The specificity and evolution of gene regulatory elements

by

Robin Carl Friedman

B.S., University of California, San Diego (2005)

Submitted to the Computational and Systems Biology Program
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2010

Author . . . .
Computational and Systems Biology Program
June 17, 2010

Certified by . . . .
Christopher B. Burge
Professor
Thesis Supervisor

Certified by . . . .
David P. Bartel
Professor
Thesis Supervisor

Accepted by . . . .
Christopher B. Burge
Computational and Systems Biology Ph.D. Program Director

The specificity and evolution of gene regulatory elements

by

Robin Carl Friedman

Submitted to the Computational and Systems Biology Program
on June 17, 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The regulation of gene expression underlies the morphological, physiological, and functional differences between human cell types, developmental stages, and healthy and disease states. Gene regulation in eukaryotes is controlled by a complex milieu including transcription factors, microRNAs (miRNAs), *cis*-regulatory DNA and RNA. It is the quantitative and combinatorial interactions of these regulatory elements that defines gene expression, but these interactions are incompletely understood. In this thesis, I present two new methods for determining the quantitative specificity of gene regulatory factors. First, I present a comparative genomics approach that utilizes signatures of natural selection to detect the conserved biological relevance of miRNAs and their targets. Using this method, I quantify the abundance of different conserved miRNA target types, including different seed matches and 3′-compensatory targets. I show that over 60% of mammalian mRNAs are conserved targets of miRNAs and that a surprising amount of conserved miRNA targeting is mediated by seed matches with relatively low efficacy. Extending this method from mammals to other organisms, I find that miRNA targeting rules are mostly conserved, although I show evidence for new types of miRNA targets in nematodes. Taking advantage of variations in 3′ UTR lengths between species, I describe general properties of miRNA targeting that are affected by 3′ UTR length. Finally, I introduce a new, high-throughput assay for the quantification of transcription factor *in vitro* binding affinity to millions of sequences. I apply this method to *GCN4*, a yeast transcription factor, and reconstruct all known properties of its binding preferences. Additionally, I discover some new subtleties in its specificity and estimate dissociation constants for hundreds of thousands of sequences. I verify the utility of the binding affinities by comparing to *in vivo* binding data and to the regulatory repsonse following *GCN4* induction.

Thesis Supervisor: Christopher B. Burge
Title: Professor

Thesis Supervisor: David P. Bartel
Title: Professor

Acknowledgments

Thanks first to my advisors, Christopher Burge and David Bartel, for letting me make mistakes, lending help when I needed it, and for teaching me to think like a scientist. After being at MIT for several years, I have come to realize what a privilege it is to learn to become a scientist in such a rich intellectual incubator. In such an environment it is easy to stay humble and to appreciate all of those who contributed to this thesis. My committee members, Philip Sharp and Ernest Fraenkel, also provided many helpful suggestions and have been remarkably warm and supportive over the years. My thanks go the US Department of Energy and the Krell institute for providing years of funding, which gave me confidence to explore my own interests. The work presented in this thesis would not have been possible without numerous collaborators, most of whom I have tried to acknowledge in the relevant chapters. Here I would like to particularly mention the Whitehead institute bioinformatics group, especially George Bell, for their excellent work on the TargetScan website. Despite all the wonderful faculty, probably the best thing about being at MIT is interacting with the incredibly intelligent students in and outside of your lab from day to day. Many thanks to all past and present members of the Burge lab, Bartel lab, and CSB PhD program for creating a wonderful community here. Thanks in particular Lawrence David, Jesse Shapiro, Charles Lin, Bjorn Millard, Razvan Nutiu, and Dima Ter-Ovanesyan for serving as friends that I also admire. My thanks also go to my latin dance teacher, Armin Kappacher, and my lovely dance partner Brittany Low, for taking me on an emotional and physical journey to parallel my intellectual one. Most importantly, I have to thank my family: Mom, Dad, Julia, Alex, and Tara. Your support and love make me feel like I can do anything.

Robin Friedman
June 3rd, 2010

Contents

Chapter 1
Introduction

Chapter 1

Introduction

1.1 Overview

A central goal of systems biology is to utilize the data enabled by new technologies, such as high-throughput sequencing, to further the understanding of biological processes. Although the amount of this data has increased dramatically in recent years, it has not been easy to translate data into quantitative or mechanistic insights into biological processes such as the regulation of gene expression. This thesis is presented with the underlying belief that the regulation of gene expression can be better understood using new methods that leverage large amounts of data to determine quantitative regulatory interactions. Importantly, a quantitative catalog of regulatory interactions is useful for any approach to understanding gene expression. It can constrain computational models of systems, help with interpretation of high-throughput screens, and provide candidates for low-throughput experimentation.

With the goal of learning generalizable principles of gene regulation, I present here both computational and experimental methods for determining the specificity of trans-acting regulatory factors, and explore some ways that this data can lead to mechanistic and evolutionary insight into the underlying biology. The rest of chapter 1 introduces current understanding of gene regulation with an emphasis on transcription factor and microRNA (miRNA) specificity. Chapter 2 describes in detail a computational method for using comparative genomics to predict the targets of miRNAs and learn the principles of their specificity. Chapter 3 extends this method from vertebrates to other clades and discusses some insights into the co-evolution of miRNAs and 3′ UTRs. Chapter 4 introduces a new experimental method for the high-throughput quantitative determination of sequence-specific transcription factor binding affinity and applies it to gain new insights into the function of *GCN4*, a yeast transcription factor. Finally, chapter 5 provides a brief discussion of the implications of the thesis work and suggests some steps for extending this research.

1.2 Regulation by transcription factors

1.2.1 Gene regulation

Despite having the same genome, cells in a human body can have amazingly different morphology, physiology, and function. These varied properties depend chiefly on the proteins within each cell. As a result, levels of gene expression are regulated in a cell-specific manner and are subject to strong purifying selection (Gilad et al., 2006; Xie et al., 2005), suggesting that disregulation could lead to disease. In fact, there are countless diseases driven by the misexpression of genes. For example, the transcription factor *RUNX1* has been implicated in over 30 different translocations that lead to acute leukemia (Blyth et al., 2005). Surprisingly, both dominant negative fusions and dominant overexpressions of *RUNX1* can be oncogenic, suggesting transcriptional effects that are cell-type dependent (Blyth et al., 2005). When beneficial, regulatory changes can contribute to adaptation and speciation. For example, it has long been puzzling that so few protein-coding differences separate chimpanzees and humans, despite significant phenotypic differences (King and Wilson, 1975). However, more substantial differences in gene expression could explain this paradox (Enard et al., 2002). Another striking example is the *KITLG* gene, which has undergone regulatory changes in stickleback fish that adjust pigmentation, an adaptation to specific freshwater lakes (Miller et al., 2007). This regulatory strategy is conserved in human populations, explaining some of the pigmentation difference between Africans and Europeans (Miller et al., 2007). A better understanding of the regulatory processes controlling gene expression would therefore help decipher how genotype becomes phenotype.

1.2.2 Control of transcription

Jacob, Monod and coworkers elucidated the first model of gene regulation, the *E. coli lac* operon (Jacob and Monod, 1961). In this context, the transcription of genes into messenger RNA may be repressed by the specific binding of a *trans*-acting factor to a *cis*-acting promoter region of DNA directly upstream ($5'$) of the gene (operon). Further work confirmed the generality of this model, in which transcription factors either activated or repressed bacterial genes by binding to *cis* promoters, and it was soon appreciated that the initiation of transcription was the most important step for regulation. Despite major differences between prokaryotic and eukaryotic cellular structure, DNA organization, polymerase machinery, and RNA processing, the model of control by transcription factors binding to specific promoter sequences has generalized surprisingly well to eukaryotes. Nonetheless, there are crucial differences in the regulation of transcription between these domains of life. Most importantly, the chromatin organization of eukaryotic DNA leads to a ground state of inactive transcription, whereas the ground state of prokaryotic promoters is activation (Struhl, 1999). Therefore, the role of transcription factors in bacteria is generally limited to either inhibiting the recruitment of RNA polymerase to the DNA or enhancing polymerase recruitment to compensate for a weak or incomplete promoter. In contrast, eukaryotic transcription factors can affect transcription by either interacting directly with polymerase or modifying chromatin to increase accessibility to the transcriptional machinery (Struhl, 1999; Lee and Young, 2000). Because of the complexity of eukaryotic gene initiation, transcription is an inherently combinatorial process, with many *trans*-factors and *cis*-elements combining to generate tightly regulated patterns of gene expression (Ravasi et al., 2010).

