number of times *in vivo*. This approach showed that miRNAs were essential for the earliest steps in the differentiation into T_{FH} cells, including upregulation of the expression of Bcl-6 and CXCR5, downregulation of the expression of CCR7 and migration to sites of interaction with B cells in secondary lymphoid organs. Those findings were in contrast to the requirement for miRNAs to restrain T_H1 differentiation[19,29] but were reminiscent of the requirement for miRNAs in supporting the differentiation and function of T_{reg} cells[30–32].

The same transfer system described above showed that miR-17~92 regulated various activities of T cells that are important for mounting effective humoral immune responses. We found an unexpectedly small effect of miR-17~92 on the proliferation of OT-II T cells *in vivo*, but it was required for optimal T_{FH} cell differentiation. Higher expression of *Pten*, a known direct target of miR-17~92 that encodes a regulator of T_{FH} cell responses[18,33], partially accounted for the defective generation of T_{FH} cells during the earliest cell divisions. Notably, costimulation via CD28 represses PTEN expression[34] and induces miR-17~92 expression in activated T cells[16,35]. We speculate that the required role of costimulation via CD28 in T_{FH} cell differentiation[36] may be mediated in part by induction of miR-17~92 expression and subsequent downregulation of PTEN expression. However, the effect of *Pten* regulation was barely detectable in this system, which indicates important roles for other direct targets of miR-17~92. Cell proliferation and the early induction of Bcl-6 expression that occurs in all activated T cells[6] was intact in the absence of miR-17~92. Those observations indicated that some of the relevant targets must affect T_{FH} cell differentiation itself rather than T cell activation in general.

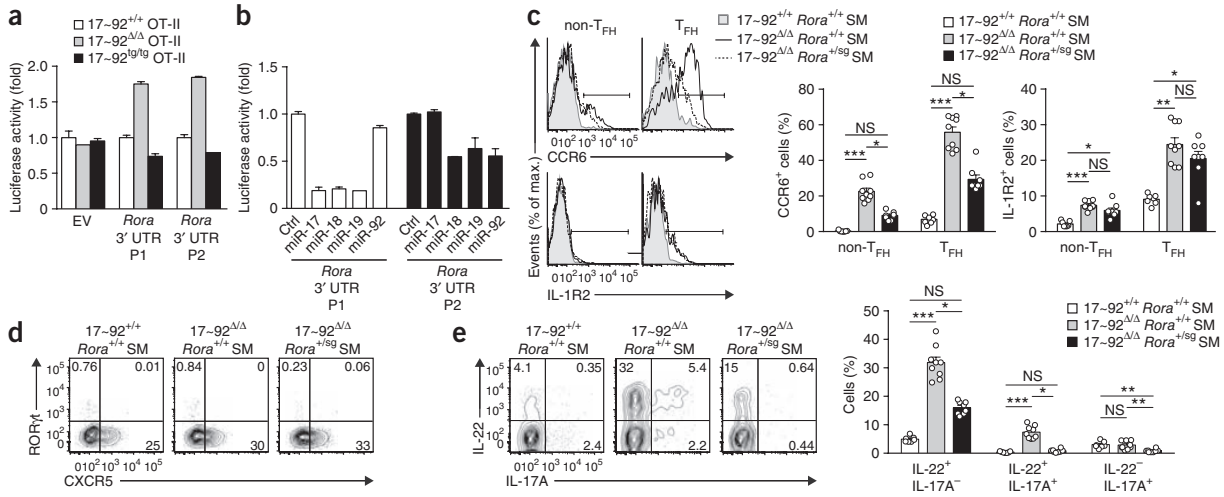


Figure 6 *Rora* is a functionally relevant target of miR-17~92. (**a**) Renilla luciferase activity of *in vitro*–stimulated 17~92$^{+/+}$ OT-II, 17~92$^{\Delta/\Delta}$ OT-II and 17~92$^{tg/tg}$ OT-II cells transfected with empty vector (EV; control) or with dual luciferase reporters for position 1 (P1) or position 2 (P2) of the 3′ untranslated (UTR) region of *Rora* (details, **Supplementary Fig. 7**), assessed 24 h after transfection and presented relative to firefly luciferase activity in transfected 17~92$^{+/+}$ OT-II cells. (**b**) Renilla luciferase activity in primary polyclonal *Dgcr8*$^{\Delta/\Delta}$ CD4$^+$ T cells transfected with luciferase reporters as in **a** together with miRNA mimics of the miR-17~92 cluster (horizontal axis), assessed 24 h after transfection; results were normalized to firefly luciferase activity and are presented relative to those of cells transfected with control miRNA mimics (Ctrl). (**c**) Flow cytometry of cells from wild-type recipient mice given adoptive transfer of naive 17~92$^{+/+}$*Rora*$^{+/+}$ SM, 17~92$^{\Delta/\Delta}$*Rora*$^{+/+}$ SM or 17~92$^{\Delta/\Delta}$*Rora*$^{+/sg}$ SM cells, followed by intraperitoneal infection of recipients with LCMV and analysis 5.5 d later (left); bracketed lines indicate CCR6$^+$ cells (top row) or IL-1R2$^+$ cells (bottom row) among non-T$_{FH}$ and T$_{FH}$ donor cells. Right, quantification of the results at left; each symbol represents an individual mouse (n = 7–9 per genotype). (**d**) Expression of RORγt and CXCR5 by 17~92$^{+/+}$*Rora*$^{+/+}$, 17~92$^{\Delta/\Delta}$*Rora*$^{+/+}$ or 17~92$^{\Delta/\Delta}$*Rora*$^{+/sg}$ SM cells; number in quadrants indicate percent cells in each. (**e**) Flow cytometry of cells obtained from recipients of 17~92$^{+/+}$*Rora*$^{+/+}$ SM, 17~92$^{\Delta/\Delta}$*Rora*$^{+/+}$ SM or 17~92$^{\Delta/\Delta}$*Rora*$^{+/sg}$ SM cells 5.5 d after infection as in **c** and then restimulated with PMA and ionomycin (left); number in quadrants indicate percent cells in each. Right, quantification of IL-22- and/or IL-17A-producing cells; each symbol represents an individual mouse (n = 7–9 mice per genotype). *P < 0.05, **P < 0.01 and ***P < 0.001 (nonparametric Kruskal-Wallis test with Dunn's *post-hoc* test). Data are representative of three independent experiments (**a**) or two independent experiments (**b**–**e**; mean and s.e.m. in **a**–**c**,**e**).

In contrast, the further upregulation of Bcl-6 expression in dividing cells characteristic of T$_{FH}$ cells was considerably blunted, and the induction of CXCR5 expression was almost completely abrogated. We also observed specific effects on T$_{FH}$ cell differentiation that could be distinguished from general activation defects during LCMV infection. The 17~92$^{\Delta/\Delta}$ SM T$_{FH}$ cells acquired an inappropriate gene-expression program reminiscent of T$_H$17 or T$_H$22 cells[4,37], including upregulation of the expression of genes encoding RORα, CCR6 and components of the IL-1 pathway (IL-1R1 and IL-1R2), and inducible production of IL-22. T$_H$17 and T$_H$22 cells are closely related helper T cell subsets with many shared features (such as CCR6 expression) but also distinct features[38,39]. However, 17~92$^{\Delta/\Delta}$ SM T$_{FH}$ cells did not convert into RORγt$^+$ and IL-17-producing T$_H$17 cells or become proper T$_H$22 cells. Instead, they maintained expression of genes of the T$_{FH}$ cell program, including the markers CXCR5, Bcl-6 and PD-1, and became a hybrid cell type with molecular features of more than one helper T cell subset. We conclude that T$_{FH}$ cells need miR-17~92 to repress inappropriate, non–T$_{FH}$ cell gene-expression programs. This requirement was selective for T$_{FH}$ cells, as expression of CCR6 and IL-1R2 was affected much less in non-T$_{FH}$ cells (which are mostly T$_H$1 cells) in the same infected spleens.

We noted that miR-17~92 deficiency did not affect the frequency of SMARTA T$_{FH}$ cells but did result in a significantly lower total number of both T$_{FH}$ cells and T$_H$1 cells in infected spleens. Thus, infection with LCMV, which induces extremely robust T cell population expansion, resulted in a defect in antigen-driven helper T cell proliferation *in vivo* predicted by published *in vitro* studies[18,19], whereas the more slowly proliferating OT-II T cells manifested a more selective defect in T$_{FH}$ cell differentiation. Polyclonal responses in T17~92$^{\Delta/\Delta}$ mice also differed in the magnitude of the defect in the frequency and number of T$_{FH}$ cells (affected more during infection with LCMV) and GC B cells (affected more after immunization with OVA). Compromised function of CD8$^+$ T cells, which also lack miR-17~92 in these mice, may indirectly affect these responses, particularly in the case of infection with LCMV[40].

CD4$^+$ helper T cell 'plasticity' has garnered considerable attention recently. The present models of T cell differentiation suggest that cell identity is less rigid than previously thought[41]. Although Bcl-6 has been identified as a subset-defining transcription factor required for T$_{FH}$ cell differentiation[3–5], it remains controversial whether or not T$_{FH}$ cells represent a stable cell lineage[1]. A new model suggests that initial helper T cell differentiation proceeds via a Bcl-6$^+$ pre–helper T cell stage with concurrent upregulation of expression of the subset-defining transcription factors T-bet, GATA-3 and/or RORγt[42]. According to this model, T$_H$1-, T$_H$2- or T$_H$17-differentiation cues downregulate Bcl-6 expression and further upregulate expression of the subset-defining factors. In contrast, higher Bcl-6 expression and suppression of RORγt, GATA-3 and T-bet yields T$_{FH}$ cells. As concomitant expression of competing transcription programs is common, repression of genes encoding molecules that lead to alternative cell fates is an important requirement during T cell differentiation[10,43]. Individual miRNAs can be powerful enough to shift a cell's transcriptome to that of a different cell type[44], and they maintain the fidelity of cell type–specific transcriptomes by repressing genetic programs of other cell lineages[45]. Deficiency in miRNAs induces proinflammatory cytokine secretion in T$_{reg}$ cells even when they continue to

express the transcription factor Foxp3 (ref. 30). The miRNA miR-10a may restrict the plasticity of several subsets of helper T cells, including both T_{reg} cells and T_{FH} cells, and may influence T_H17 differentiation[15,46]. The miRNA miR-29 prevents aberrant activation of the T_H1 program by repressing both T-bet and its homolog eomesodermin, which is usually not expressed in $CD4^+$ T cells[19]. In this study, we found that all four miRNA families in the miR-17~92 cluster targeted *Rora* to prevent expression of the gene encoding CCR6 and other genes associated with T_H17 or T_H22 cells. Thus, a paradigm is emerging in which miRNAs help to define and maintain cell identity by repressing alternative gene-expression programs and thus effectively limiting the plasticity of differentiating T cells.

Notably, although a dichotomy between the T_{FH}- and T_H17-differentiation pathways has been proposed[4], T_H17 cells can acquire a T_{FH} cell phenotype under certain conditions in Peyer's patches[47]. The unexpected identification of *Rora* as a direct target of miR-17~92 and of RORα as a functionally relevant contributor to T_{FH} cell–inappropriate gene expression suggests that differentiating (pre-)T_{FH} cells receive signals that induce *Rora* transcription but that miR-17~92 renders such induction inconsequential. Additional studies are needed to identify those signals and to determine whether miR-17~92 controls *Rora* expression in other cell types as well. A 'lineage-defining' transcription factor has not been identified for T_H22 cells, but we note that despite their similarity to T_H17 cells, human T_H22 cells do not require RORγt expression[37]. In line with that, RORγt expression was not affected in the hybrid T_{FH}-T_H17-T_H22 signature we noted in $17\sim92^{\Delta/\Delta}$ SM T_{FH} cells. Moreover, the heterogeneity of T_H17 cells poses specific challenges, and certain types of T_H17 cells might be more closely related to T_H22 cells than to conventional T_H17 cells[39]. Thus, the distinct functions of RORα and RORγt in differentiating T cells might need to be revisited. In addition, it remains uncertain whether altered migration, response to cytokines or some other trait of RORα-expressing cells limits the differentiation and function of T_{FH} cells *in vivo*. Additional studies are needed to define miR-17~92 function in the determination of early T_{FH} cell fate and to delineate cell-intrinsic effects on the molecular program from secondary effects due to altered abilities to sense the environment. Finally, we note that our data demonstrated that regulation of RORα only partially explained the hybrid gene-expression profile of miR-17~92-deficient T_{FH} cells. Additional direct targets relevant to this phenotype must exist and remain to be discovered. Nevertheless, our findings indicate that miR-17~92 and Bcl-6 act together to 'imprint' and protect the identity of developing T_{FH} cells by repressing differentiation into alternative T cell subsets.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession code. GEO: microarray data, GSE42760.

Note: Any Supplementary Information and Source Data are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Panduro for technical assistance; R. Blelloch (University of California, San Francisco) for *Dgcr8*$^{fl/fl}$ mice; R. Barbeau, J. Pollack, A. Barczak and D. Erle for assistance with microarray experiments; the 'miRNA in lymphocytes interest group' of the University of California, San Francisco, for discussions; D. Fuentes for animal husbandry; and D. Le for help with genotyping. Supported by the Burroughs Wellcome Fund (CABS 1006173 to K.M.A.), the US National Institutes of Health (R01 HL109102 and P01 HL107202 to K.M.A.; P01 AI35297 and U19 AI056388 to J.A.B.; and P30 DK63720 for core support), Juvenile Diabetes Research Foundation (J.A.B.), the Swiss National Science Foundation (PBBEP3-133516 to D.B.), the Swiss Foundation for Grants in Biology and Medicine (PASMP3-142725 to D.B.; and PASMP3-124274/1 to L.T.J.), the National Science Foundation (J.M.C.) and the Wellcome Trust (O.B.).

AUTHOR CONTRIBUTIONS

D.B. did and analyzed most of the experiments; R.K., J.M.C., M.M.M., S.P., D.d.K., O.B., M.M. and L.T.J. did and analyzed some of the experiments; J.A.B. interpreted the data; D.B., K.M.A. and L.T.J. designed the experiments, interpreted the data, and wrote the manuscript; and all authors discussed the results and commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Crotty, S. Follicular helper CD4 T cells (TFH). *Annu. Rev. Immunol.* **29**, 621–663 (2011).
2. Vinuesa, C.G., Sanz, I. & Cook, M.C. Dysregulation of germinal centres in autoimmune disease. *Nat. Rev. Immunol.* **9**, 845–857 (2009).
3. Johnston, R.J. *et al.* Bcl6 and Blimp-1 are reciprocal and antagonistic regulators of T follicular helper cell differentiation. *Science* **325**, 1006–1010 (2009).
4. Nurieva, R.I. *et al.* Bcl6 mediates the development of T follicular helper cells. *Science* **325**, 1001–1005 (2009).
5. Yu, D. *et al.* The transcriptional repressor Bcl-6 directs T follicular helper cell lineage commitment. *Immunity* **31**, 457–468 (2009).
6. Baumjohann, D., Okada, T. & Ansel, K.M. Cutting edge: distinct waves of BCL6 expression during T follicular helper cell development. *J. Immunol.* **187**, 2089–2092 (2011).
7. Ansel, K.M., McHeyzer-Williams, L.J., Ngo, V.N., McHeyzer-Williams, M.G. & Cyster, J.G. *In vivo*-activated CD4 T cells upregulate CXC chemokine receptor 5 and reprogram their response to lymphoid chemokines. *J. Exp. Med.* **190**, 1123–1134 (1999).
8. Yusuf, I. *et al.* Germinal center T follicular helper cell IL-4 production is dependent on signaling lymphocytic activation molecule receptor (CD150). *J. Immunol.* **185**, 190–202 (2010).
9. Baumjohann, D. *et al.* Persistent antigen and germinal center B cells sustain T follicular helper cell responses and phenotype. *Immunity* **38**, 596–605 (2013).
10. Oestreich, K.J. & Weinmann, A.S. Master regulators or lineage-specifying? Changing views on $CD4^+$ T cell transcription factors. *Nat. Rev. Immunol.* **12**, 799–804 (2012).
11. Nakayamada, S. *et al.* Early Th1 cell differentiation is marked by a Tfh cell-like transition. *Immunity* **35**, 919–931 (2011).
12. Pepper, M., Pagan, A.J., Igyarto, B.Z., Taylor, J.J. & Jenkins, M.K. Opposing signals from the Bcl6 transcription factor and the interleukin-2 receptor generate T helper 1 central and effector memory cells. *Immunity* **35**, 583–595 (2011).
13. Weber, J.P., Fuhrmann, F. & Hutloff, A. T-follicular helper cells survive as long-term memory cells. *Eur. J. Immunol.* **42**, 1981–1988 (2012).
14. Xiao, C. & Rajewsky, K. MicroRNA control in the immune system: basic principles. *Cell* **136**, 26–36 (2009).
15. Takahashi, H. *et al.* TGF-beta and retinoic acid induce the microRNA miR-10a, which targets Bcl-6 and constrains the plasticity of helper T cells. *Nat. Immunol.* **13**, 587–595 (2012).
16. Bronevetsky, Y. *et al.* T cell activation induces proteasomal degradation of Argonaute and rapid remodeling of the microRNA repertoire. *J. Exp. Med.* **210**, 417–432 (2013).
17. Jiang, S. *et al.* Molecular dissection of the miR-17–92 cluster's critical dual roles in promoting Th1 responses and preventing inducible Treg differentiation. *Blood* **118**, 5487–5497 (2011).
18. Xiao, C. *et al.* Lymphoproliferative disease and autoimmunity in mice with increased miR-17–92 expression in lymphocytes. *Nat. Immunol.* **9**, 405–414 (2008).
19. Steiner, D.F. *et al.* MicroRNA-29 regulates T-box transcription factors and interferon-gamma production in helper T cells. *Immunity* **35**, 169–181 (2011).
20. Ballesteros-Tato, A. *et al.* Interleukin-2 inhibits germinal center formation by limiting T follicular helper cell differentiation. *Immunity* **36**, 847–856 (2012).
21. Johnston, R.J., Choi, Y.S., Diamond, J.A., Yang, J.A. & Crotty, S. STAT5 is a potent negative regulator of TFH cell differentiation. *J. Exp. Med.* **209**, 243–250 (2012).
22. Ventura, A. *et al.* Targeted deletion reveals essential and overlapping functions of the miR-17 through 92 family of miRNA clusters. *Cell* **132**, 875–886 (2008).
23. Olive, V. *et al.* miR-19 is a key oncogenic component of mir-17–92. *Genes Dev.* **23**, 2839–2849 (2009).
24. Ebert, M.S. & Sharp, P.A. Roles for microRNAs in conferring robustness to biological processes. *Cell* **149**, 515–524 (2012).
25. Mendell, J.T. & Olson, E.N. MicroRNAs in stress signaling and human disease. *Cell* **148**, 1172–1187 (2012).
26. Friedman, R.C., Farh, K.K., Burge, C.B. & Bartel, D.P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19**, 92–105 (2009).