There are several classes of *trans*-acting factors affecting eukaryotic transcription: general transcription factors, including RNA polymerase II and associated proteins, involved in all transcription of protein-coding genes; promoter-specific transcription factors, which activate specific genes; and coactivators, which typically mediate interactions between specific transcription factors and general transcription machinery (Maston et al., 2006). The *cis*-acting elements that regulate transcription also are also quite varied: core promoters are the sites of the general transcription factor binding and define the transcriptional start site; proximal promoters flank the core promoter and typically are rich in transcription factor binding sites; enhancers are more distal elements that stimulate transcription; silencers are typically distal elements that repress transcription; and insulators block long-range effects on transcription (Maston et al., 2006). The combination of effects of all the *trans*-factors binding to all these *cis*-elements can be dynamic and cell-type specific. It is this combination that determines the rate of transcription initiation, in turn determining a large contribution to the rate of gene expression.

1.2.3 Sequence-specific transcription factor binding

Although there have been many important advances in the mechanism of transcription factor action (Struhl, 1999; Lee and Young, 2000; Maston et al., 2006), I focus here on the specificity of transcriptional control, which is for the most part determined by binding events. Transcription factors typically recognize degenerate sequences ranging from a few to tens of nucleotides, although there are generally 4-6 positions defining the binding specificity (Portales-Casamar et al., 2010). Because these motifs typically have low information content, sites predicted to bind based on sequence are found far more frequently throughout the genome than experimentally-determined binding sites (Maston et al., 2006). When factors do bind these motifs, sequence variants have a number of functional impacts: they can alter the binding affinity, select for different dimerization partners, or induce structural alterations in the protein that have a functional impact. For example, two κB binding sequences differing by a single nucleotide elicit differential responses not by affecting the ability of a particular κB dimer to bind, but instead by affecting which coactivators interact with the bound dimer (Leung et al., 2004). Gradients between strong and weak binding affinities can also specify spatial regions of gene expression. For example, in the developing

Drosophila embryo, the Dorsal morphogen activates enhancers with weak binding sites only in the ventral-most regions where Dorsal expression is highest, whereas enhancers with strong binding sites are activated throughout the neurogenic ectoderm (Jiang and Levine, 1993; Papatsenko and Levine, 2005). Gradients in binding affinities can also control the timing of gene activation in a developmental context. For example, the transcription factor PHA-4 specifies the identity of all *C. elegans* pharyngeal cells and increases in expression during development. Genes with weak PHA-4 binding sites in their promoters are activated later than genes with strong binding sites, creating a temporal gradient of activation (Gaudet and Mango, 2002). In a similar example, the Prep1 transcription factor controls the precise timing of *Drosophila* eye lens development via two low-affinity binding sites in the *Pax6* enhancer (Rowan et al., 2010). This evidence indicates that, rather than a single motif, individual transcription factors have a collection of binding sequences that have a range of properties and functional impacts.

1.2.4 Determining *in vivo* transcription factor specificity

The human genome encodes well over a thousand transcription factors, each of which is thought to bind to thousands of regions in the genome. As a result, it is challenging to determine experimentally the precise binding locations of transcription factors on a genomic scale. One method traditionally used to map regulatory elements takes advantage of the nucleosome-free state of bound regulatory regions by mapping DNase I hypersentivite regions. DNase I hypersensitivity has been applied to whole-genome analysis of binding sites and has proven useful for defining promoter regions, enhancers, and insulators (Crawford et al., 2004; Boyle et al., 2008). However, nucleosome-free regions are not specific for a particular transcription factor and do not have sufficient resolution to demarcate individual binding sites. A more specific method, called chromatin immunoprecipitation (ChIP), involves reversible crosslinking of DNA to protein followed by the selection of a particular transcription factor by a specific antibody (Ren et al., 2000). The bound regions can then be identified by polymerase chain reaction (PCR), by microarray chips (ChIP-Chip) (Ren et al., 2000) or by sequencing (ChIP-Seq) (Johnson et al., 2007). Reproducible and specific, these techniques have been used to elucidate many thousands of transcription factor binding regions (Harbison et al., 2004; ENCODE Project Consortium, 2007). Although the resolution of ChIP-Seq far exceeds ChIP-Chip (tens of nucleotides rather than hundreds), the exact binding location cannot be determined using either method. Typically, one follows a ChIP-Seq experiment with computational motif finding to help identify precise binding specificity and locations, although the variability and low information content of most binding motifs often make this challenging (Park, 2009). Additionally, there are inherent biases in the chromatin immunoprecipitation due to chromatin state, DNA fragmentation, antibody specificity, or biases in recognizing different conformational states or bound co-factors (Park, 2009). Therefore, it is difficult to attribute a quantitative affinity for individual sites.

1.2.5 Computational and *in vitro* approaches to specificity

Because it is difficult to experimentally determine the precise binding locations and sequence preferences of transcription factors *in vivo*, it is desirable to know the *in vitro* sequence preferences of transcription factors. Knowledge of sequence preferences helps to distinguish direct from indirect binding, often a problem in ChIP experiments (Gordân et al., 2009) and to separate context effects, such as chromatin structure, from binding affinity (Wasson and Hartemink, 2009). If one knows not only the sequence preferences but also the quantitative biophysical affinities, one can also constrain models of transcriptional systems (Endy and Brent, 2001). Finally, knowledge of *in vitro* affinities allows generalization of results to new biological systems with different genomes, DNA mutations, or changes in the expression of the transcription factor or any other involved protein. For all these reasons, many experimental and computational techniques have been developed for assaying transcription factor specificity *in vitro*.

A wide variety of approaches have been developed for finding *de novo* motifs that are enriched in a set of sequences, using different definitions for motifs and different models for statistical overrepresentation (Tompa et al., 2005). Typically these methods are designed to be run on promoters of coexpressed genes or on regions found by high-throughput ChIP experiments. These methods have been successfully applied in some contexts, especially in yeast (e.g. Segal et al. (2003)). However, in mammals, faced with the larger intergenic regions, increased combinatorial complexity, and larger role of chromatin effects, the sensitivity and accuracy of these methods is greatly reduced (Tompa et al., 2005). A complementary computational method is to identify regulatory motifs by their conservation across species. For example, Xie et al. (2005) identified dozens of known transcription factor motifs and over 100 new motifs by comparing human promoters to orthologous regions in other mammals. Although successful at finding consensus sequences, this approach cannot determine which factor binds which motif and cannot distinguish the relative binding strength of different sequences.

Although they were first developed decades ago, experimental assays for *in vitro* transcription factor specificity are currently undergoing accelerating progress. For example, systematic evolution of ligands by exponential enrichment (SELEX) is a general technique for selecting nucleic acids with a particular binding affinity and has been successfully applied for many transcription factors (Klug and Famulok, 1994). SELEX has recently been coupled with high-throughput sequencing to determine transcription factor binding preferences (Zhao et al., 2009; Jolma et al., 2010). Amazingly, these approaches can examine binding of transcription factors to millions of DNA sequences. Unfortunately, when multiple rounds of selection are used, SELEX provides high specificity but loses its power to quantitatively determine the affinity of weak sites. In contrast, a single round of selection leaves a high level of noise, necessitating large amounts of data and assumptions about the binding mode to develop a biophysical model of interaction (Zhao et al., 2009). Regardless, SELEX provides an excellent approach for defining a consensus binding site.

An alternative method for defining the comprehensive specificity of transcription factors is to use a protein binding microarray (PBM) (Mukherjee et al., 2004). A PBM enumerates all sequences of 8-10 nucleotides; fluorescently labeled transcription factors are bound directly to these sequences on the microarray, and visualized *in* *situ* (Berger et al., 2006). This has the advantage of directly measuring the relative affinity of a protein to each sequence, so that preferences can be determined without any assumptions about the biophysical nature of the interaction, such as the independence of motif positions. One major limitation of the PBM approach is that less than 50,000 microarray features can be used, limiting the length of sequences (k-mers) that can be completely enumerated. One can extend the length of the DNA features to compensate, but this creates a tradeoff by complicating analysis and interpretation (Berger and Bulyk, 2009). Although the measured affinities are quantitative, a single microarray only measures binding at one concentration, requiring the experimenter to sacrifice the weakest binding motifs if she wishes to differentiate between the relatively strong motifs. Stringent washes are also required for PBMs, potentially removing proteins from weak binding motifs. For more quantitative determination of thermodynamic parameters, microfluidic devices have been used to trap fluorescently-labeled protein bound to DNA at equilibrium. By adjusting the input concentration, one can thus measure thermodynamic binding constants for hundreds of protein-DNA pairs (Maerkl and Quake, 2007). Although this platform cannot handle thousands of sequences at once, it can directly measure equilibrium binding *in situ* without the need for a wash step, and is thus highly quantitative. A wide variety of methods have proven useful for determining the binding specificity of transcription factors, but there is still no method that is comprehensive, extremely high-throughput, and highly quantitative.

1.3 Regulation by miRNAs

1.3.1 miRNA genes

The first microRNA (miRNA), *lin-4*, was discovered in a screen for regulators of developmental timing in *Caenorhabditis elegans* (Lee et al., 1993). Remarkably, the Ambros and Ruvkun labs postulated several of the key characteristics of miRNAs from this single example: the *lin-4* locus encoded for functional RNA, not a protein;

the locus encoded a larger RNA, now called a precursor or pre-miRNA, processed into a roughly 22 nucleotide RNA, now called the mature miRNA; and the *lin-4* RNA downregulated a target gene, *lin-14*, by pairing to the 3′ UTR of its mRNA (Lee et al., 1993; Wightman et al., 1993). However, it was nearly seven years until a second miRNA was discovered in another screen for developmental regulators in *C. elegans* (Reinhart et al., 2000). Subsequently, it was shown that this new miRNA, *let-7*, was conserved throughout most metazoans (Pasquinelli et al., 2000). Thus it was recognized that rather than being a unique mechanism for regulating a single *C. elegans* gene, miRNAs are a class of gene regulators that is active in most metazoans, including humans. Several efforts to clone small RNAs discovered dozens of abundant miRNAs in *C. elegans*, *D. melanogaster*, and vertebrates (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). It soon became clear that miRNAs are a major gene family in virtually all animals (Bartel, 2004).

1.3.2 miRNA biogenesis

Just prior to the discovery of the prominence of miRNAs, pioneering work in *C. elegans* led to another fundamental discovery: RNA interference, or RNAi (Fire et al., 1998). Fire and colleagues observed that exogenous, long double-stranded RNA could specifically and catalytically repress the expression of genes with matching sequence. Surprisingly, a *Drosophila in vitro* system revealed that the double-stranded RNA of RNAi was processed to form 21 to 23 nucleotide RNAs, later called small interfering RNAs (siRNAs), that were the same length as mature miRNAs (Zamore et al., 2000). It was subsequently shown that an RNase III enzyme, Dicer, processes both dsRNAs into siRNAs (Bernstein et al., 2001) and pre-miRNAs like the long form of *lin-4* into mature miRNAs (Grishok et al., 2001; Hutvágner et al., 2001; Ketting et al., 2001). Combined with evidence that both functional siRNAs and miRNAs reside in the RNA-induced silencing complex (RISC) (Martinez et al., 2002), this suggested that siRNAs and miRNAs might join the same functional pathway. Indeed, it was later shown that siRNAs can have non-enzymatic, miRNA-like target repression (Doench et al., 2003), and that miRNAs can cleave target mRNAs (Hutvágner and Zamore, 2002; Yekta et al., 2004). Both siRNAs and miRNAs have functions mediated by Argonaute proteins, which are the core component of the RISC. When incorporated into the RISC with the same core Argonaute protein, all evidence points to the idea siRNAs and miRNAs are functionally equivalent. The RNAi pathway and metazoan miRNA pathway differ mainly in the prominence of highly complementary targets (discussed later), and their earlier steps in biogenesis. While dsRNAs are processed directly into siRNAs by Dicer (Bernstein et al., 2001), most miRNAs are first transcribed by RNA polymerase II as capped, polyadenylated transcripts (Cai et al., 2004). These primary miRNA transcripts, or pri-miRNAs, are then processed by an RNase III enzyme called Drosha and exported from the nucleus as a pre-miRNAs (Lee et al., 2003). In the cytoplasm, Dicer finally processes the pre-miRNA to create the mature miRNA.

1.3.3 Principles of miRNA targeting

The first miRNA targets, discovered genetically, had sequences in their 3′ UTRs with partial complementarity to their respective miRNA regulators (Lee et al., 1993; Wightman et al., 1993; Moss et al., 1997; Reinhart et al., 2000). Typically, siRNAs strongly repress RNAs with extensive complementarity by cleaving the target strand between the portion paired with nucleotides 10 and 11 of the siRNA. In contrast, animal miRNAs do not typically have enough complementarity to their targets to mediate cleavage (Elbashir et al., 2001; Hutvágner and Zamore, 2002). Without extensive complementarity between miRNAs and their targets, it was unclear what the rules of miRNA specificity might be and how to predict further targets. Lacking the strong effects of enzymatic cleavage of messages, miRNAs repress their targets by translational repression and destabilizing the message, often totaling less than 20% repression at the protein level (Baek et al., 2008; Selbach et al., 2008; Bartel, 2009). Because of the noise inherent in experimental assays for gene expression, it is difficult to evaluate such small effects for individual targets. Such experimental noise can, however, be averaged out when aggregating a group of potential targets to study principles of targeting. Because of the difficulty of experimentally verifying individual miRNA targets, many turned to computational methods to determine the rules of miRNA specificity and to predict new targets.

A first hint of target specificity was found when it was noticed that the $5'$ ends of some *Drosohpila* miRNAs were perfectly complementary to motifs known as the K box, Brd box, and GY box, which had previously been shown to confer post-transcriptional repression (Lai, 2002). Subsequently, several groups attempted to leverage what was known about miRNA targets to computationally predict new targets (Stark et al., 2003; Enright et al., 2003; Lewis et al., 2003; Rajewsky and Socci, 2004; John et al., 2004). This first generation of methods based predictions on some combination of complementarity often weighted towards the $5'$ end of the miRNA, free energy of pairing to the entire miRNA, and evolutionary conservation of the target. These sets of predictions had little overlap and two major weaknesses of these methods were soon apparent: they used an extremely small number of experimentally-determined targets as training and/or test sets and were based on assumptions with little or no experimental support. For example, the free-energy-of-pairing calculation assumes two free RNA molecules in solution, but the miRNA is tightly and stably incorporated into an Argonaute protein, which constrains the possible structures that could be formed (Rajewsky, 2006; Bartel, 2009). Although most first-generation prediction programs were validated using a limited set of experimentally-determined targets, Lewis et al. (2003) used conservation as an elegant and independent tool for determining the specificity of miRNA targeting. Reasoning that bona fide miRNA target sequences that are vital for the survival of an organism would be conserved over evolutionary time, they tested potential targeting rules by comparing the conservation of miRNA complementary sequences to those complementary to shuffled miRNAs. Using the resulting "signal to noise ratio", or ratio of conserved sequences complementary to miRNAs to conserved sequences complementary to shuffled controls, the authors found that matches to the $5'$ end of miRNAs were conserved at a much higher rate than matches to the $3'$ end. They defined nucleotides 2 through 8 of the miRNA as the "seed" sequence. They next predicted targets by selecting sequences with conserved perfect complementarity to the seed sequence (seed matches), and passing a free energy criterion for predicted pairing to the rest of the miRNA.