27. Chung, Y. *et al.* Critical regulation of early Th17 cell differentiation by interleukin-1 signaling. *Immunity* **30**, 576–587 (2009).
28. Yamazaki, T. *et al.* CCR6 regulates the migration of inflammatory and regulatory T cells. *J. Immunol.* **181**, 8391–8401 (2008).
29. Muljo, S.A. *et al.* Aberrant T cell differentiation in the absence of Dicer. *J. Exp. Med.* **202**, 261–269 (2005).
30. Zhou, X. *et al.* Selective miRNA disruption in T reg cells leads to uncontrolled autoimmunity. *J. Exp. Med.* **205**, 1983–1991 (2008).
31. Liston, A., Lu, L.F., O'Carroll, D., Tarakhovsky, A. & Rudensky, A.Y. Dicer-dependent microRNA pathway safeguards regulatory T cell function. *J. Exp. Med.* **205**, 1993–2004 (2008).
32. Chong, M.M., Rasmussen, J.P., Rudensky, A.Y. & Littman, D.R. The RNAseIII enzyme Drosha is critical in T cells for preventing lethal inflammatory disease. *J. Exp. Med.* **205**, 2005–2017 (2008).
33. Rolf, J. *et al.* Phosphoinositide 3-kinase activity in T cells regulates the magnitude of the germinal center reaction. *J. Immunol.* **185**, 4042–4052 (2010).
34. Buckler, J.L., Walsh, P.T., Porrett, P.M., Choi, Y. & Turka, L.A. Cutting edge: T cell requirement for CD28 costimulation is due to negative regulation of TCR signals by PTEN. *J. Immunol.* **177**, 4262–4266 (2006).
35. de Kouchkovsky, D. *et al.* miR-17~92 regulates interleukin-10 production by Tregs and control of experimental autoimmune encephalomyelitis. *J. Immunol.* (in the press).
36. Linterman, M.A. & Vinuesa, C.G. Signals that influence T follicular helper cell differentiation and function. *Semin. Immunopathol.* **32**, 183–196 (2010).
37. Duhen, T., Geiger, R., Jarrossay, D., Lanzavecchia, A. & Sallusto, F. Production of interleukin 22 but not interleukin 17 by a subset of human skin-homing memory T cells. *Nat. Immunol.* **10**, 857–863 (2009).
38. Rutz, S., Eidenschenk, C. & Ouyang, W. IL-22, not simply a Th17 cytokine. *Immunol. Rev.* **252**, 116–132 (2013).
39. Basu, R., Hatton, R.D. & Weaver, C.T. The Th17 family: flexibility follows function. *Immunol. Rev.* **252**, 89–103 (2013).
40. Wu, T. *et al.* Temporal expression of microRNA cluster miR-17–92 regulates effector and memory $CD8^+$ T-cell differentiation. *Proc. Natl. Acad. Sci. USA* **109**, 9965–9970 (2012).
41. O'Shea, J.J. & Paul, W.E. Mechanisms underlying lineage commitment and plasticity of helper $CD4^+$ T cells. *Science* **327**, 1098–1102 (2010).
42. Ma, C.S., Deenick, E.K., Batten, M. & Tangye, S.G. The origins, function, and regulation of T follicular helper cells. *J. Exp. Med.* **209**, 1241–1253 (2012).
43. Zhu, J., Yamane, H. & Paul, W.E. Differentiation of effector CD4 T cell populations. *Annu. Rev. Immunol.* **28**, 445–489 (2010).
44. Lim, L.P. *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769–773 (2005).
45. Stark, A., Brennecke, J., Bushati, N., Russell, R.B. & Cohen, S.M. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell* **123**, 1133–1146 (2005).
46. Jeker, L.T. *et al.* MicroRNA 10a marks regulatory T cells. *PLoS ONE* **7**, e36684 (2012).
47. Hirota, K. *et al.* Plasticity of TH17 cells in Peyer's patches is responsible for the induction of T cell-dependent IgA responses. *Nat. Immunol.* **14**, 372–379 (2013).

ONLINE METHODS

Mice. OT-II mice (004194), mice with *lox*P-flanked alleles encoding miR-17~92 (008458), *Rosa26*-miR-17~92–transgenic mice (008517), *Rosa26*-EYFP–transgenic mice (006148) and mice heterozygous for the *Rora*^{sg} mutation (002651) were from The Jackson Laboratory. CD4-Cre mice (4196) were from Taconic. OT-II mice and SMARTA mice[48] were crossed with B6.SJL-Ptprca Pepcb/BoyJ mice (002014) to obtain offspring with congenic alleles encoding CD45. Mice with *lox*P-flanked *Dgcr8* alleles[49] were provided by R. Blelloch. C57BL/6 (The Jackson Laboratory) or congenic B6-LY5.2/Cr (National Cancer Institute) mice were used as recipients. Mice with *lox*P-flanked *Pten* alleles have been described[50]. All experiments were done according to the Institutional Animal Care and Use Committee guidelines of the University of California, San Francisco.

Adoptive cell transfer, infection and immunization. Cell suspensions from spleens and lymph nodes were pre-enriched for OT-II and SMARTA cells with a CD4$^+$ negative isolation kit (Invitrogen), and naive T cells (CD4$^+$CD8$^-$CD25$^-$CD44^{lo}CD62L^{hi}) were further purified on a FACSAria II (BD Biosciences). To obtain true *Dgcr8*-deficient OT-II cells, naive cells were additionally sorted according to expression of a yellow fluorescent protein reporter driven by the ubiquitous *Rosa26* promoter, in which efficient excision of a *lox*P-flanked stop cassette by Cre recombinase activity driven by the *Cd4* promoter results in a bright yellow fluorescent signal. For cell proliferation experiments, naive T cells were labeled with 5 μM CellTrace Violet (Invitrogen) as described[6]. NP_{18}-OVA (18 molecules of nitrophenyl linked to OVA; Biosearch Technologies) was mixed with Imject Alum (Pierce) and 5 μg NP_{18}-OVA were injected subcutaneously into each hind footpad or 50 μg were injected subcutaneously in the base of tail and flank. In some experiments, mice were infected intraperitoneally with LCMV, Armstrong strain (2×10^5 plaque-forming units).

Flow cytometry. Spleen and lymph node cells were gently disrupted between the frosted ends of microscope slides and single-cell suspensions were filtered through fine mesh. T_{FH} cells were stained as described[51]. Antibodies were as follows: anti-CD4 (RM4-5), anti-CD8α (53-6.7), anti-CD19 (1D3), anti-CD25 (PC61.5), anti-CD45.1 (A20), anti-CD44 (IM7), anti-CD62L (MEL-14), anti-B220 (RA3-6B2), anti-GL-7, anti-IgD (11-26c), anti-IL-17A (eBio17B7), anti-IL-22 (1H8PWSR), anti-PD-1 (J43 or RMP1-30) and anti-Foxp3 (FJK-16s; all from eBioscience); anti-IL-1R2 (4E2), anti-Fas (Jo2) and anti-CCR6 (140706; all from BD Biosciences); and anti-CD45.2 (104; BioLegend). Nonspecific binding was blocked with anti-CD16/CD32 (2.4G2; Cell Culture Facility of the University of California, San Francisco) plus 2% normal mouse serum and 2% normal rat serum. Biotinylated anti-CXCR5 (2G8; BD Biosciences) was visualized with streptavidin-allophycocyanin (eBioscience) or streptavidin–Brilliant Violet 421 (Biolegend). Staining for 30 min at 37 °C with biotinylated anti-CCR7 (4B12; eBioscience) was followed by regular surface staining, including streptavidin-allophycocyanin at 4 °C. Monoclonal antibody (mAb) to Bcl-6 (K112-91), mAb to PTEN (A2B1) and mAb to RORγt (Q31-378) were from BD Biosciences. Intracellular Bcl-6, Foxp3 and RORγt were stained with the Foxp3 Staining Set (eBioscience). Cytofix Fixation Buffer and Perm Buffer III (BD) were used for intracellular staining of PTEN. For intracellular cytokine staining, lymph node or spleen cells were stimulated for 4 h with PMA (phorbol 12-myristate 13-acetate) and ionomycin (both from Fisher Scientific), with the addition of brefeldin A (Sigma-Aldrich) for the final 2 h. Cells were fixed with 4% paraformaldehyde, followed by permeabilization with saponin (Sigma-Aldrich). A chimera of mouse IL-21R and human crystallizable fragment (IL-21R–Fc; R&D Systems) was detected with phycoerythrin-labeled $F(ab')_2$ fragments specific to the Fc region of human IgG (Jackson ImmunoResearch). Samples were acquired on a LSR II cytometer (BD Biosciences) and were analyzed with FlowJo software (Tree Star), with gating out of doublets as well as non-T cells or non-B cells, where appropriate, in a dump channel. Dead cells were excluded with 7-aminoactinomycin D (eBioscience) or Fixable Viability Dye eFluor780 (eBioscience).

***In vitro* costimulation and proliferation assay.** Naive T cells from control and T17~92$^{Δ/Δ}$ mice were activated for 48 h and 72 h *in vitro* with plate-bound anti-CD3 (2C11; produced in-house) and anti-CD28 (PV1; produced in-house). Cells were labeled with CFSE (carboxyfluorescein diacetate succinimidyl ester) as described[52]. Proliferation was analyzed with the proliferation analysis function in FlowJo for Mac V9.2 and higher. To normalize for interexperimental differences, we normalized all data to the control in the first experiment (defined as a proliferative index of 1).

RNA extraction and quantitative real-time PCR. RNA extraction and quantitative PCR analysis of miRNAs was done as described[46].

Microarray. Naive SMARTA cells purified by flow cytometry from T17~92$^{+/+}$ or T17~92$^{Δ/Δ}$ donor mice were adoptively transferred into wild-type mice. Recipients were infected intraperitoneally with LCMV, Armstrong strain, and spleens were dissected 5.5 d later. Spleen cells were pooled for each condition (n = 3–5 mice) and samples were enriched for CD4$^+$ T cells with a CD4$^+$ negative isolation kit (Invitrogen), and congenically marked SMARTA T_{FH} cells (7AAD$^-$CD4$^+$CD8$^-$CD19$^-$CXCR5^{hi}PD-1^{hi}) were sorted directly into Trizol LS reagent and stored at –80 °C until further processing. RNA from four independent experiments was purified with RNeasy columns (Qiagen). Sample preparation, labeling and array hybridizations were done according to standard protocols from the Shared Microarray Core Facilities of the University of California, San Francisco, and Agilent Technologies. Total RNA quality was assessed with a Pico Chip on an Agilent 2100 Bioanalyzer (Agilent Technologies). RNA was amplified with Sigma whole transcriptome amplification kits according to the manufacturer's protocol (Sigma-Aldrich), and subsequent labeling of CTP with indocarbocyanine was done with NimbleGen one-color labeling kits (Roche-NimbleGen). Indocarbocyanine-labeled cDNA was assessed with the Nanodrop ND-8000 (Nanodrop Technologies), and equal amounts of indocarbocyanine-labeled target were hybridized to Agilent whole-mouse genome 8X60K in-jet arrays. Hybridizations were done for 17 h, according to the manufacturer's protocol. Arrays were scanned with the Agilent microarray scanner and raw signal intensities were extracted with Feature Extraction v10.6 software.

Cloning of the 3′ untranslated region of *Rora*, T cell transfection and luciferase assay. Two different constructs of the 3′ untranslated region of *Rora* were cloned into the psiCHECK-2 luciferase reporter construct (Promega) as described in **Supplementary Fig. 7**. Primer sequences were: P1 F: 5′-TAGTAGCTCGAGATGTCGCGCCCGAGCACTTC-3′; P1 R: 5′-TAGTAGGCGGCCGCAAACAGCAGCATAAATACCTCCCAACG-3′; P2 F: 5′-TAGTAGCTCGAGCCCCCAAAGTCTTTAACATCCTGA-3′; P2 R: 5′-TAGTAGGCGGCCGCAGTCAACCATAAGGTGCTTATTACTATTA-3′.) Transfection of T cells and luciferase assays were done as described[13]. CD4$^+$ T cells from spleen and lymph nodes were isolated by magnetic bead selection (Dynal) and were stimulated with anti-CD3 and anti-CD28. Cells were transfected with the Neon electroporation transfection system (Invitrogen). The miRIDIAN miRNA mimics miR-17, miR-18a, miR-19a and miR-92a and controls were from Dharmacon. Activated CD4$^+$ T cells were transfected with reporter constructs, and luciferase activity was measured 24 h after transfection with the Dual Luciferase Reporter Assay System (Promega) and a FLUOstar Optima plate-reader (BMG Labtech).

Enzyme-linked immunosorbent assay. First, 96-well half-area plates (Costar) were coated overnight at 4 °C with 10 μg/ml NP_{24}-BSA (Biosearch Technologies) in PBS. Nonspecific binding to plates was blocked with 1% BSA in PBS; serial dilutions of serum were incubated at 21 °C; horseradish peroxidase–conjugated polyclonal antibody to mouse IgG1 (1070–05; Southern Biotech) and Super AquaBlue ELISA Substrate (eBioscience) were used as the detection reagents. Absorbance was measured at 410 nm with a FLUOstar Optima plate-reader (BMG Labtech). Absolute values were calculated according to reference serum from hyperimmunized mice and results are presented in arbitrary units. For measurement of LCMV-specific antibodies, plates were coated with lysates of LCMV-infected baby hamster kidney cells. After blockade of nonspecific binding with 10% FBS in PBS, serially diluted serum was added. Horseradish peroxidase–conjugated polyclonal antibody to mouse IgG (Southern Biotech) was used as detection antibody, with 3,3′,5,5′-tetramethylbenzidine as the substrate. Antibody titers were determined as the reciprocal of the dilution that gave an absorbance value (at 450 nm) of more than twofold above that of naive control serum.

doi:10.1038/ni.2642

Immunohistochemistry. Draining popliteal lymph nodes were dissected, embedded in Tissue-Tek optimum cutting temperature compound (Sakura Finetek) and 'flash-frozen' in liquid nitrogen. Frozen tissues were stored at −80 °C until further processing. Cryosections (7 µm in thickness) were air-dried for 1 h before and after fixation in cold acetone for 10 min, then were rehydrated for 10 min in Tris-buffered saline (TBS), pH 7.6, containing 0.1% BSA. Slides were stained for 3 h at 20–25 °C in a humidified chamber in TBS containing 0.1% BSA, 1% normal mouse serum and 1% normal rat serum with a mixture of the following diluted primary antibodies: fluorescein isothiocyanate–conjugated anti-CD45.2 (104; Biolegend) and polyclonal goat antibody to mouse IgD (GAM/IGD(FC)/7S; Cedarlane Labs). After being washed for 5 min in TBS, slides were incubated for 1h with cocktails of the following secondary reagents in 0.1% BSA in TBS: alkaline phosphatase-conjugated mouse antibody to fluorescein isothiocyanate (200-052-037; Jackson ImmunoResearch), horseradish peroxidase–conjugated polyclonal donkey anti-goat (705-035-147; Jackson ImmunoResearch) and streptavidin–horseradish peroxidase. Enzyme conjugates were developed with DAB and Fast-blue (both from Sigma-Aldrich).

Statistics. Data were analyzed with Prism 5 (GraphPad Software). The two-tailed non-parametric Mann-Whitney test was used for comparison of two unpaired groups. The nonparametric Kruskal-Wallis test was used for comparison of three or more unpaired groups, followed by Dunn's post-hoc test for calculation of the *P* value for each group. Two-way analysis of variance was used together with Bonferroni *post-hoc* tests for comparison of replicates in each cell division of CTV-labeled OT-II cells.

48. Oxenius, A., Bachmann, M.F., Zinkernagel, R.M. & Hengartner, H. Virus-specific MHC-class II-restricted TCR-transgenic mice: effects on humoral and cellular immune responses after viral infection. *Eur. J. Immunol.* **28**, 390–400 (1998).
49. Rao, P.K. *et al.* Loss of cardiac microRNA-mediated regulation leads to dilated cardiomyopathy and heart failure. *Circ. Res.* **105**, 585–594 (2009).
50. Suzuki, A. *et al.* T cell-specific loss of Pten leads to defects in central and peripheral tolerance. *Immunity* **14**, 523–534 (2001).
51. Baumjohann, D. & Ansel, K.M. Identification of T follicular helper (Tfh) cells by flow cytometry. *Nat. Protoc.* doi.org/10.1038/protex.2013.060 (18 June 2013).
52. Tang, Q. *et al.* In vitro-expanded antigen-specific regulatory T cells suppress autoimmune diabetes. *J. Exp. Med.* **199**, 1455–1465 (2004).

doi:10.1038/ni.2642

Title: Functional activity of RNA-binding protein binding sites drives vertebrate 3′ UTR evolution

Authors: Adam J. Litterman[1,2], Robin Kageyama[1,2], Olivier Le Tonqueze[3,4], Wenxue Zhao[3,4], Hani Goodarzi[5,6,7], David J. Erle[3,4], K. Mark Ansel[1,2*]

Affiliations:

[1] Department of Microbiology & Immunology, University of California San Francisco, San Francisco, CA.

[2] Sandler Asthma Basic Research Center, University of California San Francisco, San Francisco, CA

[3] Department of Medicine, University of California San Francisco, San Francisco, CA.

[4] Lung Biology Center, University of California San Francisco, San Francisco, CA.

[5] Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA.

[6] Department of Urology, University of California San Francisco, San Francisco, CA.

[7] Helen Diller Family Comprehensive Cancer Center, San Francisco, CA.

*Correspondence to: Mark.Ansel@ucsf.edu

Abstract: Untranslated regions (UTRs) of RNA transcripts are bound by RNA binding proteins (RBPs) during constitutive RNA metabolism and gene specific regulatory interactions. Messenger RNA 3′ UTRs exhibit strong evidence of selection on microRNA binding sites and other cis-regulatory elements, but the overall pattern of 3′ UTR evolution is largely undescribed. To understand the evolution of functional elements in 3′ UTRs and their impact on gene expression, we developed GCLiPP, a biochemical technique for detecting RBP occupancy transcriptome-wide. Using GCLiPP, we identified >25,000 RBP binding sites in 3′ UTRs of T lymphocyte mRNAs. We then measured their effects on gene expression using a massively parallel reporter assay. GC rich regions of 3′ UTRs were highly bound by RBPs, rapidly evolving, destabilizing of reporter mRNAs, and more likely to be folded in vivo. By reducing gene expression, these RBP occupied sequences act as a rapidly evolving substrate for gene regulatory interactions.