Although the evidence for targeting by seed matches was initially based only on conservation, experimental evidence quickly confirmed their utility (Doench and Sharp, 2004). A striking analysis of *in vivo* silencing of fluorescent protein in transgenic *Drosophila melanogaster* showed that seed matches confer repression on targets without strong pairing to the 3′ end of the miRNA (Brennecke et al., 2005). However, extensive pairing to the 3′ end of the miRNA could supplement a weak seed match to enhance repression or could compensate for imperfect seed matches, such as those with G:U wobble pairs. Comparing the number of conserved sequences complementary to miRNA seed regions to those complementary shuffled miRNAs revealed that the vast majority of miRNA targets lacked substantial 3′ pairing in *Drosophila* (Brennecke et al., 2005). The next generation of mammalian target predictions redefined the seed as nucleotides 2 through 7 of the miRNA and found that in this system as well, dropping any free energy criterion (leaving only the seed matches) provided better sensitivity without sacrificing specificity (Lewis et al., 2005). This led to the following simple strategy for predicting miRNA targets: start with perfect seed matches in 3′ UTRs and filter them for perfect conservation between several species, e.g. human, mouse, rat, and dog, yielding a set of predictions that are enriched for true targets. The second generation of target predictions was based on these principles but had some variations: Lewis et al. (2005) found evidence for the conservation of an adenosine opposite position 1 of the miRNA and required perfect seed matches, whereas other methods counted Watson-Crick matches at position 1 and searched for 3′ compensatory sites as well (Krek et al., 2005; Brennecke et al., 2005). The degree of overlap for several current target prediction programs is quite high because all of them now require stringent seed pairing (Bartel, 2009). However, there are also minor differences due to the use of different alignments, UTR annotations, and miRNA annotations. Chapter 2 describes the development of a much more sensitive and accurate set of unique features for a conservation-based target prediction method that are currently not implemented elsewhere.

1.3.4 Evolutionary impact of miRNAs

Several experimental approaches have provided genome-scale confirmation of the importance of miRNA seed matches. Transfections of miRNA mimics or siRNAs into human cell lines followed by microarray analysis repress hundreds of messages containing seed matches but more extended seed pairing has little additional effect (Lim et al., 2005; Birmingham et al., 2006; Jackson et al., 2006; Grimson et al., 2007; Nielsen et al., 2007). Mass spectrometry approaches have confirmed this result at the protein level (Baek et al., 2008; Selbach et al., 2008). Many of the seed matches conferring repression in these cases are conserved, but the majority are not, suggesting that nonconserved miRNA repression is widespread. If miRNAs impact general expression in the cells in which they are present, then abundant mRNAs must avoid complementarity to co-expressed miRNAs in order to maintain their high expression. These hypothetical mRNAs avoiding miRNA targeting were termed "anti-targets" (Bartel and Chen, 2004). Two remarkable studies showed that anti-targets were common phenomena: Stark et al. (2005) used *in situ* hybridization evidence to show that miRNAs were expressed in exclusive spatial patterns from their targets, and Farh et al. (2005) used microarray data to show that miRNA seed matches were depleted in genes that were highly and specifically coexpressed with miRNAs. Ubiquitously expressed (housekeeping) genes have short 3′ UTRs and also strongly avoid miRNA seed matches at the sequence level (Stark et al., 2005). Additionally, Farh et al. (2005) showed that non-conserved miRNA seed matches often conferred strong repression in luciferase reporter assays, signifying that genes coexpressed with highly expressed miRNAs could be repressed if they did not avoid seed matches in their 3′ UTRs. In one striking example of widespread repression *in vivo*, Giraldez et al. (2006) observed that miR-430, crucial for the zebrafish maternal to zygotic transition, targeted hundreds of maternal mRNAs with nonconserved seed matches including many weak 6mer seed matches. Taken together, these results show that widespread miRNA targeting impacts the evolution of metazoan 3′ UTRs and genes in a profound way even in the absence of extended pairing or strong conservation.

1.3.5 Beyond conserved miRNA targeting

Reporter assays have suggested that seed matches, whether conserved or not, were a necessary but not sufficient condition for seed match repression in mammals (Farh et al., 2005; Grimson et al., 2007). Especially in light of the large number of 6-8 nucleotide seed matches in mammalian 3′ UTRs, there must therefore be other sequence elements beyond the mere presence or absence of a seed match that determine the extent of repression for miRNA targets. Several of these context features have been found by the analysis of global expression datasets. The first feature is the type of the seed match itself, which has a large effect on target repression. A match to positions 2-7 of a miRNA (a 6mer seed match) typically has only a small effect on messages unless flanked by a Watson-Crick match opposite position 8 (a 7mer-M8), an adenosine opposite position 1 (a 7mer-A1), or both (an 8mer) (Grimson et al., 2007; Nielsen et al., 2007; Baek et al., 2008; Selbach et al., 2008). Another important feature is positioning at least 15 nucleotides after the stop codon. Seed matches in 5′ UTRs and ORFs are much less effective than those in 3′ UTRs, likely because the scanning or translating ribosome interferes with RISC binding (Grimson et al., 2007). In fact, the "ribosome shadow" of 15 nucleotides past the stop codon is also subject to steric interference by the ribosome, decreasing the efficacy of miRNA target repression in this region (Grimson et al., 2007). AU-rich composition of the sequence surrounding a seed match also improves target efficacy by decreasing secondary structure and increasing accessibility of the site (Grimson et al., 2007; Nielsen et al., 2007; Kertesz et al., 2007). Finally, seed matches located between 8 and 40 nucleotides of another seed match tend to act cooperatively, providing a potent increase in efficacy (Grimson et al., 2007).

Combining these effects into a target prediction framework enables accurate prediction of efficacy (Grimson et al., 2007; Baek et al., 2008). Therefore these features provide not only insight into miRNA targeting mechanism but also a useful complement to predictions based on conservation, since the efficacy of repression conferred by a potential miRNA target is not perfectly correlated with conservation. Also, many bona fide miRNA targets presumably represent species-specific adaptations that are by definition not conserved. Filtering seed matches for conservation provides a greatly-reduced set of target predictions that are enriched for targets that are subject to purifying selection, so target predictions based on context features alone will have much higher sensitivity. However, conservation is still an extremely useful tool for target prediction. Repression by miRNAs in a heterologous assay does not imply endogenous targeting or function. A message with a seed match might never be co-expressed with a miRNA under physiological conditions, precluding any chance of interaction. Or, some repression of a gene might simply be noise if the relevant biological process is subject to canalization, which could provides robustness to small changes in expression. In contrast, given an appropriate null model and statistical framework, a significant signal for evolutionary conservation likely corresponds to function that is important for the fitness of the organism in some way, sidestepping problems of noisy gene expression and prioritizing vital interactions for further study.

1.3.6 miRNA functions *in vivo*

While tremendous progress has been made in predicting target genes for miRNAs, it has been much less clear how these targets translate into biological function. The classical miRNAs, *lin-14* and *let-7*, acted as switches to clearly delineate *C. elegans* developmental state by repressing key targets (Reinhart et al., 2000). Since these discoveries, miRNAs have been implicated in virtually every metazoan biological process, including development, tissue definition, genetic disease, immune function, and countless others. However, it has been difficult to find individual switch targets that mediate many of these functions. It was hypothesized early on that miRNAs could affect targets in other ways, for example by "micromanaging" gene expression as rheostats for individual genes (Bartel and Chen, 2004). This hypothesis seemed increasingly likely as evidence mounted for the widespread impact of both conserved and non-conserved targets in animals. One example of this type is miR-8 targeting the *Drosophila* transcription factor *atrophin*. miR-8-mediated repression of *atrophin* expression is necessary to prevent apoptosis in the brain and behavioral defects, but excessive *atrophin* repression leads to the deleterious formation of extra wing veins (Karres et al., 2007). Thus, miR-8 fine-tunes *atrophin* expression to a beneficial level in the cells in which they are coexpressed. In addition to tuning targets, it is likely many other modes of miRNA/target interaction arise from their relative expression patterns. For example, (Shkumatava et al., 2009) used fluorescent-activated cell sorting of cells from zebrafish embryos to assay the expression patterns of miRNAs and their targets. The authors found that for most targets, miRNAs acted in concert with other regulatory mechanisms to reinforce coherent patterns of expression, but that some targets were switch-like and others were preferentially co-expressed with their cognate miRNAs. Clearly, the *in vivo* roles of miRNAs are complex and a rich source for further study.

1.4 References

D. Baek, J. Villén, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, Sep 2008. doi: 10.1038/nature07242.

D. P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116 (2):281–97, Jan 2004.

D. P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2): 215–33, Jan 2009. doi: 10.1016/j.cell.2009.01.002.

D. P. Bartel and C.-Z. Chen. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nat Rev Genet*, 5(5):396–400, May 2004. doi: 10.1038/nrg1328.

M. F. Berger and M. L. Bulyk. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc*, 4(3):393–411, Jan 2009. doi: 10.1038/nprot.2008.195.

M. F. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep, and M. L. Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11):1429–35, Nov 2006. doi: 10.1038/nbt1246.

E. Bernstein, A. A. Caudy, S. M. Hammond, and G. J. Hannon. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818):363–6, Jan 2001. doi: 10.1038/35053110.

A. Birmingham, E. M. Anderson, A. Reynolds, D. Ilsley-Tyree, D. Leake, Y. Fedorov, S. Baskerville, E. Maksimova, K. Robinson, J. Karpilow, W. S. Marshall, and A. Khvorova. 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nat Methods*, 3(3):199–204, Mar 2006. doi: 10.1038/nmeth854.

K. Blyth, E. R. Cameron, and J. C. Neil. The RUNX genes: gain or loss of function in cancer. *Nat Rev Cancer*, 5(5):376–87, May 2005. doi: 10.1038/nrc1607.

A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–22, Jan 2008. doi: 10.1016/j.cell.2007.12.014.

J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen. Principles of microRNA-target recognition. *PLoS Biol*, 3(3):e85, Mar 2005. doi: 10.1371/journal.pbio.0030085.

X. Cai, C. H. Hagedorn, and B. R. Cullen. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA (New York, NY)*, 10(12):1957–66, Dec 2004. doi: 10.1261/rna.7135204.

G. E. Crawford, I. E. Holt, J. C. Mullikin, D. Tai, R. Blakesley, G. Bouffard, A. Young, C. Masiello, E. D. Green, T. G. Wolfsberg, F. S. Collins, and N. I. O. H. I. S. Center. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proceedings of the National Academy of Sciences of the United States of America*, 101(4):992–7, Jan 2004. doi: 10.1073/pnas.0307540100.

J. G. Doench and P. A. Sharp. Specificity of microRNA target selection in translational repression. *Genes Dev*, 18(5):504–11, Mar 2004. doi: 10.1101/gad.1184404.

J. G. Doench, C. P. Petersen, and P. A. Sharp. siRNAs can function as miRNAs. *Genes Dev*, 17(4):438–42, Feb 2003. doi: 10.1101/gad.1064703.

S. M. Elbashir, W. Lendeckel, and T. Tuschl. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev*, 15(2):188–200, Jan 2001.

W. Enard, P. Khaitovich, J. Klose, S. Zöllner, F. Heissig, P. Giavalisco, K. Nieselt-Struwe, E. Muchmore, A. Varki, R. Ravid, G. M. Doxiadis, R. E. Bontrop, and S. Pääbo. Intra- and interspecific variation in primate gene expression patterns. *Science*, 296(5566):340–3, Apr 2002. doi: 10.1126/science.1068996.

ENCODE Project Consortium. Identification and analysis of functional elements in 1human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, Jun 2007. doi: 10.1038/nature05874.

D. Endy and R. Brent. Modelling cellular behaviour. *Nature*, 409(6818):391–5, Jan 2001. doi: 10.1038/35053181.

A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks. MicroRNA targets in Drosophila. *Genome Biol*, 5(1):R1, Jan 2003. doi: 10.1186/gb-2003-5-1-r1.

K. K.-H. Farh, A. Grimson, C. Jan, B. P. Lewis, W. K. Johnston, L. P. Lim, C. B. Burge, and D. P. Bartel. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, 310(5755):1817–21, Dec 2005. doi: 10.1126/science.1121158.

A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–11, Feb 1998. doi: 10.1038/35888.

J. Gaudet and S. E. Mango. Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4. *Science*, 295(5556):821–5, Feb 2002. doi: 10.1126/science.1065175.

Y. Gilad, A. Oshlack, and S. A. Rifkin. Natural selection on gene expression. *Trends Genet*, 22(8):456–61, Aug 2006. doi: 10.1016/j.tig.2006.06.002.

A. J. Giraldez, Y. Mishima, J. Rihel, R. J. Grocock, S. V. Dongen, K. Inoue, A. J. Enright, and A. F. Schier. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science*, 312(5770):75–9, Apr 2006. doi: 10.1126/science.1122689.

R. Gordân, A. Hartemink, and M. Bulyk. Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res*, Sep 2009. doi: 10.1101/gr.094144.109.

A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell*, 27(1):91–105, Jul 2007. doi: 10.1016/j.molcel.2007.06.017.

A. Grishok, A. E. Pasquinelli, D. Conte, N. Li, S. Parrish, I. Ha, D. L. Baillie, A. Fire, G. Ruvkun, and C. C. Mello. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, 106(1):23–34, Jul 2001.

C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. MacIsaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, Sep 2004. doi: 10.1038/nature02800.

G. Hutvágner and P. D. Zamore. A microRNA in a multiple-turnover RNAi enzyme complex. *Science*, 297(5589):2056–60, Sep 2002. doi: 10.1126/science.1073827.

G. Hutvágner, J. McLachlan, A. E. Pasquinelli, E. Bálint, T. Tuschl, and P. D. Zamore. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, 293(5531):834–8, Aug 2001. doi: 10.1126/science.1062961.

A. L. Jackson, J. Burchard, J. Schelter, B. N. Chau, M. Cleary, L. Lim, and P. S. Linsley. Widespread siRNA ”off-target” transcript silencing mediated by seed region sequence complementarity. *RNA (New York, NY)*, 12(7):1179–87, Jul 2006. doi: 10.1261/rna.25706.

F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318–56, Jun 1961.

J. Jiang and M. Levine. Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. *Cell*, 72 (5):741–52, Mar 1993.

B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks. Human MicroRNA targets. *PLoS Biol*, 2(11):e363, Nov 2004. doi: 10.1371/journal.pbio.0020363.

D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–502, Jun 2007. doi: 10.1126/science.1141319.

A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanpää, M. Bonke, K. Palin, S. Talukder, T. R. Hughes, N. M. Luscombe, E. Ukkonen, and J. Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome research*, Apr 2010. doi: 10.1101/gr.100552.109.

J. S. Karres, V. Hilgers, I. Carrera, J. Treisman, and S. M. Cohen. The conserved microRNA miR-8 tunes atrophin levels to prevent neurodegeneration in Drosophila. *Cell*, 131(1):136–45, Oct 2007. doi: 10.1016/j.cell.2007.09.020.

M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal. The role of site accessibility in microRNA target recognition. *Nat Genet*, 39(10):1278–84, Oct 2007. doi: 10.1038/ng2135.

R. F. Ketting, S. E. Fischer, E. Bernstein, T. Sijen, G. J. Hannon, and R. H. Plasterk. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev*, 15(20):2654–9, Oct 2001. doi: 10.1101/gad.927801.

M. C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–16, Apr 1975.

S. J. Klug and M. Famulok. All you wanted to know about SELEX. *Mol Biol Rep*, 20(2):97–107, Jan 1994.

A. Krek, D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky. Combinatorial microRNA target predictions. *Nat Genet*, 37(5):495–500, May 2005. doi: 10.1038/ng1536.

M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 294(5543):853–8, Oct 2001. doi: 10.1126/science.1064921.

E. C. Lai. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet*, 30(4):363–4, Apr 2002. doi: 10.1038/ng865.

N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294 (5543):858–62, Oct 2001. doi: 10.1126/science.1065062.

R. C. Lee and V. Ambros. An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294(5543):862–4, Oct 2001. doi: 10.1126/science.1065329.