Introduction

The life cycle of protein coding RNA transcripts involves their transcription from DNA, 5′ capping, splicing, 3′ polyadenylation, nuclear export, targeting to the correct cellular

compartment, translation and degradation (*1–3*). RNA binding proteins (RBPs) coordinately regulate these processes through interaction with RNA cis-regulatory elements, often in the 5′ and 3′ untranslated regions (UTRs) whose sequences are not constrained by a functional coding sequence (*4*). Mammalian genomes encode hundreds of RBPs (*5*), and mutations in individual RBPs or even individual binding sites can induce strong developmental, autoimmune and neurological defects in human patients and mouse models (*6–8*).

Methods like DNase I hypersensitivity and ATAC-seq that query regulatory element accessibility and occupancy without prior knowledge of their protein binding partners have proven remarkably powerful not only for systematic mapping of the genome, but also for uncovering specific regulatory events and fundamental principles of genome regulation (*9–11*). As much as half of the extensive gene expression changes that occur during T cell activation occur post-transcriptionally (*12*), and several RBPs are known to be critical determinants of immune function and homeostasis (*6*). Yet our understanding of post-transcriptional regulatory circuits and the evolution of untranslated regions of transcribed genes remain rudimentary, due in part to a lack of systematic methods for mapping the transcriptome's cis-regulatory landscape.

Although RBP binding sites are often deeply conserved and 3′ UTRs are more conserved than other non-coding sequences in vertebrate genomes (*13*) , the overall level of 3′ UTR sequence conservation varies widely between genes and within different regions of the same gene. Phylogenetic analyses have suggested that conservation and nucleotide composition are related, with AU-rich 3′ UTRs exhibiting greater conservation than GC-rich 3′ UTRs (*14*, *15*). The reason for this association remains unclear, as only a small fraction of UTR sequence space has been functionally annotated and information about RBP occupancy is largely limited to interrogation by biochemical purification of individual RBPs (*16*).

Results

Transcriptome-wide analysis of RBP occupancy in mouse T cells

To achieve transcriptome-wide RBP binding site profiling, we developed a protocol for Global Cross-linking Protein Purification (GCLiPP) suitable for use in mammalian cells, and applied this technique in cultured primary mouse T cells (Figure 1A). GCLiPP is an adaptation of previously described biochemical methods for crosslinking purification of all mRNA-RBP complexes (*17–19*). The key features of GCLiPP include: crosslinking of endogenous ribonucleoproteins complexes using high energy UV light (no photo-crosslinkable ribonucleotide analogues); oligo-dT pulldown prior to biotinylation to enrich for mRNA species; chemical biotinylation of primary amines using a water soluble reagent with a long, flexible linker; brief RNase digestion with the guanine specific RNase T1; and on-bead linker ligation with radiolabeled 3′ linker to facilitate downstream detection of ligated products. We used the guanine specific ribonuclease T1 to favor larger average fragment sizes than using an RNA endonuclease with no nucleotide specificity (such as RNase A). We hypothesized that this would facilitate detection of RBP binding to sites in unstructured, GC-poor regions which would tend to be digested into larger fragments relative to the protein-protected site due to their relative lack of guanine residues. Sequencing libraries of GCLiPP fragments yielded highly reproducible patterns in $CD4^+$ and $CD8^+$ T cells (Fig. 1B and 1C), with read coverage strongly enriched within mature mRNAs and long non-coding RNAs (Figure 1—Figure Supplement 1).

To validate that GCLiPP detects *bona fide* RBP-occupied sites in cellular RNA, we examined known regulatory interactions of various functional and structural categories. GCLiPP recapitulated previously described *cis* regulatory elements bound by known RBPs (Figures 2 and

3 and Figure 2—figure supplement 1-5) that mediate constitutive RNA metabolism (Figure 2A and Figure 2—figure supplement 1), transcript localization (Figure 2B, Figure 2—figure supplement 2), regulation of gene expression (Fig. 2C and Figure 2—figure supplement 3), and translation (Figure 2—figure supplement 4), including both structured elements and single-stranded RNA determinants.

Constitutive mRNA metabolism cis-*elements*

The canonical polyadenylation signal AAUAAA is a known linear sequence motif that binds to a number of RBPs in the polyadenylation complex, including CPSF and PABP (*20*), as part of constitutive mRNA metabolism. We examined T cell lineage-defining transcripts with well resolved GCLiPP profiles (due to their high expression levels), including *Cd3g* (Figure 2A), *Cd3e, Cd4*, and *Cd8b1* (Figure 2—figure supplement 1). The only canonical polyadenylation signal sequences in these transcripts were contained within called GCLiPP peaks, often as the peak with the highest GCLiPP read density in the entire transcript. Interestingly, in the *Cd8b1* transcript we could observe direct biochemical evidence for alternative polyadenylation signal usage (Figure 2—figure supplement 1C), a phenomenon that has previously been described to be important in activated T cells (*21*). GCLiPP peaks appear in multiple canonical polyadenylation signal sequences in *Cd8b1*, coincident with clear evidence for both short and long 3′ UTR isoform usage indicated by lower RNAseq read counts after the initial canonical polyadenylation signal. A similar pattern was also apparent in *Hif1a* (Figure 2—figure supplement 1D) and a number of other highly expressed transcripts.

Transcript subcellular localization elements

Known cis-regulatory elements involved in transcript localization were also represented by local regions of GCLiPP read density. The Beta-actin "zipcode" element is responsible for localization of *Actb* mRNA to the cellular leading edge in chicken embryo fibroblasts (*22*) and contains conserved linear sequence elements separated by a variable linker. These conserved sequence elements are thought to form the RNA/protein contacts in a complex involving the actin mRNA and the RNA binding protein Igf2bp1 (previously known as Zbp1) where the non-conserved sequence winds around the RBP (*23*). This sequence corresponds to the center of the second highest peak of GCLiPP read density in the *Actb* transcript (Figure 2B). Similarly, peaks of GCLiPP read density corresponded to perinuclear localization signals in the 3′ UTRs of the transcripts of Vimentin and Myc as defined previously by sufficiency for perinuclear localization of heterologous reporter constructs (Figure 2—figure supplement 2) (*31*, *32*).

Cis-*regulatory elements that control mRNA stability and translation*

Some RBPs regulate the half-life and/or translation of the mRNAs that they bind. The mRNA-destabilizing Roquin/Regnase binding site in the 3′ UTR of *Ier3* is a straightforward example of this functional category of RNA/RBP interaction detected as a region of GCLiPP read density (Figure 2C). The known mRNA stabilizing iron response elements in the transferrin receptor (*Tfrc*) 3′ UTR and ferritin light chain (*Ftl*) 5′ UTR showed a more complex mode of detection by GCLiPP. These elements consist of a stereotypic hairpin structure containing the motif RRCAGUGNYY that binds to a cytoplasmic factor, originally called IRF and now called Aco1, which stabilizes the transferrin receptor mRNA and inhibits translation of the ferritin mRNA in low iron conditions (*26–28*). We compared GCLiPP with transcriptome wide

measurements of RNA structure from icSHAPE (*29*) and observed that paired peaks of GCLiPP read density surrounded exposed, unfolded guanine residues in the loop of this motif (Figure 2—figure supplement 3).

The insertion of the selenium containing amino acid selenocysteine into selenoproteins represents a unique case of RBP regulation of protein translation. Selenoproteins are redox enzymes that use selenocysteine at key reactive residues (*30*, *31*). Selenocysteine is encoded by the stop codon UGA, and this recoding occurs only in mRNAs that contain 3′ UTR *cis*-regulatory elements (termed SECIS elements) that bind to RBPs that recruit the elongation factor Eefsec and selenocysteine-tRNA (*32*, *33*). SECIS elements were prominent peaks of GCLiPP read coverage in selenoprotein mRNAs. For example, the SECIS element (*34*) in the 3′ UTR of *Gpx4* was entirely covered by GCLiPP reads (Figure 2—supplement 4A). Indeed, a canonical polyadenylation signal and the full hairpin structure containing the SECIS element account for essentially all of the GCLiPP reads in the *Gpx4* 3′ UTR. Comparing transcriptome-wide in vivo icSHAPE and GCLiPP data suggests that the folded, RBP bound structure is even larger than that predicted by SECISearch 3, with regions of GCLiPP read density and apposed high and low icSHAPE signal spanning almost the entire 3′ UTR (Figure 2—figure supplement 4B,C). icSHAPE also revealed that a conserved stretch of adenines in the apical loop were exposed to icSHAPE tagging (Figure 2—figure supplement 4D).

Protein occupancy in non-coding RNAs

GCLiPP also corroborated RBP-RNA interactions in both long and short non-coding RNAs. For example, we observed GCLiPP read density at Pumilio protein binding sites in the abundant lncRNA *Norad* (also known as *2900097C17Rik* in mouse and *LINC00657* in humans).

Norad deletion is associated with genomic instability and aneuploidy (*35*, *36*). The lncRNA functions by sequestering Pumilio family RBPs, reducing their ability to regulate mRNAs that also contain Pumilio binding sites. We examined *Norad* for high-scoring linear Pumilio motifs using a 20 nt position weight matrix previously determined by SELEX experiments using mouse Pum2 (*37*). The eight highest scoring motifs containing a canonical UGUA minimal Pumilio binding sequence marked the regions of greatest RBP occupancy as indicated by GCLiPP read density in *Norad* (Figure 3A).

7SK is an abundant small nuclear (sn)RNA that forms distinct ribonucleoprotein complexes with several RBP partners. This highly structured non-coding RNA is transcribed by PolIII and is present in the nucleus of all eukaryotic cells, where it scaffolds a complex between the PolII regulatory kinase P-TEFb and other RNA binding proteins (*38*). The entirety of the 7SK RNA is recovered in GCLiPP reads (Figure 3B), with clear regions of higher GCLiPP coverage corresponding to the conserved structural motifs M1, M7 and M8 (*39*). This GCLiPP profile clearly delineates previously identified structural motifs and protein interaction domains. The M8 stem-loop binds to the 7SK stabilizing protein LaRP7 (*40*). The key interaction with P-TefB also occurs through direct or indirect binding at the M8 motif. The M1 and M7 motifs interact with various hnRNPs (*38*), and the M3 hairpin binds Hexim proteins that are critical for transcriptional elongation (*41*). The apical loops of the M3 and M5 motifs show lower GCLiPP coverage, presumably due to cleavage of the exposed loops of protein-bound hairpins.

Distinct from these complexes that act near gene promoters, 7SK snRNA also interacts with the BAF chromatin remodeling complex at enhancers (*42*). Interestingly, while the Hexim1 binding motif M3 was in a region of low GCLiPP read density, the M1 region of the transcript adopts an icSHAPE accessible conformation when 7SK is in complex with Hexim1 and had low

icSHAPE accessibility and high GCLiPP coverage (Figure 3B). Further, the stem of the M7 motif is highly represented in GCLiPP reads, and the loop of this structure represents an area whose icSHAPE accessibility is higher when 7SK is in association with the BAF complex (Figure 3—figure supplement 1). These data suggest that the M7 motif may represent a site of association with BAF.

Taken together, these data provide evidence that GCLiPP read density reflects the abundance of a wide variety of crosslinked RBP-RNA species in cellular transcripts. GCLiPP reads are abundantly recovered from interactions between RBPs and single stranded as well as double stranded RNAs. The single stranded RNA binding sites identified can occur either via binding to simple linear sequence motifs (such as canonical polyadenylation signals and Pum motifs) or in more complex structures involving interactions of multiple spatially separated sequence elements that have specific interactions with RBPs (such as the *Actb* zipcode). Double stranded RNA binding events also show several modes of representation in GCLiPP data. Some small hairpin loops were cloned whole and show large peaks of GCLiPP read abundance (such as the Roquin binding loop in *Ier3*). Other structures are more susceptible to RNase T1 digestion and are typically cloned as two fragments (such as the iron response elements in iron responsive transcripts) or as many separate fragments of a larger secondary structure (such as SECIS elements and the 7SK snRNA).

RBP-RNA interactions identified by GCLiPP occur in 5′ UTRs, coding sequences, 3′ UTRs and non-coding RNAs. For the purpose of this study, we restricted functional analyses to 3′ UTRs and used 70 nucleotide called peaks of GCLiPP read density to define putative functional units to test in a heterologous reporter assay of post-transcriptional regulation of gene expression. The large number of known RBP-RNA interactions of diverse structural types which

are localized to local peaks of GCLiPP read density give us confidence that a large number of our called peaks represent a sufficient amount of sequence context to recapitulate the functional effects of many RNA-RBP interactions.

Relationship between RBP occupancy in GCLiPP and nucleotide composition

GCLiPP identified thousands of putative cis-regulatory sequences and revealed striking global relationships between RBP occupancy, nucleotide composition, and evolutionary sequence conservation in 3′ UTRs. To facilitate functional analysis, we defined ~27000 70 nt "GCLiPP peaks" of cross-linked 3′ UTR sequence reads and 5000 "conserved voids" that were absent or very low in GCLiPP but abundantly expressed by mRNA-seq (Figure 4A). Overall, GCLiPP peaks had much higher GC content than the annotated 3′ UTRs from which they are derived (Figure 4B, $p<10^{-307}$, t=50.4, Welch's two sample t-test), and voids had even lower GC content (Figure 4B, $p<10^{-307}$, t=59.3, Welch's two sample t-test). Consistent with the poor conservation of GC rich sequences within 3′ UTRs (*15*), GCLiPP peak GC content negatively correlated with evolutionary conservation among placental mammals (Figure 4C, t = -76.4, ρ=-0.411). Thus, GC rich sequences are enriched for RBP occupancy despite the fact that they are not strictly conserved and are often subject to lineage-specific selection.

Fast-UTR massively parallel reporter assay of GCLiPP defined RBP occupied sites

High-throughput functional analysis further revealed how these features dictate mRNA stability. We transduced mouse T cells with a retroviral fast-UTR library (*43*) containing all 27000 GCLiPP peaks, 5000 voids, and 7000 additional control sequences, and inferred the effect of each insert on mRNA stability from the relative ratios of sequenced amplicons from reverse

transcribed RNA and genomic DNA templates. As expected, inserts containing seed binding sequences for highly expressed miRNAs were destabilizing compared with inserts containing scrambled variants of the same seed sequence (Figure 4—figure supplement 1). An unexpected and striking pattern in the data set was the strong negative correlation between insert GC content and fast-UTR mRNA stability (Figure 4D, t = -143.19, ρ=-0.658). For inserts representing GCLiPP peaks, there was a corresponding positive correlation between fast-UTR mRNA stability and evolutionary conservation (t = 44.1, ρ=0.260), with stepwise decreases in stability observed for inserts binned from the most strictly conserved to the most rapidly evolving sequences (Fig. 4E). For rapidly evolving GCLiPP peaks (examples showing dissimilarity of sequences between human and mouse shown in Figure 4—supplement 2A), there was only a modest correlation between the mouse and syntenic human sequence effects on fast-UTR mRNA stability (Figure 4—figure supplement 2B, t = 6.3, ρ=0.188). However, much of the species-specific differences in destabilizing activity could be explained by changes in GC content (Figure 4—figure supplement 2C, t =-14.606, ρ=-0.408). These data indicate that acquisition of GC rich sequences within protein-bound regions of 3′ UTRs confers destabilizing functional activity.

Validation of relationship between 3′ UTR GC content and effect on gene expression

The association between nucleotide composition and effect on gene expression observed for isolated GCLiPP peaks holds for longer sequences. GC-rich full length 3′ UTRs containing numerous strong GCLiPP peaks (*Cd4, Dusp2, Ier2*) reduced reporter luciferase activity, whereas an AU-rich 3′ UTR with little protein binding (*Cnn3*) had no effect (Figure 4F). Protein production was similarly affected in previous experiments with a lentiviral fast-UTR library of

160 nt human 3′ UTR segments downstream of *EGFP* (*43*). FACS-sorted cells with low EGFP fluorescence harbored 3′ UTR inserts with significantly higher GC content than inserts from cells with high EGFP expression (Figure 4G, $p<10^{-28}$, Welch's two-sample t-test). These inserts also reduced mRNA half-life (Figure 4H) and lowered steady-state mRNA abundance in three human cell lines (Fig. 4I). In all of these prior experiments, 3′ UTR GC content strongly correlated with reduced gene expression (Figure 4—figure supplement 3).

RNA secondary structure in RBP-occupied regions detected by GCLiPP

RNA folding influences global patterns of RBP occupancy and functional activity. The predicted folding energy of GCLiPP peaks was significantly lower than that of conserved voids (Figure 5—figure supplement 1A), as expected since folding energy is strongly correlated with GC content (Figure 5—figure supplement 1B). To look for evidence of in vivo folding in RBP occupied sites genome wide, we compared GCLiPP data from mouse T cells with published icSHAPE in vivo folding data from mouse ES cells across all transcripts expressed in both cell types. The predicted secondary structure of individual RBP-bound GC rich GCLiPP peaks tended to be supported by low icSHAPE signal at nucleotides in predicted folds, immediately adjacent to icSHAPE tagged nucleotides predicted to be in an accessible loop or bulge (Figure 5A). Conversely, conserved voids of protein binding tended to be GC-poor with higher predicted folding energies and structures unsupported by icSHAPE profiles. In these sequences, nucleotides in predicted folds were similarly accessible to icSHAPE tagging as unfolded nucleotides, indicating that in vivo structure differed from the in silico predicted conformation (Figure 5B).

We took advantage of these patterns in the icSHAPE signals of structured and unstructured sequences to assess the in vivo folding of all GCLiPP peaks and voids. We reasoned that structures held in a folded conformation in vivo exhibit higher local variability of icSHAPE signal due to the apposition of minimally tagged (tightly folded) nucleotides and maximally tagged (bulge, loop or flanking) nucleotides with high icSHAPE accessibility. Therefore, we examined the standard deviation and multi-modality of icSHAPE signal as a proxy for in vivo folding within GCLiPP peaks and voids. Consistent with their GC content and predicted folding energies, GCLiPP voids had lower icSHAPE standard deviations than GCLiPP peaks (Figure 5C). Among these peaks, the most rapidly evolving RBP-occupied sites also had the highest icSHAPE standard deviations (Figure 5C). A similar pattern held for the dip statistic (a continuous measurement of multi-modality) computed using Hartigan's dip test of unimodality (53) (Figure 5—figure supplement 2). Thus, RBP-occupied sequences identified by GCLiPP are enriched in structural elements that adopt a folded conformation in vivo.

We identified enriched short structural motifs associated with fast-UTR destabilizing activity using TEISER (*43*) (Figure 6A). icSHAPE profiles were concordant with the predicted structure for most individual examples of these motifs. Maximal or near maximal icSHAPE tagging occurred at predicted bulge, loop or linear nucleotides near the base of the hairpin, and minimal tagging was observed at predicted folded nucleotides (Figure 6B). We confirmed the inhibitory effect of these motifs in T cells transfected with in vitro transcribed reporter mRNAs linked to short 3′ UTRs containing the exemplary in vivo folded sequences (Figure 6C). The relationship between structured 3′ UTR elements and inhibition of gene expression was not entirely dependent on GC content, as better than expected folders were more destabilizing than poor folders with similar GC content (Figure 5—figure supplement 1C). Thus, linked GCLiPP

and fast-UTR analyses revealed specific structural motifs associated with gene regulation, and a global relationship between 3′ UTR sequence structure, protein occupancy, and mRNA stability.