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

T. I. Lee and R. A. Young. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet*, 34:77–137, Jan 2000. doi: 10.1146/annurev.genet.34.1.77.

Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Rådmark, S. Kim, and V. N. Kim. The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425(6956):415–9, Sep 2003. doi: 10.1038/nature01957.

T. H. Leung, A. Hoffmann, and D. Baltimore. One nucleotide in a kappaB site can determine cofactor specificity for NF-kappaB dimers. *Cell*, 118(4):453–64, Aug 2004. doi: 10.1016/j.cell.2004.08.007.

B. P. Lewis, I. hung Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–98, Dec 2003.

B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, Jan 2005. doi: 10.1016/j.cell.2004.12.035.

L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027): 769–73, Feb 2005. doi: 10.1038/nature03315.

S. J. Maerkl and S. R. Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–7, Jan 2007. doi: 10.1126/science.1131007.

J. Martinez, A. Patkaniowska, H. Urlaub, R. Lührmann, and T. Tuschl. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell*, 110(5): 563–74, Sep 2002.

G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, 7:29–59, Jan 2006. doi: 10.1146/annurev.genom.7.080505.115623.

C. T. Miller, S. Beleza, A. A. Pollen, D. Schluter, R. A. Kittles, M. D. Shriver, and D. M. Kingsley. cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell*, 131(6):1179–89, Dec 2007. doi: 10.1016/j.cell.2007.10.055.

E. G. Moss, R. C. Lee, and V. Ambros. The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. *Cell*, 88(5):637–46, Mar 1997.

S. Mukherjee, M. F. Berger, G. Jona, X. S. Wang, D. Muzzey, M. Snyder, R. A. Young, and M. L. Bulyk. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet*, 36(12):1331–9, Dec 2004. doi: 10.1038/ng1473.

C. B. Nielsen, N. Shomron, R. Sandberg, E. Hornstein, J. Kitzman, and C. B. Burge. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*, 13(11):1894–910, Nov 2007. doi: 10.1261/rna.768207.

D. Papatsenko and M. Levine. Quantitative analysis of binding motifs mediating diverse spatial readouts of the Dorsal gradient in the Drosophila embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 102(14): 4966–71, Apr 2005. doi: 10.1073/pnas.0409414102.

P. J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–80, Oct 2009. doi: 10.1038/nrg2641.

A. E. Pasquinelli, B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Müller, J. Spring, A. Srinivasan, M. Fishman, J. Finnerty, J. Corbo, M. Levine, P. Leahy, E. Davidson, and G. Ruvkun. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–9, Nov 2000. doi: 10.1038/35040556.

E. Portales-Casamar, S. Thongjuea, A. T. Kwon, D. Arenillas, X. Zhao, E. Valen, D. Yusuf, B. Lenhard, W. W. Wasserman, and A. Sandelin. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic acids research*, 38(Database issue):D105–10, Jan 2010. doi: 10.1093/nar/gkp950.

N. Rajewsky. microRNA target predictions in animals. *Nat Genet*, 38 Suppl:S8–13, Jun 2006. doi: 10.1038/ng1798.

N. Rajewsky and N. D. Socci. Computational identification of microRNA targets. *Dev Biol*, 267(2):529–35, Mar 2004. doi: 10.1016/j.ydbio.2003.12.003.

T. Ravasi, H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, P. Carninci, C. O. Daub, A. R. R. Forrest, J. Gough, S. Grimmond, J.-H. Han, T. Hashimoto, W. Hide, O. Hofmann, A. Kamburov, M. Kaur, H. Kawaji, A. Kubosaki, T. Lassmann, E. van Nimwegen, C. R. MacPherson, C. Ogawa, A. Radovanovic, A. Schwartz, R. D. Teasdale, J. Tegnér, B. Lenhard, S. A. Teichmann, T. Arakawa, N. Ninomiya, K. Murakami, M. Tagami, S. Fukuda, K. Imamura, C. Kai, R. Ishihara, Y. Kitazume, J. Kawai, D. A. Hume, T. Ideker, and Y. Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–52, Mar 2010. doi: 10.1016/j.cell.2010.01.044.

B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772):901–6, Feb 2000. doi: 10.1038/35002607.

B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290 (5500):2306–9, Dec 2000. doi: 10.1126/science.290.5500.2306.

S. Rowan, T. Siggers, S. A. Lachke, Y. Yue, M. L. Bulyk, and R. L. Maas. Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity. *Genes Dev*, 24(10):980–5, May 2010. doi: 10.1101/gad.1890410.

E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, 19 Suppl 1:i273–82, Jan 2003.

M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455 (7209):58–63, Sep 2008. doi: 10.1038/nature07228.

A. Shkumatava, A. Stark, H. Sive, and D. P. Bartel. Coherent but overlapping expression of microRNAs and their targets during vertebrate development. *Genes Dev*, 23(4):466–81, Feb 2009. doi: 10.1101/gad.1745709.

A. Stark, J. Brennecke, R. B. Russell, and S. M. Cohen. Identification of Drosophila MicroRNA targets. *PLoS Biol*, 1(3):E60, Dec 2003. doi: 10.1371/journal.pbio.0000060.

A. Stark, J. Brennecke, N. Bushati, R. B. Russell, and S. M. Cohen. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123(6):1133–46, Dec 2005. doi: 10.1016/j.cell.2005.11.023.

K. Struhl. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, 98(1):1–4, Jul 1999. doi: 10.1016/S0092-8674(00)80599-1.

M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. D. Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Régnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–44, Jan 2005. doi: 10.1038/nbt1053.

T. Wasson and A. J. Hartemink. An ensemble model of competitive multi-factor binding of the genome. *Genome Res*, 19(11):2101–12, Nov 2009. doi: 10.1101/gr.093450.109.

B. Wightman, I. Ha, and G. Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5):855–62, Dec 1993.

X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–45, Mar 2005. doi: 10.1038/nature03441.

S. Yekta, I. hung Shih, and D. P. Bartel. MicroRNA-directed cleavage of HOXB8 mRNA. *Science*, 304(5670):594–6, Apr 2004. doi: 10.1126/science.1097434.

P. Zamore, T. Tuschl, P. Sharp, and D. Bartel. RNAi: Double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell*, Jan 2000.

Y. Zhao, D. Granas, and G. D. Stormo. Inferring binding energies from selected binding sites. *PLoS Comput Biol*, 5(12):e1000590, Dec 2009. doi: 10.1371/journal.pcbi.1000590.

Chapter 2

Most mammalian mRNAs are conserved targets of microRNAs

Robin C. Friedman, Kyle Kai-How Farh, Christopher B. Burge, and David P. Bartel

This chapter is presented in the context of its contemporary science and originally appeared in *Genome Research 19: 92-105 (2009).*

K.K.F. performed many of these analyses in preliminary form, R.C.F., C.B.B. and D.P.B. designed the algorithms, R.C.F. performed all analysis, and R.C.F., C.B.B. and D.P.B. wrote the paper.

Corresponding Supplementary Material can be found in Appendix 1.

Chapter 2

Most mammalian mRNAs are conserved targets of microRNAs

2.1 Introduction

MicroRNAs are 22-nucleotide (nt) endogenous RNAs that derive from distinctive hairpin precursors in plants and animals (Bartel, 2004). After incorporation into a silencing complex, which contains at its core an Argonaute protein, an miRNA can pair to an mRNA and thereby specify the post-transcriptional repression of that protein-coding message, either by transcript destabilization, translational repression, or both. MicroRNAs constitute one of the more abundant classes of gene-regulatory molecules in animals, with hundreds of distinct miRNAs confidently identified in both human and mouse (Landgraf et al., 2007). A central goal for understanding the functions of all these small regulatory RNAs has been to determine which messages are targeted for repression.