Evolutionary pressures shape 3′ UTRs differently in distinct classes of genes

Phylogenetic analysis revealed that post-transcriptional regulation shapes vertebrate 3′ UTR evolution, with purifying selection of GC poor sequences and accelerated lineage-specific evolution of GC rich RBP binding cis-regulatory regions. GC content and PhyloP scores were strongly negatively correlated for GCLiPP peaks, but not for a control set of P300-bound (*45*) transcriptional enhancers (Figure 7—figure supplement 1A,B). Rapidly evolving sequence content (PhyloP <0) also strongly correlated with GC content (Figure 7—figure supplement 1C,D). Overall, 3′ UTRs that exhibit strong evidence of purifying selection have lower GC content than other 3′ UTRs across nine vertebrate species (Figure 7A). Conversely, 3′ UTRs exhibiting lineage specific selection have higher GC content across species (Figure 7B). Thus, across the vertebrate lineage, 3′ UTRs that differ in nucleotide content face different regimes of selection, with purifying selection on GC poor UTRs and lineage-specific, accelerated evolution of GC rich UTRs. Furthermore, the biological categories of genes selected under these different regimes did not appear to be randomly distributed. Genes whose 3′ UTRs exhibit significant evidence of purifying selection and low GC content are enriched in gene ontology categories involving body plan development and organ morphogenesis (Figure 7C), whereas genes whose 3′ UTRs are rapidly evolving and GC rich are over-represented among genes related to inflammation and metabolism (Figure 7D).

Discussion

This study outlines a previously undescribed paradigm that shapes untranslated transcript sequence throughout vertebrate genomes. By starting with an unbiased biochemical method to identify RBP occupied sequences in mRNAs, we discovered that RBP occupied sites in 3′ UTRs tend to have a higher GC content (~48%) than the 3′ UTRs themselves (~43.5%). These GC rich RBP binding sites tend to be less strictly conserved across vertebrates than less bound, GC poor regions of 3′ UTRs. Across multiple experiments in primary mouse cells and human cell lines, we found that these GC rich UTR sequences were associated with lower gene expression as measured by the steady state mRNA level, amount of protein and half-life of transcripts. In turn, it appears that the protein bound sequences that are GC rich are more likely to be folded, as demonstrated both by theoretical local folding energies for the bound and unbound sequences as well as in vivo measurements of RNA conformation. Across vertebrate 3′ UTRs, there is a tendency towards purifying selection of GC poor UTRs, whereas GC rich UTRs tend to exhibit evidence of rapid selection in specific lineages.

Taken together, these results paint a global picture wherein purifying selection winnows away local regions with high GC content in certain classes of genes. These GC rich regions are more likely to be folded in vivo and to form a binding site for an RBP, and reduce the amount of protein produced from a given amount of transcript. When GC rich regions invade the 3′ UTR of a gene in a specific lineage, they tend to exhibit accelerated evolution. That is, they become fixed in the lineage at a rate greater than would be predicted by neutral drift. GC biased gene conversion may contribute to this phenomenon, favoring the creation of novel post-transcriptional regulatory elements, but it affects only a small portion of vertebrate genomes—approximately 0.3% of the human genome (*46*). Only 13/1198 (~1.1%) of rapidly evolving

vertebrate 3′ UTRs we identified in this study overlapped with regions of the human genome undergoing GC biased gene conversion.

Our results leave unresolved the specific mechanism of regulation of gene expression by differences in UTR nucleotide composition. In general, our data suggest that GC rich sequences are more likely to be folded and to be bound by RBPs than GC poor sequences, and that transcripts containing GC rich UTRs are less stable. However, the effect of individual RBP-binding elements may vary in their native sequence context. Similarly, our experiments do not resolve whether or how multiple regulatory elements may cooperate or compete to achieve precise gene regulation. There are several well known examples of RBPs that bind to locally folded structures and lead to transcript degradation, such as Roquin (*47*), Regnase (*48*) and Staufen (*49*). Although Roquin and Regnase bind to conserved stereotypic hairpin structures, Staufen binds promiscuously to tightly folded sequences in 3′ UTRs of transcripts throughout the genome, typified by runs of guanines base-paired with pyrimidine rich tracts. There may be other RBPs like Staufen that use promiscuous binding to dsRNA to mediate transcript degradation (*50*). However, other RBP interactions with locally folded structures are associated with mRNA stabilization, such as iron regulatory element binding proteins that stabilize and sequester mRNAs involved in iron metabolism. A broad regulatory interactome is hinted at by the lack of known binding partners for active structural motifs that we identified by functional analysis of GCLiPP peaks. Future work should leverage these data to discover and classify gene regulatory interactions between *cis* regulatory sequences and *trans* acting RBPs.

The general lack of conservation of regulatory motifs suggests that the major selective pressure on most 3′ UTRs has not been to develop regulatory elements that induce transcript degradation, but rather to select for sequences associated with longer mRNA stability and

consequently greater protein production. Highly conserved strongly destabilizing elements such as canonical AU rich element nonamers (*51*), the *Tnf* constitutive decay element and other conserved Roquin binding loops (*47*), and CUG repeats bound by CUG binding proteins (*52*) are exemplary counterexamples to this general pattern in 3′ UTR evolution (*13*). These examples are likely driven by selection for stringent controls on genes whose overexpression is deleterious, such as inflammatory mediators and proto-oncogenes.

Among RBP-occupied regions detected by GCLiPP, the most strongly destabilizing sequences tended to be very GC-rich and rapidly evolving, often exhibiting highly divergent sequences between mouse and human (two vertebrates that are only separated by ~75 million years from their last common ancestor). For these rapidly evolving sites, species-specific acquisition of high GC content is associated with acquisition of destabilizing activity. Therefore, acquisition or loss of high local GC content in UTRs may be a major mechanism for the diversification of gene expression levels across species. The GC content of synonymous wobble bases in coding sequences correlates with the GC content of UTRs of the same gene (*53*), suggesting that selection on gene expression level may be a strong pressure driving mRNA sequence variation in general.

The classes of genes that exhibit these patterns of selection are not random, as genes involved in organismal development exhibited evidence of purifying selection for GC poor 3′ UTRs, whereas genes involved in metabolism and immune response tended to vary dramatically in 3′ UTR sequence. Thus, it appears that there are specific classes of genes that are functionally selected to create a large amount of protein per transcript, and that these genes tend to be involved in core developmental processes. Other genes are post-transcriptionally constrained by cis-regulatory elements whose sequences vary between species. These genes likely evolved

highly tailored expression programs in response to evolutionary pressure to contain pathogens or metabolize xenobiotics.

Materials and Methods

Cells

Primary CD4+ and CD8+ mouse T cells were isolated from C57/BL6J peripheral lymph nodes and spleen using positive and negative selection Dynabeads, respectively, according to the manufacturer's instructions (Invitrogen). Cells were stimulated with immobilized biotinylated anti-CD3 (clone 2C11, 0.25 ug/mL, BioXcell) and anti-CD28 (clone 37.51, 1 ug/mL, BioXcell) bound to Corning 10 cm cell culture dishes coated with Neutravidin (Thermo) at 10 ug/mL in PBS for 3 h at 37 degrees C. Cells were left on stimulation for 3 days before being taken off of stimulation and split into non-coated dishes in T cell medium supplemented with recombinant human IL-2 (20 U/mL). Th2 cells were polarized in medium containing 1:100 dilution of IL-4 conditioned medium, anti-Ifn-γ (10 ug/mL). CD8 cells were grown in medium containing 10 ng/mL recombinant murine IL-12 (10 ng/mL). T cells were grown in medium as described previously (*54*). For re-stimulation, cells were treated with PMA and Ionomycin (20 nM and 1 μM, respectively) for 4 hours before harvest.

GCLiPP and RNAseq

~100e6 mouse T cells cultured from 3 mice were washed and resuspended in ice cold PBS and UV crosslinked with a 254 nanometer UV Stratagene crosslinker in three doses of 4000 mJ, 2000 mJ and 2000 mJ, swirling on ice between doses. Cells were pelleted and frozen at -80 degrees C. Thawed pellets were resuspended in 400 μL PXL buffer without SDS (1X PBS with

0.5% deoxycholate, 0.5% NP-40, Protease inhibitor cocktail) supplemented with 2000 U RNasin (Promega) and 10 U DNase (Invitrogen). Pellets were incubated at 37 degrees C with shaking for 10 minutes, before pelleting of nuclei and cell debris (17000xg for 5 min). Supernatants were biotinylated by mixing at room temperature for 30 minutes with 500 μL of 10 mM EZ-Link NHS- SS-Biotin (Thermo) and 100 μL of 1 M sodium bicarbonate. Supernatants were mixed with 1 mg of washed oligo-dT beads (New England Biolabs) at room temperature for 30 minutes and washed 3 times after pulldown with a magnet. Oligo-dT selected RNA was eluted from beads by heating in poly-A elution buffer (NEB) at 65 degrees C with vigorous shaking for 10 minutes. An aliquot of eluted RNA was treated with proteinase K and saved for RNAseq analysis by using Illumina TruSeq Stranded Total RNA Library Prep Kit according to the manufacturer's instructions.

The rest of the crosslinked, biotinylated mRNA-RBP complexes were captured on 250 μL of washed M-280 Streptavidin Dynabeads (Invitrogen) for 30 minutes at 4 degrees C with rotation. Beads were washed 3X with PBS and resuspended in 40 μL of PBS containing 1000 U of RNase T1 (Thermo) for 1 m at room temperature. RNase activity was stopped by addition of concentrated (10%) SDS to a final concentration of 1% SDS and beads were washed successively in 1X PXL buffer, 5X PXL buffer and twice in PBS. 24 pmol of 3′ radiolabeled RNA linker was ligated to RBP bound RNA fragments by resuspending beads in 20 μL ligation buffer containing 10 U T4 RNA Ligase 1 (New England Biolabs) with 20% PEG 8000 at 37 degrees for 3 h. Beads were washed 3X with PBS and free 5′ RNA ends were phosphorylated with polynucleotide kinase (New England Biolabs). Beads were washed 3X with PBS and resuspensded in ligation buffer containing 10 U T4 RNA Ligase 1, 50 pmol of 5′ RNA linker and 20% PEG 8000 and incubated at 15 degrees C overnight with intermittent mixing. Beads were

again washed 3X in PBS and linker ligated RBP binding sites were eluted by treatment with proteinase K in 20 μL PBS with high speed shaking at 55 degrees C. Beads and supernatant were mixed 1:1 with bromophenol blue formamide RNA gel loading dye (Thermo) and loaded onto a 15% TBE-Urea denaturing polyacrylamide gel (BioRad). Ligated products with insert were visualized by autoradiography and compared to a control ligation (19 and 24 nt markers). Gel slices were crushed and soaked in gel diffusion buffer (0.5 M ammonium acetate; 10 mM magnesium acetate; 1 mM EDTA, pH 8.0; 0.1% SDS) at 37 degrees for 30 min with high speed shaking, ethanol precipitated and resuspended in 20 μL of RNase free water. Ligated RNAs were reverse transcribed with Superscript III reverse transcriptase (Invitrogen) and amplified with Q5 polymerase (New England Biolabs). PCR was monitored using a real time PCR thermal cycler and amplification was ceased when it ceased to amplify linearly. PCR products were run on a 10% TBE polyacrylamide gel, size selected for an amplicon with the predicted 20-50 bp insert size to exclude linker dimers, cut from the gel, and gel purified. Cleaned up library DNA was quantified on a bioanalyzer (Agilent) before being deep sequenced. All GCliPP and RNAseq sequencing runs were carried out on an Illumina HiSeq 2500 sequencer.

GCLiPP bioinformatics analysis pipeline

FastQ files were de-multiplexed and trimmed of adapters. Each experiment was performed on three technical replicates per condition (resting and stimulated) per experiment. Cloning replicates and experiments were pooled in subsequent analyses. Trimmed sequence reads were aligned to the mm10 mouse genome assembly using bowtie2. After alignment, PCR amplification artifacts were removed by de-duplication using the 2-nt random sequence at the 5′ end of the 3′ linker using a custom script that counted only a single read containing a unique

linker sequence and start and end position of alignment per sequenced sample. Custom scripts were used first to identify 3′ UTRs of expressed genes from RNAseq (expanding 3′ UTRs where there was clear evidence of expression beyond the refseq defined end) and then to identify peaks and voids of GCLiPP read density by convolving a normal distribution against a sliding window of the observed read distribution and looking for local maxima of correlation, with a read depth above a transcript minimum based on the average read depth in the UTR in GCLiPP and a global minimum set manually. Peak calling was performed on all data pooled and repeated on data separated into reads derived from Th2 and CD8 T cells, and all these called peaks were carried forward into subsequent analyses. Peak names indicate whether they were derived from calling on combined data, Th2 data or CD8 data. Sequencing data and summaries are available at Gene Expression Omnibus (accession number GSE94554). No power calculation was performed to pre-determine sample number. Samples were accumulated until high reproducibility (as assessed by Pearson correlation of read density at called peaks) between replicates of samples was observed.

Fast-UTR vector assembly, library construction and assay

A multiple cloning site consisting of MluI, I-SceI, and PacI was inserted into the 3′ UTR of the GFP gene in a previously described T cell retroviral microRNA sensor plasmid (*54*). A DNA oligonucleotide library consisting of RBP binding sites, voids and control randomers was synthesized by CustomArray (Bothell, WA). Three sets of randomer control sequences with different dinucleotide frequency were used, either using a background dinucleotide frequency that was essentially random (~3000 sequences, all di-nucleotides except for CpG represented equally, CpG represented at approximately the same frequency as the mouse genome), using the

dinucleotide frequency of the mouse genome (~2000 sequences) or using the dinucleotide frequency of mouse 3′ UTRs (~2000 sequences). Additionally, we included ~1000 human syntenic regions for the most rapidly evolving protein binding peaks in mouse 3′ UTRs. These sequences were determined by using the kentUtils liftover program on a single nucleotide at the very center of the 70 nt peak, and taking the 70 nucleotides adjacent to the lifted over nucleotide for mouse peaks where a liftover could successfully be determined.

The library was amplified using Q5 polymerase and a real-time PCR cycler through the linear range, cleaned up with a PCR cleanup kit (Qiagen), cut with MluI and PacI and run on a 10% TBE polyacrylamide gel to isolate double cut PCR fragments. The vector fragment was cut with MluI, I-SceI and PacI and run out on a 1% agarose gel and cleaned up with a gel purification kit (Qiagen). Both vector and insert were quantified and ligated at a ~10:1 molar insert:vector ratio using quick ligation kit (New England Biolabs). Ligation mixture was purified by PCR cleanup kit and electroporated into TG1 electrocompetent bacteria (Lucigen). Colonies were plated on bioassay plates and plasmid was prepared from pooled colonies scraped from bioassay plates into LB media. Ecotropic retrovirus was made by transient transfection of plasmid into Plat-E packaging cells by calcium phosphate method. Plat-E cells were grown in complete DMEM supplemented with 5% FBS. After transfection, packaging cells were left with DNA overnight, then aspirated and incubated with fresh collection media containing 10% fetal bovine serum and 1X ViralBoost reagent (ALStem). Virus containing supernatant was collected off cells, mixed with 8 μg/mL polybrene and put onto Day 2 mouse Th2 cells for 6 hours. Transduced cells were grown until Day 5, washed thrice with PBS and then collected in Trizol reagent (Thermo). RNA was collected according to the manufacturer's protocol, but after the first aqueous phase was removed, the same volume of back-extraction buffer (4M Guanidine

Thiocyanate, 50 mM Sodium Citrate, 1 M Tris base) was added and re-separated, and DNA precipitated from the second aqueous phase. RNA was reverse transcribed, and both cDNA and genomic DNA were used as a template for PCR amplification of fast-UTR inserts. Amplicons were run on a 2% agarose gel, size selected for the appropriate insert size, and gel extracted. Cleaned DNA was run on a bioanalyzer (Agilent) before being sequenced on a HiSeq 4000, run with 50% PhiX spike-in to allow clustering of non-diverse 5′ (vector) ends.

Fast-UTR data analysis and TEISER analysis

Sequencing reads trimmed of vector sequences and were aligned to the oligonucleotide library using bowtie2. Each fast-UTR insert was amplified from the oligonucleotide library with a hexanucleotide random barcode inserted in the 3′ amplification primer. A ratio of RNA reads to genomic DNA reads was computed for each barcoded insert, and a weighted average of ratios (weighted by the number of genomic DNA reads, considered to be a proxy for the amount of expansion of each independently transduced clone). This weighted ratio for each insert was divided by the ratio for the median insert and this median normalized RNA/DNA ratio (steady state mRNA level) was used for analysis of insert stability (values < 1 are less stable than the median insert, >1 more stable). All statistical analyses of fast-UTR sequences were performed with R programming language (https://www.r-project.org/). TEISER structural motif discovery was performed on GCLiPP peak sequences ranked by their steady state mRNA levels as previously described (*44*).

Reporter assays

For luciferase assays 3′ UTRs were cloned into the dual luciferase reporter plasmid psiCHECK-2 (Promega) downstream of the Renilla luciferase stop codon into the XhoI and NotI sites. Cultures of Th2 cells from three or four mice (biological replicates) were grown in vitro for 4 days and 4e5 cells were transfected in triplicates (technical replicates) with 1 μg of plasmid DNA. 24 h later firefly and renilla luciferase activity were measured with Dual-Luciferase Reporter Assay System (Promega) according to manufacturer's instructions. The mean of technical replicates was used as the ratio of renilla to firefly for each culture. No power calculation was performed to pre-determine sample size. Sample size necessary to observe significant differences in protein production was estimated based on previous reporter assays evaluating effects of microRNAs.