The search for biological targets of metazoan miRNAs has benefited greatly from the comparative analysis of orthologous mRNAs. Targets of miRNAs can be predicted above the background of false-positive predictions by requiring conserved Watson-Crick pairing to the 5′ region of the miRNA, known as the miRNA seed (Lewis et al., 2003). Because so many messages have preferentially preserved their pairing to miRNA seeds, targets can be predicted simply by searching for conserved 6-8mer matches to miRNA seed region (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005). Four types of seed-matched sites are known to be selectively conserved (Lewis et al., 2005): the 6mer site, which perfectly matches the 6-nt miRNA seed, the 7mer-m8 site, which comprises the seed match supplemented by a Watson-Crick match to miRNA nucleotide 8, the 7mer-A1 site, which comprises the seed match supplemented by an A across from miRNA nucleotide 1, and the 8mer site, which comprises the seed match supplemented by both the m8 and the A1 (Fig. 2-1A). Supporting the validity of seed-matched target predictions, cellular messages that either decrease following miRNA addition or increase following miRNA disruption preferentially contain seed matches (Lim et al., 2005; Giraldez et al., 2006; Rodriguez et al., 2007), with the following hierarchy of site efficacy: 8mer > 7mer-m8 > 7mer-A1 > 6mer (Grimson et al., 2007; Nielsen et al., 2007). The same is true when examining protein levels (Baek et al., 2008; Selbach et al., 2008).

In addition to its utility for predicting the identities of the regulatory targets, comparative sequence analysis has provided fundamental insights regarding features of mRNA sites required for effective miRNA recognition. For example, a systematic analysis of matches to 7-nt segments spanning the length of the miRNAs showed that only those matching the 5′ region of the miRNA are conserved more than expected by chance, thereby defining the seed region as the key determinant of miRNA specificity (Lewis et al., 2003). Additional analyses of preferential conservation uncovered the importance of non-Watson-Crick recognition of an A across from miRNA nucleotide 1 and of an A or U across from nucleotide 9 (Lewis et al., 2005; Nielsen et al., 2007). Comparative analyses revealed targeting in open reading frames (ORFs) (Lewis et al., 2005; Stark et al., 2007a) but also supported experimental findings that sites are more effective if they fall outside the path of the ribosome (Grimson et al., 2007). Comparative analyses supported the importance of other features of site context, including positioning within high local AU composition (Grimson et al., 2007; Nielsen et al., 2007), away from the centers of long UTRs (Gaidatzis et al., 2007; Grimson et al., 2007; Majoros and Ohler, 2007), and near to nucleotides that can pair to miRNA nucleotides 13-16 (Grimson et al., 2007).

Comparative analysis has also revealed the wide scope of metazoan miRNA targeting, indicating that many genes of mammals, flies, and worms are miRNA targets (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005; Xie et al., 2005; Lall et al., 2006). For example, combining the 3′ UTR conservation attributed to miRNA seed matching with that from ORFs indicates that over one third of human protein-coding genes have been under selective pressure to maintain pairing to miRNAs (Lewis et al., 2005). Moreover, the selective depletion of seed-matching sites in messages highly expressed in the same tissues as the miRNAs implies frequent nonconserved targeting (Farh et al., 2005; Stark et al., 2005).

Ever since the availability of whole-genome multiple alignments (Blanchette et al., 2004), sites have been considered conserved if they are retained at orthologous locations in every genome under consideration and considered nonconserved or poorly conserved if they are missing or have changed in one of the genomes. This binary approach has been very productive but becomes less suitable now that the alignments include more than a few genomes. Requiring conservation in every species of a 28-genome alignment would exclude sites that are under strong selective pressure to be conserved in many genomes yet are missing at the orthologous position in some genomes either because of lineage-specific loss, gain, or substitution, or because of imperfections in sequencing, assembly, or alignment. To capture more of the conserved sites, a quantitative approach has been developed that makes the reasonable assumption that aligned sites within orthologous genes have a single origin and measures the portion of the phylogenetic tree that retains each site by summing the branch length over which each site has been preserved (Kheradpour et al., 2007). Because it represents an estimate of the amount of evolutionary time over which a site has been conserved, this branch-length score yields a multivalued metric that accounts for phylogenetic relationships between the species studied (Kheradpour et al., 2007). The score is interpreted by selecting a branch-length cutoff that separates more conserved and less conserved sites. Sliding the branch-length cutoff from zero to the total length of all branches enables tuning of sensitivity and specificity. This method has been applied to the 12-genome alignments of flies to predict conserved miRNA sites with sensitivity substantially improved over the previous binary approach (Kheradpour et al., 2007; Ruby et al., 2007) but has yet to be applied to mammalian site conservation.

While whole-genome alignments have made it simple to detect the conservation of sites in orthologous locations of genes, it is much more difficult to distinguish those sites or motifs under selective pressure to be maintained from those conserved by chance. A general attempt to detect preferential conservation of any motif used a simple Z-score test but did not control for genomic location or sequence characteristics (Xie et al., 2005). Another approach, developed for detecting maintenance of miRNA sites, has been to generate cohort sets of miRNA-like sequences, then determine the number of conserved sites that match these control sequences and use this as the estimate of chance conservation (Lewis et al., 2003). When choosing these controls carefully so as to avoid sites underrepresented in mRNA sequences, this approach has been effective for evaluating sets of miRNA sites in aggregate (Lewis et al., 2003, 2005; Brennecke et al., 2005; Krek et al., 2005; Stark et al., 2005; Lall et al., 2006; Kheradpour et al., 2007; Ruby et al., 2007). As previously implemented, however, this approach breaks down when examining individual miRNA-site interactions because of a failure to account adequately for differing mutational biases, dinucleotide conservation rates, and local conservation rates.

Here we develop an improved method for quantitatively evaluating site conservation and apply it to the study of vertebrate miRNA targeting. The improved sensitivity uncovered classes of sites, including offset seed matches and 3′-compensatory sites, whose conservation previously had not been detected with confidence. Overall, we find three times as many preferentially conserved sites as detected previously, thereby increasing the known scope and density of conserved miRNA regulatory interactions.

2.2 Results and discussion

2.2.1 Detection of seed match conservation with increased sensitivity and statistical power

When using a branch-length metric to evaluate motif conservation, the first step is to build a phylogenetic tree based on the genomic regions under investigation (Kheradpour et al., 2007), which in our case was $3'$ UTRs. One major innovation of our method was to build the phylogeny in a way that controlled for the conservation of individual UTRs. Because mutation, gene conversion, and crossover rates vary throughout the genome (Wolfe et al., 1989; Hwang and Green, 2004; Kauppi et al., 2004), different UTRs have substantially different background conservation levels. Moreover, some vertebrate genomes have low coverage and are missing a substantial fraction of genes, also affecting the apparent background conservation. Most methods for detecting positive selection take into account local conservation rates (Yang and Bielawski, 2000) (for example, Ks in Ka/Ks), but genome-scale methods for detecting purifying selection have thus far not accounted for this factor. In addition to differences in basal conservation rates, UTRs have sequence-dependent functions apart from miRNAs, which can influence conservation levels. A site falling within a UTR with high overall conservation is far less likely to be conserved due to miRNA targeting than is one falling within a rapidly evolving UTR (Lewis et al., 2005). Any method that treats all the UTRs the same greatly overestimates purifying selection of sites in well-conserved UTRs and underestimates purifying selection of sites in poorly conserved UTRs.

Starting with a 28-way vertebrate whole-genome alignment that included 18 placental mammals and 10 other vertebrates (Miller et al., 2007), we extracted the human $3'$ UTRs and homologous regions from the 22 non-fish genomes. The five fish genomes were excluded because they lacked a sufficient amount of aligned $3'$ UTR sequence. To help control for individual UTR conservation, $3'$ UTRs were separated by conservation rate into 10 equally sized bins, and a unique set of branch lengths based on

3′ UTR sequence alignments was constructed for each bin (Fig. 2-1B; Supplemental Fig. 1). The conservation of a given sequence (e.g., an 8mer miR-1 site in a particular 3′ UTR) was then assessed by summing the total branch length in the phylogenetic tree connecting the subset of species having the sequence perfectly aligned, using the tree representing the bin of the 3′ UTR under investigation. This branch-length value had no units, with a value of 1.0 corresponding to the average conservation of a single nucleotide in similar UTRs, and thus resembled a non-normalized version of the branch-length score described by Kheradpour et al. (2007). In our analyses, however, a site in a more divergent UTR needed to be conserved in fewer orthologs to achieve the same branch-length value because the branch lengths in a phylogeny representing the more divergent UTRs were longer than those of one representing more conserved UTRs. For example, the 8mer miR-1 site found within in the human *SLC35B4* UTR received the same value as the site within the *SPRED1* UTR, even though it was present in fewer aligned genomes (Fig. 2-1B).