For protein production assays the Kikume Green Red coding sequence was PCR amplified from the pCAG-KikGR plasmid (AddGene) with primers designed to add a T7 in vitro transcription signal to the 5′ end of the coding sequence and a sequence of interest (flanked by 6 U's on both the 5′ and 3′ sides) to the 3′ end (*55*). The control poly-U 32 and Tnf constitutive decay element sequences were from (*51*) and (*47*), respectively. KikGR PCR products were cleaned up and used as templates for in vitro transcription using the HiScribe T7 ARCA mRNA Kit with tailing according to the manufacturer's instructions (New England Biolabs). mRNA integrity and Poly-A tailing was assessed by agarose gel electrophoresis. Cultures of CD8 cells from three or for mice were grown in vitro for 3 days and 2e6 cells were transfected with 2 μg of mRNA for each construct. To measure protein production in a specific time frame Kikume protein was photoconverted from green to red after 4 h by exposure to violet flashlight for 10 m, then cultured for another 4 h before being analyzed by flow cytometry. Transfected cells were

run on an LSRII with a yellow-green laser (BD Biosciences) gated on KikG+ cells and the MFI of KikR (produced before photoconversion) was determined using Cytobank software (http://www.cytobank.org).

icSHAPE and Phylogenetic analyses

Predicted RNA folding energies were computed using the ViennaRNA rnafold program (http://www.tbi.univie.ac.at/RNA/) (*56*) and visualized using forna (http://rna.tbi.univie.ac.at/forna/) (*57*). For icSHAPE we used a published bigwig file of locally normalized icSHAPE signal intensity generated in mouse ES cells (*29*). To measure conservation of loci in the mouse genome in placental mammals, we used the placental mammal PhyloP bigwig file from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenpath/mm10/phyloP60way/). We computed mean PhyloP scores across given GCLiPP peaks and called P300 enhancer peaks (*45*) by using custom perl scripts calling the kentutils bigWigSummary program. A similar script was used to obtain distributions of icSHAPE signals at selected loci, and Hartigan's dip test of unimodality was performed using the R 'diptest' package from the CRAN depository (http://cran.r-project.org).

To examine conservation across vertebrate 3′ UTRs we downloaded genomes and 3′ UTR annotations for nine vertebrate species (*Bos taurus*, *Canis familiaris*, *Danio rerio*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Macaca mulatta*, *Rattus norvegicus*, *Xenopus tropicalis*) from UCSC genome browser (https://genome.ucsc.edu/). To avoid aligning non-syntenic sequences we only used 3′ UTRs for which a single annotated 3′ UTR existed. For genes with 4 or more 3′ UTRs we performed multiple sequence alignments with Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/) and computed p-values for conservation or

acceleration of those multiple sequence alignments using a standard vertebrate phylogenetic model (vertebrate.mod available at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP46way/) using the phyloP program (http://compgen.cshl.edu/phast/help-pages/phyloP.txt). We used custom perl scripts to score the GC content of the individual aligned sequences and analyzed the collated data in R. UTRs were then classified on the basis of whether they exhibited strong evidence ($p < 0.001$) or weak evidence ($p>0.1$) of strict conservation or accelerated evolution. The genes in each of these categories were then analyzed for biological categories using the Metascape online interface (http://metascape.org) using the default settings.

Oligonucleotide and primer sequences

GCLiPP 3′ RNA linker: 5′-NNGUGUCUUUACACAGCUACGGCGUCG-3′

GCLiPP 5′ RNA linker: 5′-CGACCAGCAUCGACUCAGAAG-3′

GCLiPP Reverse transcription primer: 5′-CAAGCAGAAGACGGCATACGAGATNNNNNNCGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCGACGCCGTAGCTGTGTAAA-3′ (NNNNNN is barcode for demultiplexing)

GCLiPP 3′ PCR primer: 5′-CAAGCAGAAGACGGCATACGAGAT-3′

GCLiPP 5′ PCR primer: 5′-AATGATACGGCGACCACCGAGATCTACACTGGTACTCCGACCAGCATCGACTCAGAAG-3′

Read1seq sequencing primer for GCLiPP: 5′-ACACTGGTACTCCGACCAGCATCGACTCAGAAG-3′

Index sequencer primer for GCLiPP: 5′-

GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3′

Fast-UTR amplification forward primer: 5′-CAAGCAGAAGACGGCATACGAGAT NNNNNN

GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCTAGACGCGTAGGTTCAGA-3′

(NNNNNN is sample barcode for demultiplexing)

Fast-UTR reverse primer: 5′-

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC

T-3′

References

1. R. Reed, *Curr. Opin. Cell Biol.* **15**, 326–331 (2003).
2. K. C. Martin, A. Ephrussi, *Cell*. **136**, 719–730 (2009).
3. C. A. Beelman, R. Parker, *Cell*. **81**, 179–183 (1995).
4. J. D. Keene, *Nat. Rev. Genet.* **8**, 533–543 (2007).
5. A. Castello *et al.*, *Cell*. **149**, 1393–1406 (2012).
6. P. Kafasla, A. Skliris, D. L. Kontoyiannis, *Nat. Immunol.* **15**, 492–502 (2014).
7. G. J. Bassell, S. Kelic, *Curr. Opin. Neurobiol.* **14**, 574–581 (2004).
8. J. Schwerk, R. Savan, *J. Immunol.* **195**, 2963–2971 (2015).
9. R. E. Thurman *et al.*, *Nature*. **489**, 75–82 (2012).
10. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, *Nat. Methods*. **10**, 1213–1218 (2013).
11. D. M. Moskowitz *et al.*, *Sci. Immunol.*, in press, doi:10.1126/sciimmunol.aag0192.
12. A. Raghavan *et al.*, *Nucleic Acids Res.* **30**, 5529–5538 (2002).
13. A. Siepel *et al.*, *Genome Res.* **15**, 1034–1050 (2005).

14. L. Duret, F. Dorkeld, C. Gautier, *Nucleic Acids Res.* **21**, 2315–2322 (1993).

15. S. A. Shabalina, A. Y. Ogurtsov, D. J. Lipman, A. S. Kondrashov, *Nucleic Acids Res.* **31**, 5433–5439 (2003).

16. Y.-C. T. Yang *et al.*, *BMC Genomics*. **16**, 51 (2015).

17. M. A. Freeberg *et al.*, *Genome Biol.* **14**, R13 (2013).

18. A. G. Baltz *et al.*, *Mol. Cell*. **46**, 674–690 (2012).

19. D. D. Licatalosi *et al.*, *Nature*. **456**, 464–469 (2008).

20. S. Millevoi, S. Vagner, *Nucleic Acids Res.*, in press, doi:10.1093/nar/gkp1176.

21. R. Sandberg, J. R. Neilson, A. Sarma, P. A. Sharp, C. B. Burge, *Science*. **320**, 1643–1647 (2008).

22. E. H. Kislauskis, X. Zhu, R. H. Singer, *J. Cell Biol.* **127**, 441–451 (1994).

23. J. A. Chao *et al.*, *Genes Dev.* **24**, 148–158 (2010).

24. G. Bermano, R. K. Shepherd, Z. E. Zehner, J. E. Hesketh, *FEBS Lett.* **497**, 77–81 (2001).

25. J. Hesketh, G. Campbell, M. Piechaczyk, J. M. Blanchard, *Biochem. J.* **298**, 143–148 (1994).

26. E. W. Müllner, L. C. Kühn, *Cell*. **53**, 815–825 (1988).

27. E. W. Müllner, B. Neupert, L. C. Kühn, *Cell*. **58**, 373–382 (1989).

28. L. C. Kühn, *Metallomics*. **7**, 232–243 (2015).

29. R. C. Spitale *et al.*, *Nature*. **519**, 486–490 (2015).

30. L. V. Papp, J. Lu, A. Holmgren, K. K. Khanna, *Antioxid. Redox Signal.* **9**, 775–806 (2007).

31. L. Johansson, G. Gafvelin, E. S. J. Arnér, *Biochim. Biophys. Acta*. **1726**, 1–13 (2005).

32. M. J. Berry, L. Banu, J. W. Harney, P. R. Larsen, *EMBO J.* **12**, 3315–3322 (1993).

33. R. M. Tujebajeva *et al.*, *EMBO Rep.* **1**, 158–163 (2000).

34. M. Mariotti, A. V. Lobanov, R. Guigo, V. N. Gladyshev, *Nucleic Acids Res.* **41**, e149 (2013).

35. S. Lee *et al.*, *Cell*. **164**, 69–80 (2016).

36. A. Tichon *et al.*, *Nat. Commun.* **7**, 12209 (2016).

37. D. Ray *et al.*, *Nature*. **499**, 172–177 (2013).

38. B. M. Peterlin, J. E. Brogie, D. H. Price, *Wiley Interdiscip. Rev. RNA*. **3**, 92–103 (2012).

39. M. Marz *et al.*, *Mol. Biol. Evol.* **26**, 2821–2830 (2009).

40. K. Fujinaga, Z. Luo, B. M. Peterlin, *J. Biol. Chem.* **289**, 21181–21190 (2014).

41. M. Barboric *et al.*, *EMBO J.* **24**, 4291–4303 (2005).

42. R. A. Flynn *et al.*, *Nat. Struct. Mol. Biol.* **23**, 231–238 (2016).

43. W. Zhao *et al.*, *Nat. Biotechnol.* **32**, 387–391 (2014).

44. H. Goodarzi *et al.*, *Nature*. **485**, 264–268 (2012).

45. G. Vahedi *et al.*, *Cell*. **151**, 981–993 (2012).

46. J. A. Capra, M. J. Hubisz, D. Kostka, K. S. Pollard, A. Siepel, *PLOS Genet.* **9**, e1003684 (2013).

47. K. Leppek *et al.*, *Cell*. **153**, 869–881 (2013).

48. T. Uehata *et al.*, *Cell*. **153**, 1036–1049 (2013).

49. Y. Sugimoto *et al.*, *Nature*. **519**, 491–494 (2015).

50. E. Park, L. E. Maquat, *Wiley Interdiscip. Rev. RNA*. **4**, 423–435 (2013).

51. A. M. Zubiaga, J. G. Belasco, M. E. Greenberg, *Mol. Cell. Biol.* **15**, 2219–2230 (1995).

52. D. Beisang, B. Rattenbacher, I. A. Vlasova-St Louis, P. R. Bohjanen, *J. Biol. Chem.* **287**, 950–960 (2012).

53. F. Mignone, C. Gissi, S. Liuni, G. Pesole, *Genome Biol.*, in press, doi:10.1186/gb-2002-3-3-reviews0004.

54. D. F. Steiner *et al.*, *Immunity*. **35**, 169–181 (2011).

55. S. Nowotschin, A.-K. Hadjantonakis, *BMC Dev. Biol.* **9**, 49 (2009).

56. R. Lorenz *et al.*, *Algorithms Mol. Biol.* **6**, 26 (2011).

57. P. Kerpedjiev, S. Hammer, I. L. Hofacker, *Bioinforma. Oxf. Engl.* **31**, 3377–3379 (2015).

Acknowledgements:

We thank Matija Peterlin and Alex Marson for critical reading of the manuscript. AL was supported by Cancer Research Institute Irvington Fellowship and the UCSF Immunology T32 training grant T32AI007334. This work was supported by the US National Institutes of Health (HL107202, HL109102), the Sandler Asthma Basic Research Center, and a Scholar Award (K.M.A.) from The Leukemia & Lymphoma Society. The authors declare no conflicts of interest.

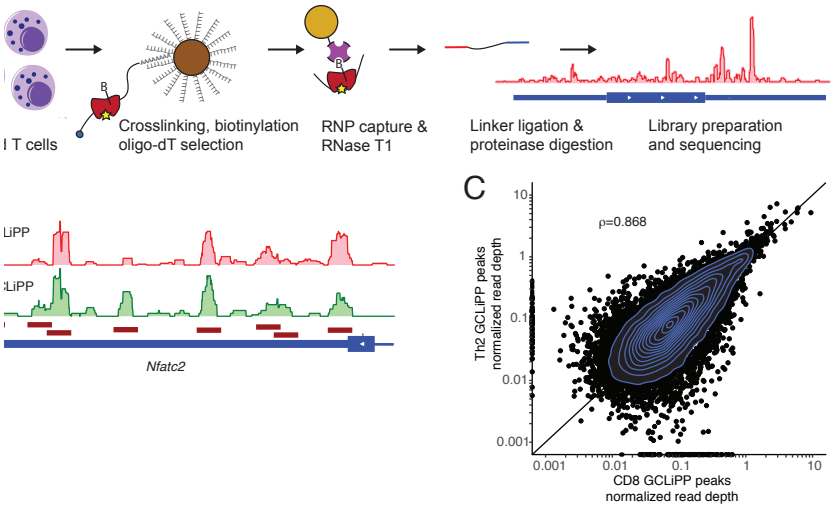

l T cells
Crosslinking, biotinylation
oligo-dT selection
RNP capture &
RNase T1
Linker ligation &
proteinase digestion
Library preparation
and sequencing
LiPP
LiPP
Nfatc2
C
ρ=0.868
Th2 GCLiPP peaks
normalized read depth
CD8 GCLiPP peaks
normalized read depth
10
1
0.1
0.01
0.001
0.001
0.01
0.1
1
10

Figure 1. GCLiPP sequencing reveals RNA transcript protein occupancy. (**A**) GCLiPP method of global RBP profiling. T cell RNAs are crosslinked to RBPs and lysates are biotinylated on primary amines. mRNAs are enriched with oligo-dT beads, and RBP protected sites are digested, captured, sequenced and aligned to the genome. (**B**) GCLiPP tracks for Th2 and CD8 T cells with peaks of GCLiPP read density indicated; each track represents combined data from three independent experiments in Th2 and two independent experiments in CD8 T cells. (**C**) Normalized GCLiPP read depth (fraction of reads in called peak relative to all GCLiPP reads in annotated 3′ UTR) in Th2 and CD8 T cells. ρ represents Pearson correlation.

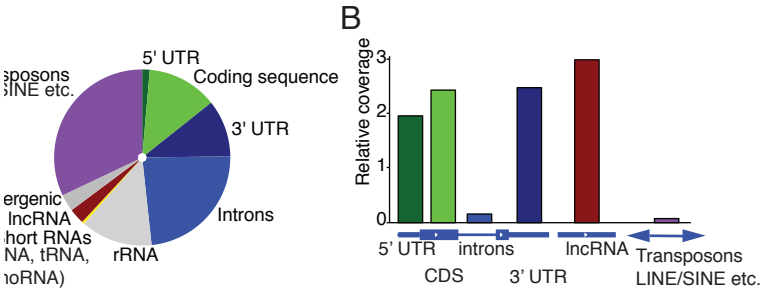

B
5' UTR
Coding sequence
3' UTR
Introns
rRNA
sposons
INE etc.
ergenic
lncRNA
hort RNAs
NA, tRNA,
oRNA)
Relative coverage
3
2
1
0
5' UTR
CDS
introns
3' UTR
lncRNA
Transposons
LINE/SINE etc.

Figure 1—figure supplement 1. Genomic origin of GCLiPP reads. (**A**) Genomic origin of aligned GCLiPP reads. (**B**) Relative coverage of genomic features in GCliPP sequencing reads.

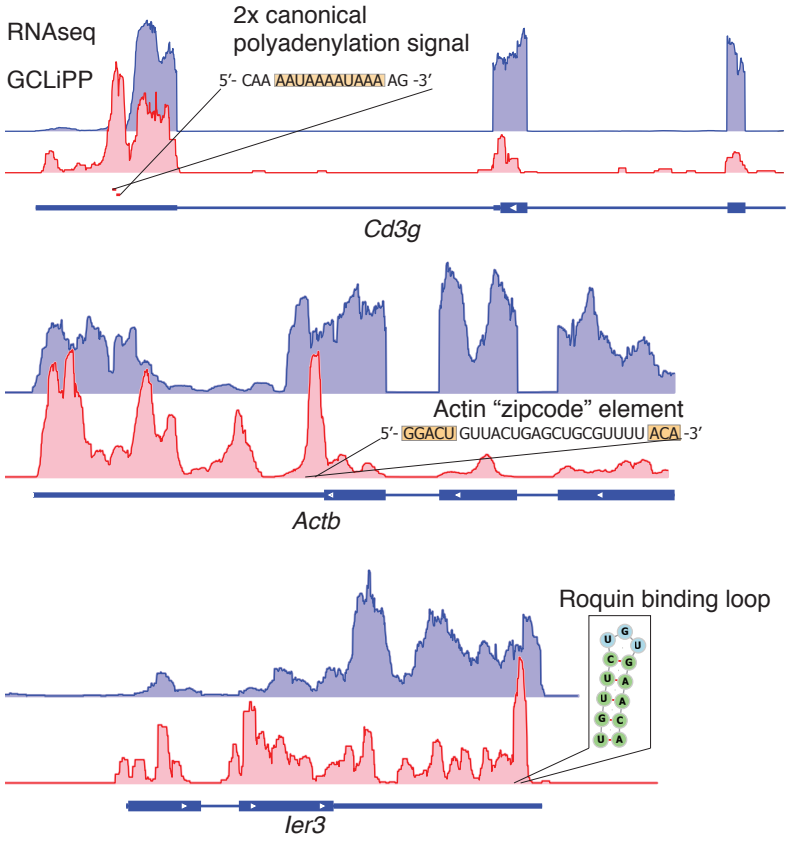

RNAseq
GCLiPP
2x canonical
polyadenylation signal
5′- CAA AAUAAAAUAAA AG -3′
Cd3g
Actin “zipcode” element
5′- GGACU GUUACUGAGCUGCGUUUU ACA -3′
Actb
Roquin binding loop
Ier3

Figure 2. GCLiPP recapitulates previously described mRNA-RBP interactions. RNAseq and GCLiPP tracks for **(A)** *Cd3g* (**B**) *Actb* **(C)** *Ier3*. RNAseq track is from resting Th2 cells. GCLiPP is sum of five experiments, three in Th2 and two in CD8 T cells. Location of known RBP binding determinants are shown as insets.

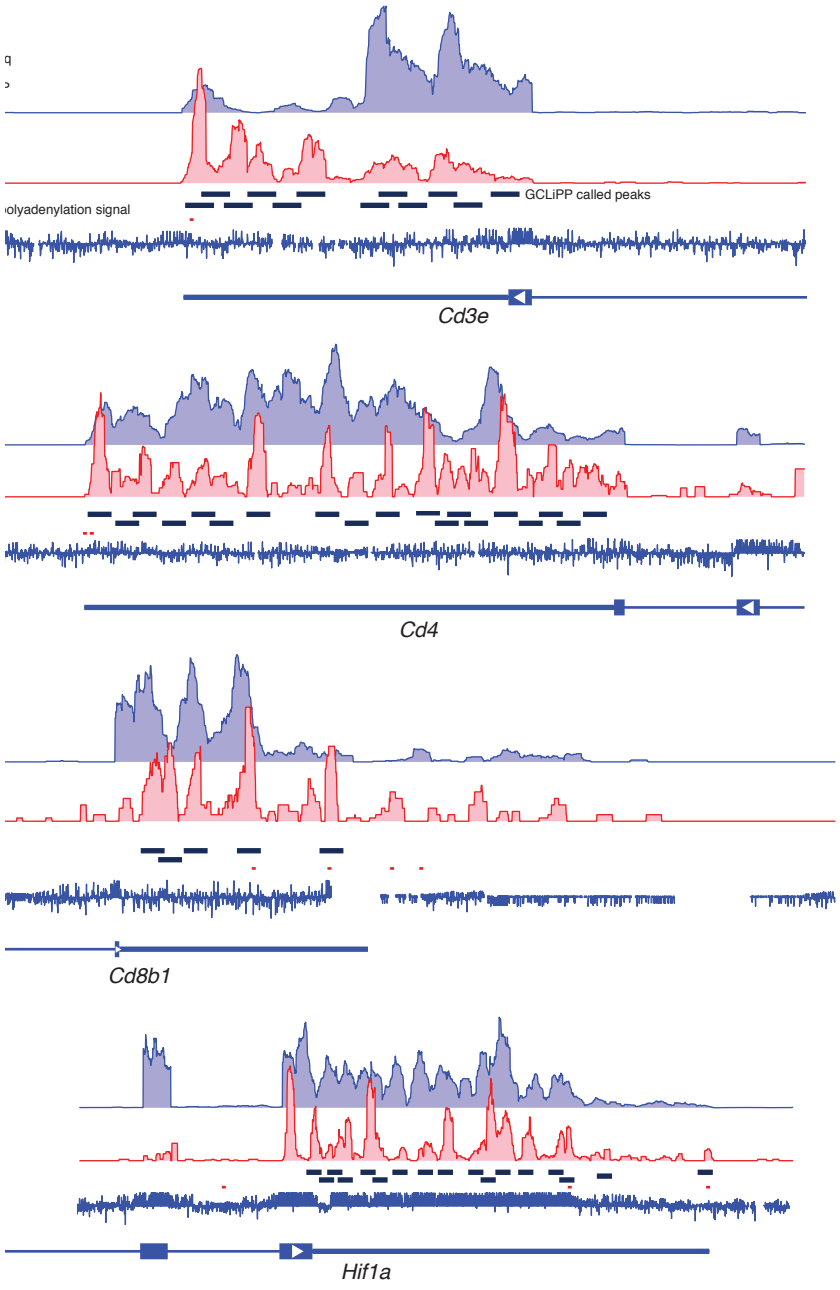

GCLiPP called peaks
olyadenylation signal
Cd3e
Cd4
Cd8b1
Hif1a

Figure 2—figure supplement 1. GCLiPP detects RBP binding of canonical polyadenylation signal. (**A-D**) RNAseq and GCLiPP read densities, conservation, called GCLiPP peaks and location of canonical polyadenylation signals (red lines) for indicated genes.

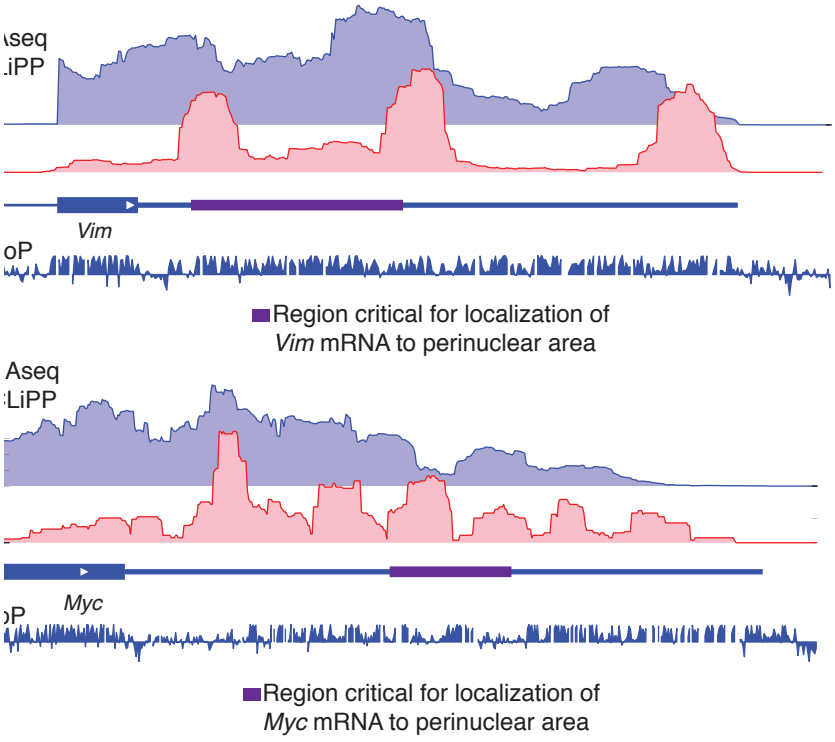

Vim
Region critical for localization of
Vim mRNA to perinuclear area
Myc
Region critical for localization of
Myc mRNA to perinuclear area

Figure 2—figure supplement 2. GCLiPP detects RBP binding in *cis* regulatory elements responsible for localizing *Vim* and *Myc* mRNAs. RNAseq read density, GCLiPP read density and conservation for (**A**) *Vim* and (**B**) *Myc* transcript. Previously identified regions of 3′ UTR critical for localization of a reporter to perinuclear area are depicted in purple.

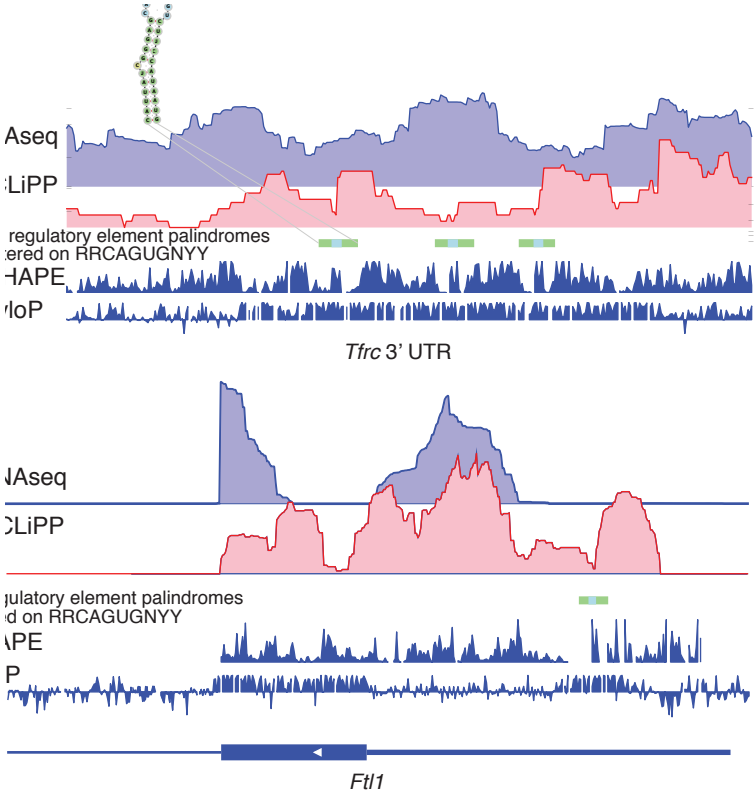

Aseq
LiPP
regulatory element palindromes
ered on RRCAGUGNYY
HAPE
loP
Tfrc 3' UTR
NAseq
CLiPP
gulatory element palindromes
d on RRCAGUGNYY
APE
P
Ftl1

Figure 2—figure supplement 3. Signature of RBP binding to iron responsive elements in *Tfrc* and *Ftl* mRNAs. RNAseq and GCLiPP read densities, icSHAPE tagging in mouse ES cells and conservation for (**A**) Transferrin receptor (*Tfrc*) and (**B**) ferritin light polypeptide 1 *Ftl1*. Iron response element binding motifs are indicated with green (stem) and blue (loop) lines. An example of IRE structure is shown.

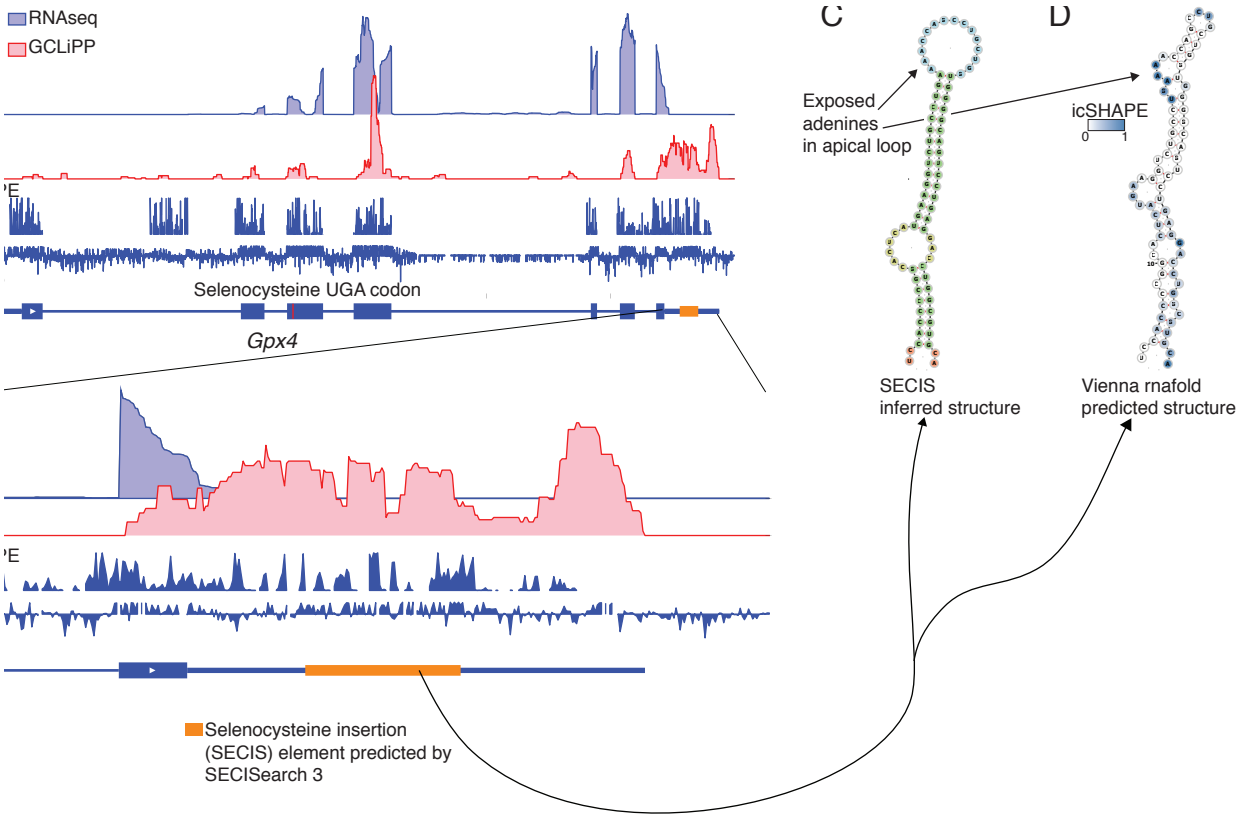

RNAseq
GCLiPP
Selenocysteine UGA codon
Gpx4
Selenocysteine insertion
(SECIS) element predicted by
SECISearch 3
C
D
Exposed
adenines
in apical loop
icSHAPE
0
1
SECIS
inferred structure
Vienna rnafold
predicted structure

Figure 2—figure supplement 4. Pervasive RBP binding in SECIS elements of the selenoprotein transcript *Gpx4* (**A**) RNAseq and GCLiPP read densities icSHAPE tagging in mouse ES cells and conservation for the selenoprotein encoding transcript *Gpx4*. Predicted SECIS element is depecting in orange. (**B**) Inset showing detail. Canonical polyadenylation signal (AAUAAA) is shown in red. (**C**) Schematic depiction of SECIS element based on conserved structural features as predicted by SECISearch. (**D**) forna predicted structure of SECIS sequence color coded with icSHAPE signal in mouse ES cells.

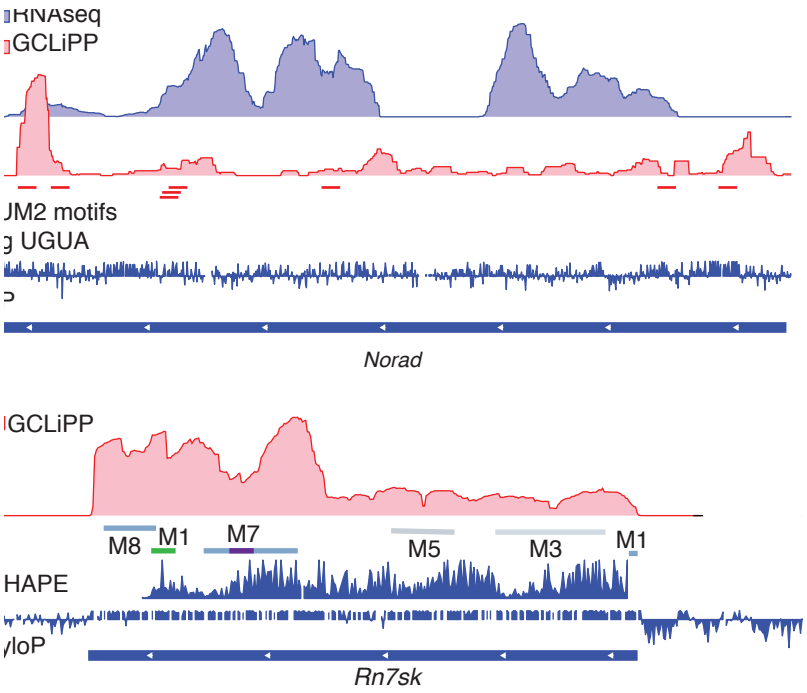

RNAseq
GCLiPP
JM2 motifs
g UGUA
Norad
GCLiPP
M8
M1
M7
M5
M3
M1
HAPE
loP
Rn7sk

Figure 3. **GCLiPP recapitulates previously described non coding RNA-RBP interactions**

(**A)** RNAseq and GCLiPP read densities and conservation for the *Norad* lncRNA. Red lines indicate the top scoring position weight matrix scores for PUM2 binding motifs containing UGUA. (**B**) GCLiPP read densities, icSHAPE tagging in mouse ES cells and conservation of the 7SK non-coding RNA. Location of conserved structural motifs within the gene is shown.

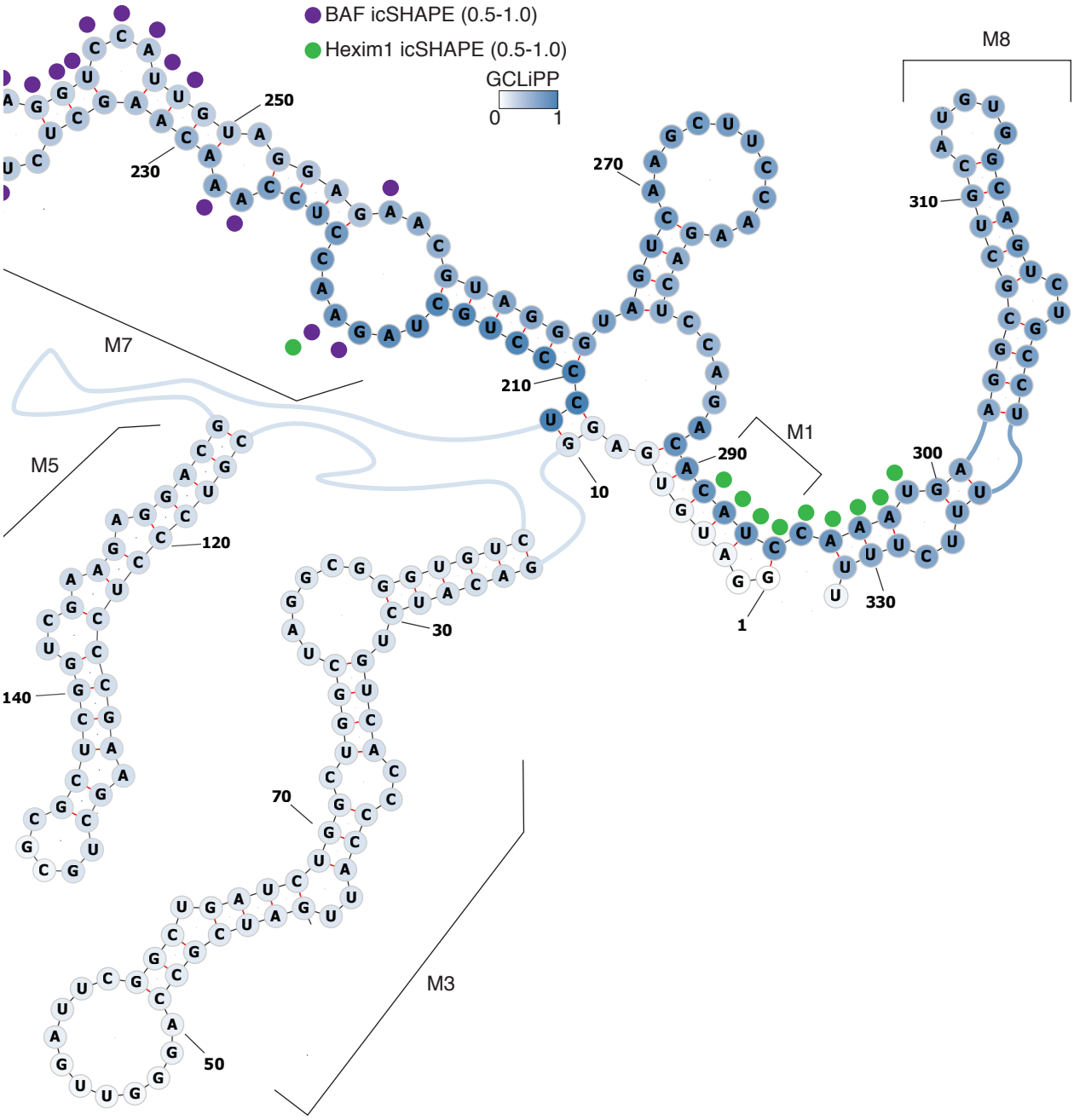

BAF icSHAPE (0.5-1.0)
Hexim1 icSHAPE (0.5-1.0)
GCLiPP
0
1
M1
M3
M5
M7
M8
1
10
30
50
70
120
140
210
230
250
270
290
300
310
330

Figure 3—figure supplement 1. Differential RBP binding between structural elements of *Rn7sk* non-coding RNA. Schematic depiction of folding of selected conserved structural motifs, color-coded by GCLiPP read density. Nucleotides with variable icSHAPE upon immunoprecipitation with antibodies against BAF or Hexim1 are depicted from Flynn et al. 2016.

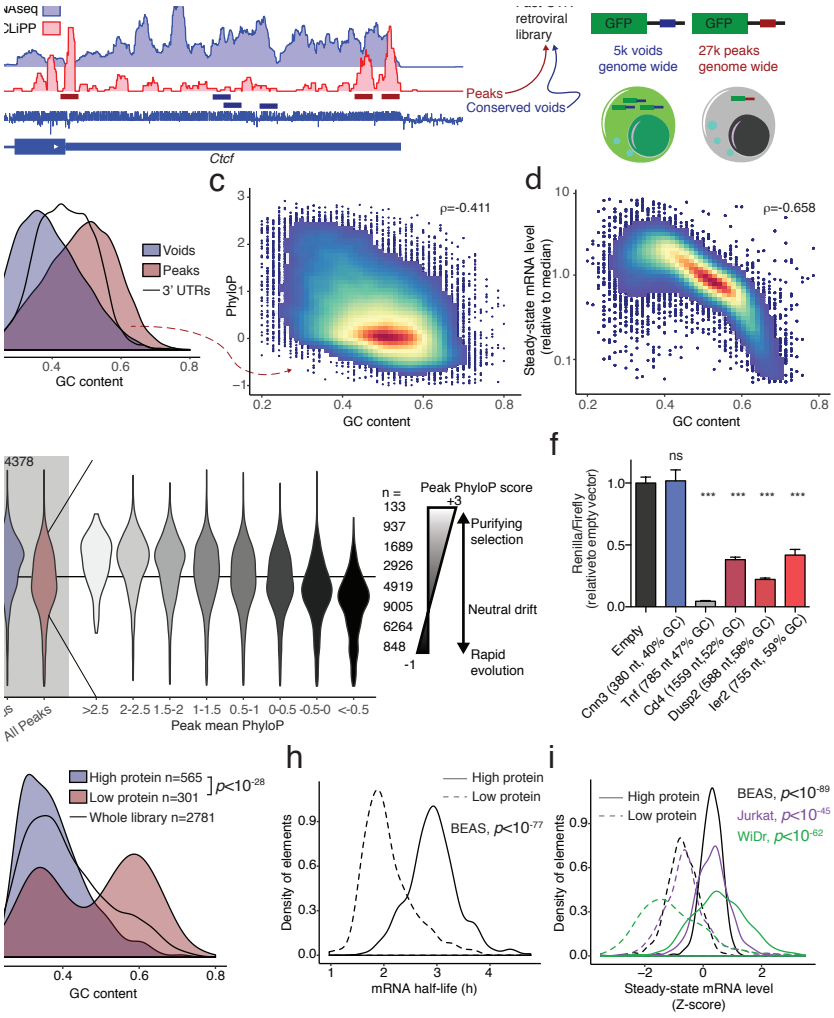

RNAseq
eCLiPP
Ctcf
Peaks
Conserved voids
retroviral library
GFP
GFP
5k voids genome wide
27k peaks genome wide
c
Voids
Peaks
3' UTRs
GC content
PhyloP
ρ=-0.411
d
Steady-state mRNA level (relative to median)
ρ=-0.658
GC content
4378
All Peaks
>2.5
2-2.5
1.5-2
1-1.5
0.5-1
0-0.5
-0.5-0
<-0.5
Peak mean PhyloP
n =
133
937
1689
2926
4919
9005
6264
848
Peak PhyloP score
+3
-1
Purifying selection
Neutral drift
Rapid evolution
f
Renilla/Firefly (relative to empty vector)
ns
***
Empty
Cnn3 (380 nt, 40% GC)
Tnf (785 nt 47% GC)
Cd4 (1559 nt,52% GC)
Dusp2 (588 nt,58% GC)
Ier2 (755 nt, 59% GC)
High protein n=565
Low protein n=301
Whole library n=2781
p<10-28
GC content
h
High protein
Low protein
BEAS, p<10-77
Density of elements
mRNA half-life (h)
i
High protein
Low protein
BEAS, p<10-89
Jurkat, p<10-45
WiDr, p<10-62
Density of elements
Steady-state mRNA level (Z-score)

Figure 4. RBP occupied sequences are GC rich, rapidly evolving and destabilize reporter mRNAs. (**A**) Example calling of GCLiPP peaks and conserved voids in *Ctcf*. PhyloP measures conservation across placental mammals. (**B**) GC content of GCLiPP peaks and conserved voids. (**C**) Relationship between GC content and conservation for GCliPP peaks. ρ represents Pearson correlation. (**D**) Relationship between GC content and steady state mRNA level in Fast-UTR reporter assay for GCLiPP peaks. (**E**) Steady state mRNA level in Fast-UTR of conserved voids and GCLiPP peaks, binned on placental mammal conservation. (**F**) Dual luciferase assay showing renilla luciferase activity relative to firefly luciferase activity, in Th2 cells transfected with plasmid with indicated 3′ UTR downstream of renilla luciferase gene and control firefly luciferase gene. A representative experiment using four replicate mice is shown. Mean and standard error of the mean are indicated by bar graph and error bars, respectively. ***, $p < 0.0001$ in unpaired t test relative to empty vector. (**G**) GC content of 3′ UTR inserts of Fast-UTR library transduced BEAS cells FACS-sorted for high or low GFP fluorescence. (**H**) mRNA half life of inserts of high or low protein expressing 3′ UTR inserts from (**G**) in BEAS cells. (**I**) Steady state mRNA level of inserts of high or low protein expressing 3′ UTR inserts from (**G**) in three human cell lines. For (**G**) through (**I**) *p* values represent Welch's unequal variance t-test between high and low protein expressing inserts.

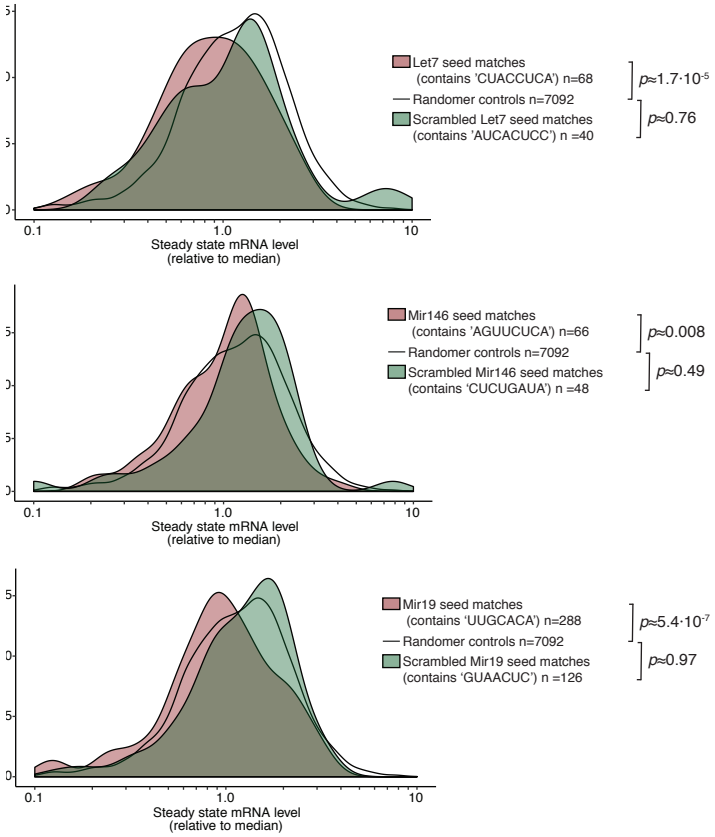
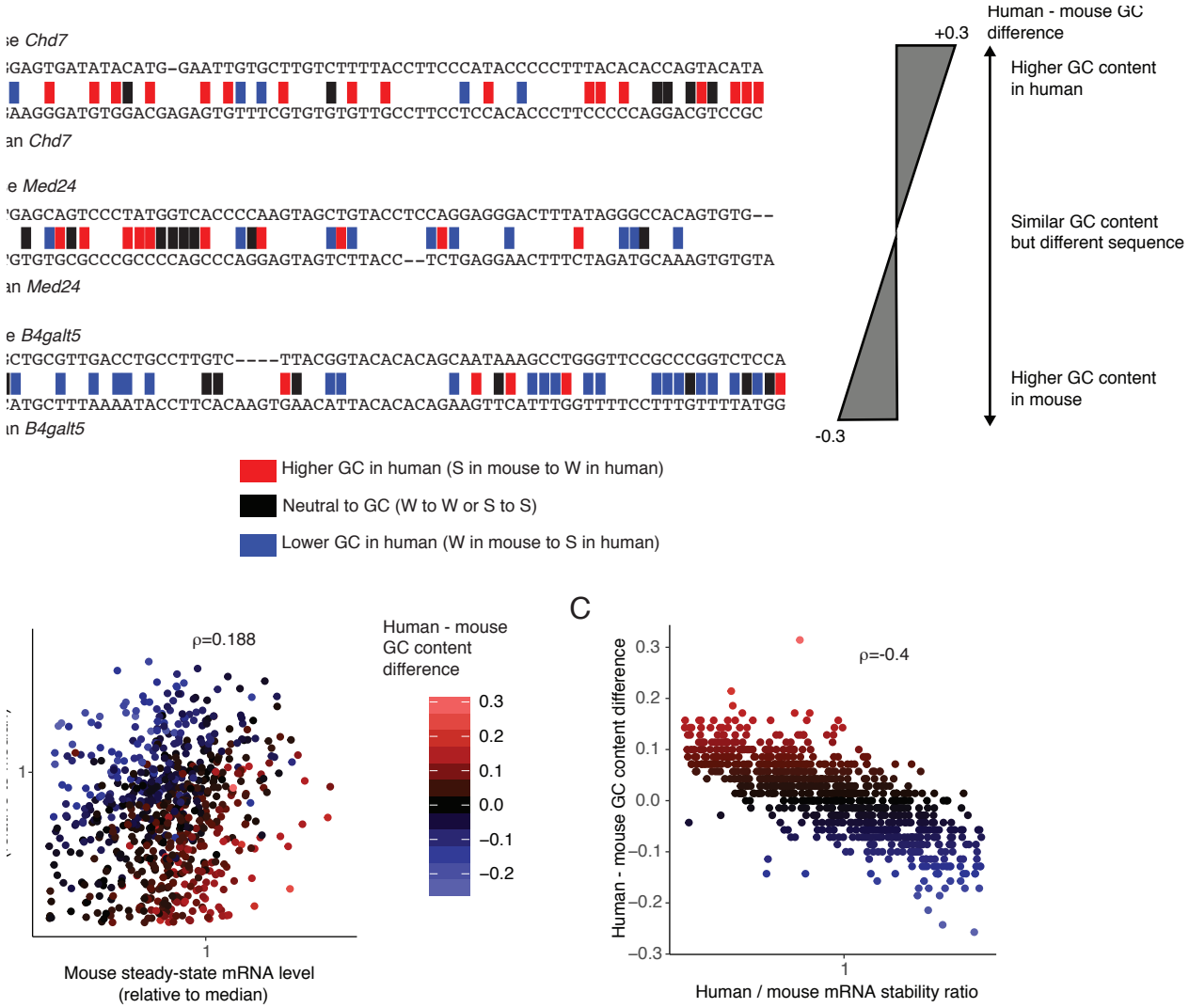

Let7 seed matches
(contains 'CUACCUCA') n=68
Randomer controls n=7092
Scrambled Let7 seed matches
(contains 'AUCACUCC') n =40
p≈1.7·10-5
p≈0.76
0.1
1.0
10
Steady state mRNA level
(relative to median)
Mir146 seed matches
(contains 'AGUUCUCA') n=66
Randomer controls n=7092
Scrambled Mir146 seed matches
(contains 'CUCUGAUA') n =48
p≈0.008
p≈0.49
0.1
1.0
10
Steady state mRNA level
(relative to median)
Mir19 seed matches
(contains 'UUGCACA') n=288
Randomer controls n=7092
Scrambled Mir19 seed matches
(contains 'GUAACUC') n =126
p≈5.4·10-7
p≈0.97
0.1
1.0
10
Steady state mRNA level
(relative to median)

Figure 4—figure supplement 1. Sequences containing highly expressed miRNA seeds have lower mRNA levels than scrambled controls in Fast-UTR reporter assay. Density plots showing median normalized steady state mRNA level for all inserts in the fast-UTR library containing specific miRNA seed sequences or indicated scrambled controls, shown against a background of all randomer control sequences (**A**) Let7 seed containing sequences. (**B**) Mir146 seed containing sequences. (**C**) Mir19 seed containing sequences.


e Chd7
GAGTGATATACATG-GAATTGTGCTTGTCTTTTACCTTCCCATACCCCCTTTACACACCAGTACATA
AAGGGATGTGGACGAGAGTGTTTCGTGTGTGTTGCCTTCCTCCACACCCTTCCCCCAGGACGTCCGC
an Chd7
e Med24
GAGCAGTCCCTATGGTCACCCCAAGTAGCTGTACCTCCAGGAGGGACTTTATAGGGCCACAGTGTG--
GTGTGCGCCCGCCCCAGCCCAGGAGTAGTCTTACC--TCTGAGGAACTTTCTAGATGCAAAGTGTGTA
an Med24
e B4galt5
CTGCGTTGACCTGCCTTGTC----TTACGGTACACACAGCAATAAAGCCTGGGTTCCGCCCGGTCTCCA
ATGCTTTAAAATACCTTCACAAGTGAACATTACACACAGAAGTTCATTTGGTTTTCCTTTGTTTTATGG
an B4galt5
Higher GC in human (S in mouse to W in human)
Neutral to GC (W to W or S to S)
Lower GC in human (W in mouse to S in human)
Human - mouse GC difference
+0.3
-0.3
Higher GC content in human
Similar GC content but different sequence
Higher GC content in mouse

ρ=0.188
Human - mouse GC content difference
0.3
0.2
0.1
0.0
−0.1
−0.2
1
1
Mouse steady-state mRNA level (relative to median)
C
ρ=-0.4
Human - mouse GC content difference
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
1
Human / mouse mRNA stability ratio

Figure 4—figure supplement 2. Changes in GC content between mouse and human protein binding sites determine mRNA stability in Fast-UTR reporter assay. (**A**) Example alignments of rapidly evolving mouse GCLiPP peaks and human syntenic regions showing higher (*Chd7*) similar (*Med24*) or lower (*B4galt5*) GC content in the mouse sequence. Correlation between (**B**) steady state mRNA levels for rapidly evolving mouse GCLiPP peaks and corresponding human syntenic regions in mouse T cell fast-UTR assay and (**C**) difference in GC content between rapidly evolving mouse GCLiPP peaks and syntenic human regions and ratio of fast-UTR steady state mRNA level for the same. ρ represents Pearson correlation.

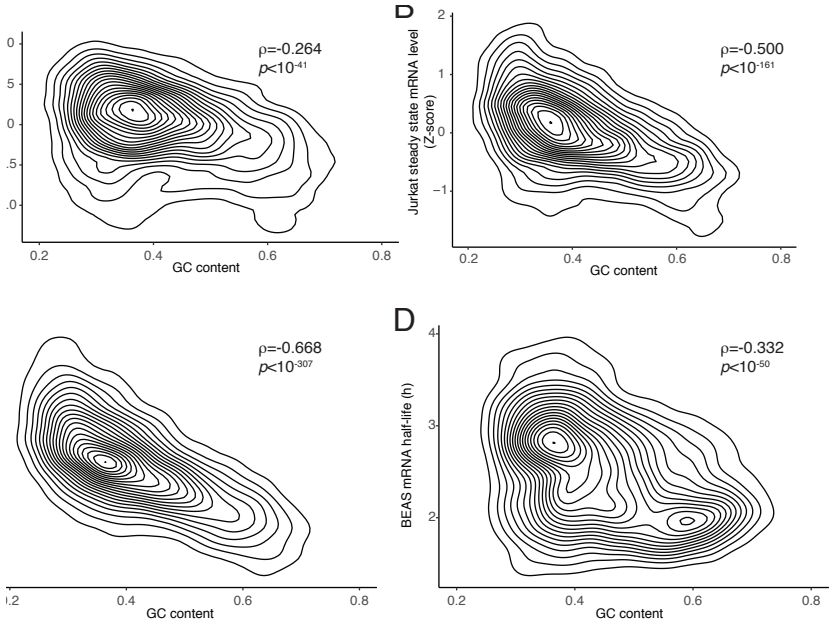

ρ=-0.264
p<10-41
GC content
B
Jurkat steady state mRNA level (Z-score)
ρ=-0.500
p<10-161
GC content
ρ=-0.668
p<10-307
GC content
D
BEAS mRNA half-life (h)
ρ=-0.332
p<10-50
GC content

Figure 4—figure supplement 3. Anticorrelation between GC content and gene expression in four Fast-UTR experiments in three cell lines. Correlation between GC content and gene expression in four fast-UTR experiments in different human cell lines (**A**) Steady-state mRNA level in BEAS cells (**B**) Steady-state mRNA level in Jurkat cells (**C**) Steady-state mRNA level in WiDr cells (**D**) mRNA half-life in BEAS cells. ρ represents Pearson correlation.

Figure 4—source data 1. Sequences and data for fast-UTR library members. This table is associated with Figures 2 and 3, showing all of the GCLiPP peaks, conserved voids, human syntenic regions and randomer controls that comprised the fast-UTR library shown in Figure 2B-2E. Table columns are defined in the table and include data about the fast-UTR expression, insert sequence, nucleotide composition, conservation and in vivo folding of each sequence. icSHAPE standard deviation is associated with Figure 3C.

Figure 4—source data 2. Custom peak calling script. This perl script calls regions of local GCLiPP read density or conserved areas with low GCLiPP read density but expressed in RNAseq.

Figure 4—source data 3. Fast-UTR scoring script. This perl script determines the ratio of RNA to DNA reads for each barcoded insert given an alignment of sequence reads to the library.

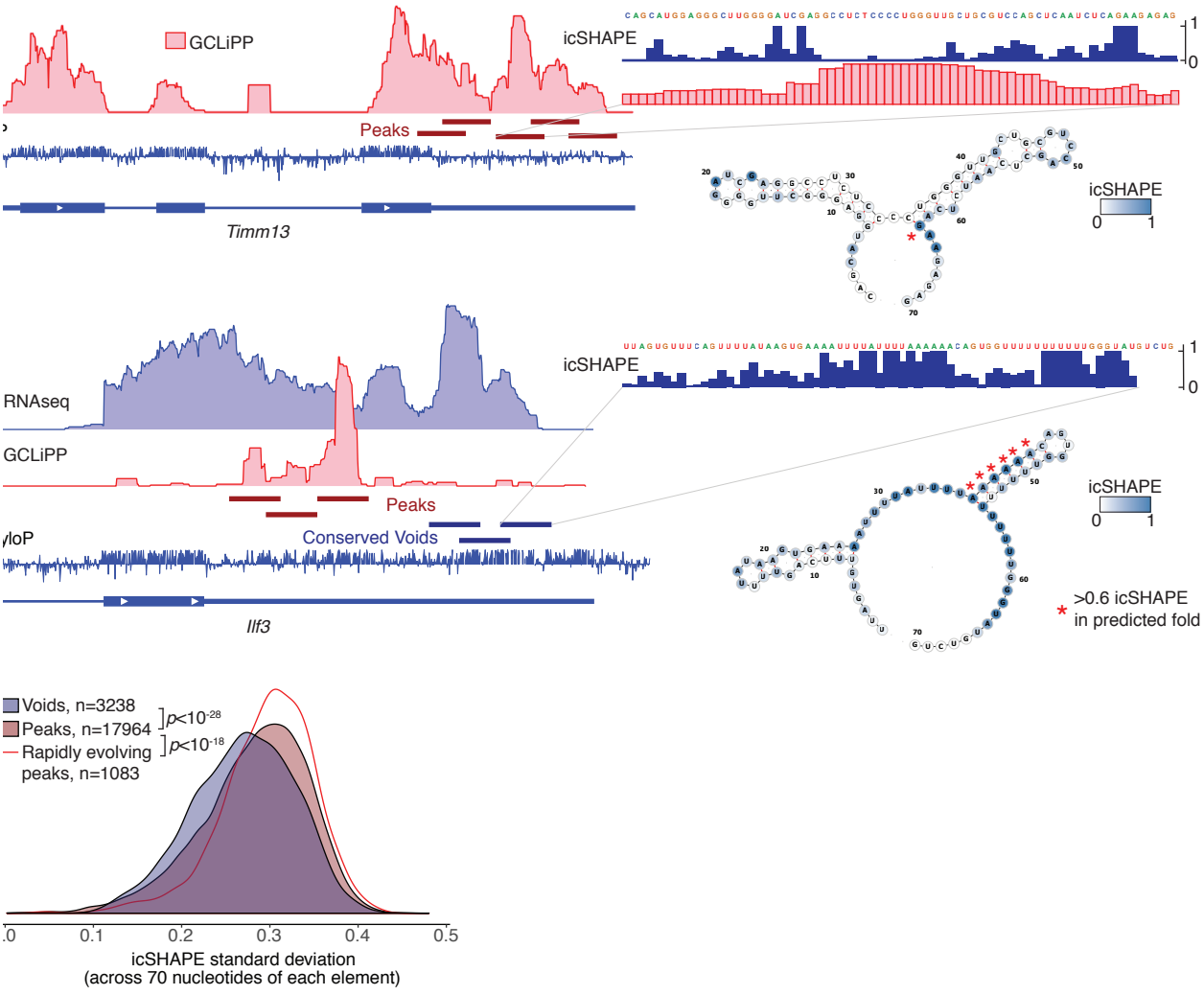

GCLiPP
Peaks
Timm13
icSHAPE
RNAseq
GCLiPP
Peaks
Conserved Voids
Ilf3
icSHAPE
icSHAPE
>0.6 icSHAPE
in predicted fold
Voids, n=3238
Peaks, n=17964
Rapidly evolving
peaks, n=1083
p<10-28
p<10-18
0.0
0.1
0.2
0.3
0.4
0.5
icSHAPE standard deviation
(across 70 nucleotides of each element)

Figure 5. GC rich RBP occupied sequences are folded in vivo. (**A**) *Timm3* locus showing GCLiPP and conservation tracks; the indicated peak is shown as an inset depicting nucleotide sequence, icSHAPE signal, GCLiPP read depth and predicted RNA structure. Colors on predicted structure shows icSHAPE signal intensity. (**B**) The same data for the *Ilf3* locus with detailing a conserved void, with high icSHAPE signal nucleotides in a predicted hairpin (red asterisks). (**C**) Standard deviation of icSHAPE signal across the 70 nucleotides of the indicated classes of elements. *P* values represent Welch's t-tests between the indicated groups

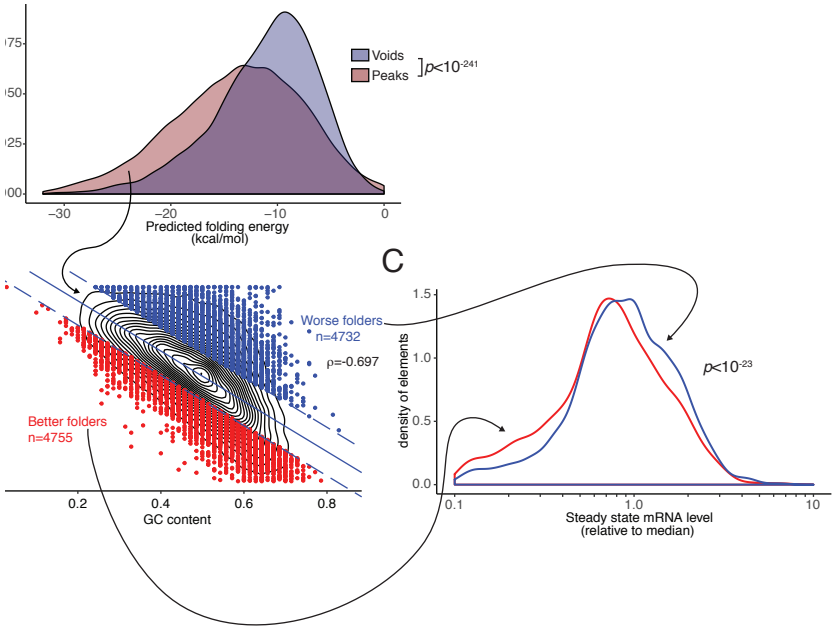

Voids
Peaks
p<10-241
Predicted folding energy
(kcal/mol)
C
Worse folders
n=4732
ρ=-0.697
Better folders
n=4755
GC content
density of elements
p<10-23
Steady state mRNA level
(relative to median)

Figure 5—figure supplement 1. Sequences with lower than expected folding energy based on GC content are associated with lower gene expression. (**A**) rnafold predicted folding energies for GCLiPP peaks and conserved voids. *p* value represents Welch's unequal variance t-test. (**B**) Correlation between GC content and predicted folding energy for GCLiPP peaks. Linear regression and ~15% outliers above and below regression line. ρ represents Pearson correlation (**C**) Steady-state mRNA level in mouse T cell fast-UTR assay for outliers of folding that are better or worse than regression. *p* value represents Welch's unequal variance t-test.

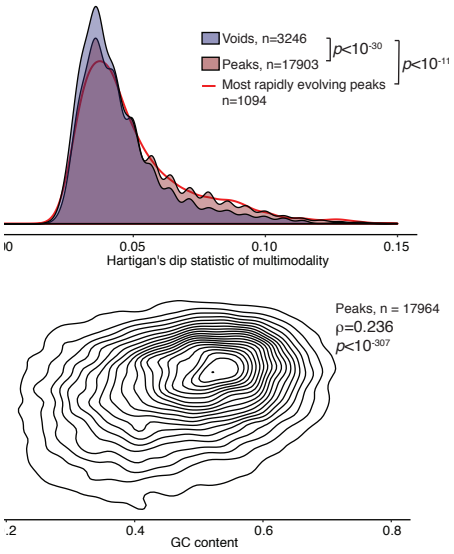
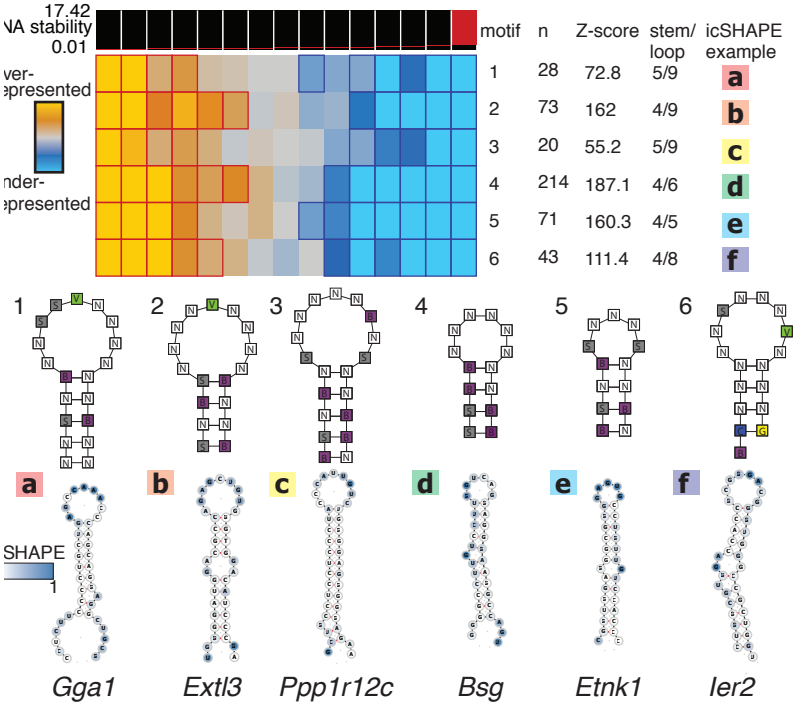

Voids, n=3246
Peaks, n=17903
Most rapidly evolving peaks
n=1094
p<10-30
p<10-11
0
0.05
0.10
0.15
Hartigan's dip statistic of multimodality
Peaks, n = 17964
ρ=0.236
p<10-307
.2
0.4
0.6
0.8
GC content

Figure 5—figure supplement 2. Rapidly evolving 3′ UTR segments are more likely to be folded in vivo (**A**) Hartigan's dip statistic of multimodality for conserved voids, GCLiPP peaks and rapidly evolving GCLiPP peaks. *p* values represent Welch's unequal variance t-test for the indicated comparisons. (**B)** Correlation between GCLiPP peak GC content and icSHAPE standard deviation. ρ represents Pearson correlation.


17.42
NA stability
0.01
ver-
epresented
nder-
epresented
motif
n
Z-score
stem/
loop
icSHAPE
example
1 28 72.8 5/9 a
2 73 162 4/9 b
3 20 55.2 5/9 c
4 214 187.1 4/6 d
5 71 160.3 4/5 e
6 43 111.4 4/8 f
SHAPE
1
Gga1
Extl3
Ppp1r12c
Bsg
Etnk1
Ier2

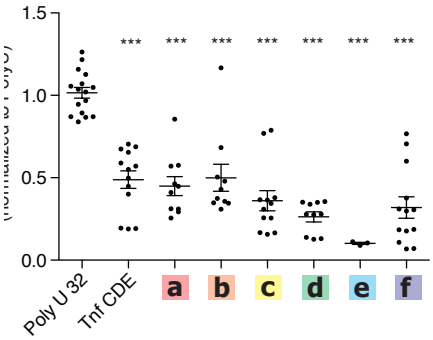

1.5
1.0
0.5
0.0
***
Poly U 32
Tnf CDE
a
b
c
d
e
f

Figure 6. TEISER identifies in vivo folded 3′ UTR structural motifs that inhibit gene expression (**A**) TEISER analysis identifies structural motifs enriched in destabilizing sequences. Columns show enrichment of motifs in deciles of GCLiPP peaks arranged by Fast-UTR steady-state mRNA level, rows represent individual motifs. (**B**) Generic motif structures and (**C**) a predicted structure for an example of each motif is depicted with icSHAPE signal indicated by color. (**D**) TEISER identified motifs lower gene expression. Kikume fluorescent protein synthesis in CD8+ T cells transfected with in vitro transcribed mRNAs with the indicated sequence inserted downstream of the stop codon. Data represent transfections of a single construct into the T cells from a single mouse pooled from 1-4 experiments using three mice each, with mean and standard error of the mean are indicated by line and error bars, respectively. ***, $p < 0.0001$ in unpaired t test relative to poly-U.

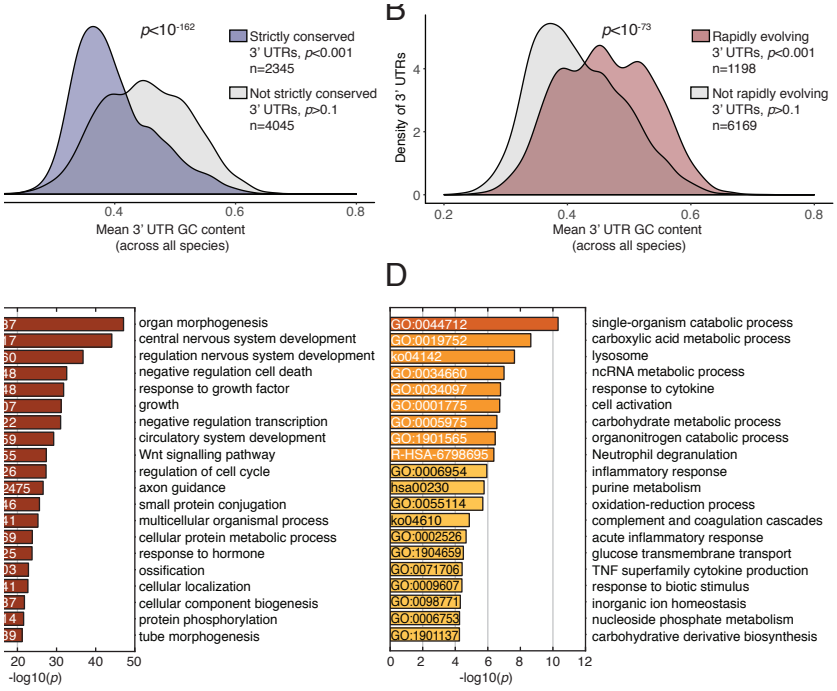

B
p<10-162
Strictly conserved
3' UTRs, p<0.001
n=2345
Not strictly conserved
3' UTRs, p>0.1
n=4045
0.4
0.6
0.8
Mean 3' UTR GC content
(across all species)
Density of 3' UTRs
0
2
4
p<10-73
Rapidly evolving
3' UTRs, p<0.001
n=1198
Not rapidly evolving
3' UTRs, p>0.1
n=6169
0.2
0.4
0.6
0.8
Mean 3' UTR GC content
(across all species)
D
organ morphogenesis
central nervous system development
regulation nervous system development
negative regulation cell death
response to growth factor
growth
negative regulation transcription
circulatory system development
Wnt signalling pathway
regulation of cell cycle
axon guidance
small protein conjugation
multicellular organismal process
cellular protein metabolic process
response to hormone
ossification
cellular localization
cellular component biogenesis
protein phosphorylation
tube morphogenesis
20
30
40
50
-log10(p)
GO:0044712
GO:0019752
ko04142
GO:0034660
GO:0034097
GO:0001775
GO:0005975
GO:1901565
R-HSA-6798695
GO:0006954
hsa00230
GO:0055114
ko04610
GO:0002526
GO:1904659
GO:0071706
GO:0009607
GO:0098771
GO:0006753
GO:1901137
single-organism catabolic process
carboxylic acid metabolic process
lysosome
ncRNA metabolic process
response to cytokine
cell activation
carbohydrate metabolic process
organonitrogen catabolic process
Neutrophil degranulation
inflammatory response
purine metabolism
oxidation-reduction process
complement and coagulation cascades
acute inflammatory response
glucose transmembrane transport
TNF superfamily cytokine production
response to biotic stimulus
inorganic ion homeostasis
nucleoside phosphate metabolism
carbohydrative derivative biosynthesis
0
2
4
6
8
10
12
-log10(p)

Figure 7. **Genes with strictly conserved, AT rich 3′ UTRs and rapidly evolving, GC rich 3′ UTRs represent different biological categories**. (**A**) Mean GC content of 3′ UTRs across 10 vertebrate species amongst genes that were found to be strictly conserved in a multiple sequence alignment versus other genes or (**B**) the same data comparing genes that were rapidly evolving in a multiple sequence alignment versus other genes. *p* values represent Welch's unequal variance t-test between genes that exhibit strong evidence of conservation/rapid evolution and those that do not. (**C**) Enriched gene ontology categories for genes with strictly conserved 3′ UTRs or (**D**) rapidly evolving 3′ UTRs. *p* values are for enrichment of the indicated GO category

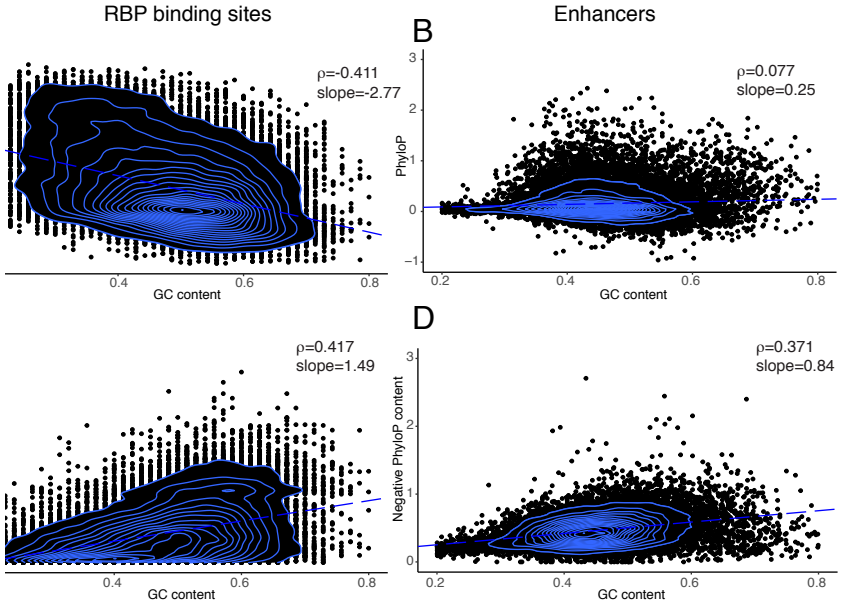

RBP binding sites
Enhancers
B
ρ=-0.411
slope=-2.77
ρ=0.077
slope=0.25
PhyloP
GC content
D
ρ=0.417
slope=1.49
ρ=0.371
slope=0.84
Negative PhyloP content
GC content

Figure 7—figure supplement 1. RBP sites have greater selection on GC content than enhancers. (**A**) and (**B**) Correlation between GC content and conservation (mean phyloP score) for (**A**) GCLiPP peaks and (**B**) p300 peaks in Th2 cells. (**C**) and (**D**) Correlation between GC content and rapid evolution (negative phyloP content) for (**C**) GCLiPP peaks and (**D**) p300 peaks in Th2 cells. ρ represents Pearson correlation.

Figure 7—source data 1. **Summary of vertebrate 3′ UTR multiple alignments**. This table is associated with Figure 4. This is a list of genes for which enough unambiguous annotated 3′ UTRs were found across the 10 vertebrate genomes (at least 4) to perform a multiple alignment. For each gene symbol, the number of aligned sequences, mean and standard deviation of GC content of aligned sequences, and the –log(p) value for strict conservation and accelerated evolution computed by the phyloP program are shown.

Figure 7—source data 2. Metascape gene list analysis report for genes with strictly conserved 3′ UTRs. This table is associated with Figure 4A and C. Metascape report showing enriched biological categories of genes with strictly conserved 3′ UTRs across 10 vertebrate species.

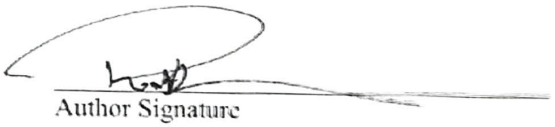
Figure 7—source data 3. Metascape gene list analysis report for genes with rapidly evolving 3′ UTRs. This table is associated with Figure 4B and D. Metascape report showing enriched biological categories of genes with accelerated evolution in 3′ UTRs across 10 vertebrate species.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature

6/16/17

Date