Because sequences can be conserved by chance or for many reasons other than functional miRNA targeting, the branch-length values were only interpretable when considered within the context of the estimated background conservation. Our method attempted to control for many factors that can affect the conservation level of a short sequence of length k (a *k*-mer), including GC content, dinucleotide content, the interrelation of miRNA seed-match types, genome alignment quality, and the local conservation rate. The combined effects of all of these factors on background conservation were estimated based on em-pirically observed conservation of *k*-mers as opposed to theoretical calculations. As done previously (Lewis et al., 2003, 2005; Brennecke et al., 2005; Krek et al., 2005; Stark et al., 2005; Lall et al., 2006; Kheradpour et al., 2007; Ruby et al., 2007), the expected fraction of sites conserved due to miRNA recognition was estimated using a cohort of *k*-mers with similar properties, which were presumed to be subject to the same evolutionary pressures except for the possible miRNA regulatory relationship. Because we did not allow conserved *k*-mers that were seed matches for miRNAs with any known conservation, and because the discovery rate of highly conserved vertebrate miRNAs has dropped dramatically in recent years, the control k-mers can be assumed devoid of conservation due to miRNA targeting. Three substantial improvements to the estimation of background conservation were introduced. First, we matched control k-mers using an expected conservation based on both the k-mers GC content and the expected conservation of its constituent dinucleotides, which enabled a more rigorous and accurate estimate of background conservation levels for individual miRNAs (Fig. 2-1C). Previous methods matched k-mers based on their abundance in human 3′ UTRs, which is adequate when analyzing large groups of miRNAs, but this variable is poorly correlated to conservation for individual miRNAs (Fig. 2-1C) and can be affected by evolutionary avoidance of k-mers, a known property of miRNA seed matches (Farh et al., 2005; Stark et al., 2005). Second, we created mutually exclusive seed-match classes by subtracting the signal and the background of larger seed matches (e.g., 8mers) from the smaller seed matches that could be contained within them (e.g., 7mers, Fig. 2-1D). This protected against double-counting conservation while increasing sensitivity by more closely matching control k-mer sizes to the observed conservation (see Supplemental material for discussion). Third, the estimate of background conservation controlled for the conservation of individual UTRs, in that control cohorts were analyzed using the same 10 phylogenetic trees and the same 10 UTR data sets were employed for analysis of authentic sites. Without this improvement, different members of the control cohort had widely varying conservation. By reducing this variability, more precise background estimates were achieved, which enabled more sensitive detection of site conservation. Thus, we calculated 10 distributions of branch-length values for both signal and background using both the k-mer and its set of controls and then summed these distributions to compile the overall signal and background distributions for each k-mer (Fig. 2-1E; Supplemental Discussion). These three innovations all helped to control for the background conservation specific to individual seed-match sites, enabling statistically sound comparisons between the conservation of seed-match types, between seed matches to different miRNAs, and even between individual sites.

2.2.2 At least 44,000 sites are selectively maintained because of miRNA targeting

We first looked for excess conservation of seed matches for a set of highly conserved miRNAs that appear to have been present since the last common ancestor of all 23 vertebrate species under consideration, defined as those mammalian miRNAs also found in chicken, lizard, or fish, which fell into 87 families based on the identity of nucleotides 2-8 (Supplemental Table 1). For each k-mer, representing a single seed-match type for a particular miRNA, the distribution of branch-length values was compiled for sites present in human $3'$ UTRs. As the branch-length-value cutoff was increased from zero, the number of sites that matched control sequences decreased faster than did the number matching authentic miRNA seed matches (Fig. 2-2A). At any particular branch-length cutoff, if the number of conserved sites of a k-mer (the signal) was higher than that of control sequences (the background), the excess conservation was attributed to purifying selection. We use the term background instead of noise because the latter term may connote variance in the background estimate as opposed to the estimate itself. The number of sites conserved above background reflects the sensitivity of the analysis, whereas the ratio of signal to background reflects its specificity.

We first considered the three 7-8mer seed-match types (8mer, 7mer-m8, 7mer-A1), which correlate most strongly with targeting efficacy (Grimson et al., 2007; Nielsen et al., 2007) and are among the miRNA matches currently used to predict conserved targets of metazoan miRNAs (Fig. 2-2B) (Brennecke et al., 2005; Grün et al., 2005; Krek et al., 2005; Lewis et al., 2005; Lall et al., 2006; Ruby et al., 2006, 2007; Gaidatzis et al., 2007). At a branch-length cutoff of 2.0, a large majority of these sites were in excess of the background (Fig. 2-2B, right). However, this high specificity came at a price, with many more sites detected above background at a less stringent cutoff of 1.0 (Fig. 2-2B).

Our more precise estimate of background conservation enabled robust detection of purifying selection for 6mer seed matches that were not part of the larger, 7-8mer seed-matched sites (Fig. 2-2B). Ten thousand sites were conserved above background – a high number when considering the marginal efficacy of these 6mer sites, as measured by monitoring mRNA destabilization or protein output after adding or disrupting miRNAs (Grimson et al., 2007; Nielsen et al., 2007; Baek et al., 2008; Selbach et al., 2008). Analysis of mRNA expression following ectopic addition of miRNAs into HeLa cells indicated that an offset 6mer matching miRNA positions 3-8 (Fig. 2-1A) mediated mRNA destabilization approaching that of the seed-matched 6mer, matching positions 2-7 (Fig. 2-2C), although the effects of the seed-match 6mer were still significantly stronger (P = 0.03, two-sided KS test). This marginal yet detectable activity prompted us to explore the possibility that these offset 6mer sites might also be selectively maintained. Our analysis, subtracting conservation due to 7- or 8-nt seed-matched sites as well as that attributed to matching seeds of related miRNAs, indicated that a small but detectable fraction of these offset 6mers were indeed selectively maintained (Fig. 2-2B). Because these 6mer sites are so abundant in 3′ UTRs, this small fraction corresponded to thousands of sites under purifying selection, which have been missed by algorithms that search for only seed-matched sites.

The result for the offset 6mer raised the question of whether 6mer matches to nearby miRNA segments might also be selectively maintained. Analysis of matches to miRNA segments 1-6, 4-9, and 5-10, excluding those sites that also possessed seed matches, revealed no 6mer segments with appreciable signal above background (Supplemental Fig. 2). Parallel analyses of the mRNA expression data also failed to reveal 6mer sites with efficacy approaching that of the 6mer site corresponding to segment 3-8. We therefore focused on the selective conservation of the five types of sites that matched the seed region, one 8mer, two 7mers, and two 6mers, which we refer to as the 6mer and the offset 6mer (Fig. 2-1A).

When examined over a broad range of branch-length cutoffs, signal-to-background ratios plateaued at a branch-length cutoff of about 3 (Fig. 2-2D), which exceeded the maximal branch length of the more highly conserved UTR bins. Larger signal-to-background ratios implied higher fractions of seed matches under selection. For example, a signal-to-background ratio of 4.0 corresponds to 75% of matches being under purifying selection and thus presumably having conserved function. Regardless of the cutoff, the hierarchy of signal-to-background ratios remained constant, with 8mer > 7mer-m8 > 7mer-A1 > 6mer > offset 6mer. Moreover, the signal-to-background ratio of the five site types, which indicated the fraction of sites under selection, corresponded well with the minimal fraction of sites conferring transcript destabilization following microRNA transfection, indicating a striking correlation between the selective maintenance of site types and their efficacy (Fig. 2-2E).

When considering the number of selectively maintained sites, a moderate branch-length cutoff of 1.0 yielded the highest signal above background (Fig. 2-2F). Increasing cutoffs from 1.0 to 2.0 yielded a tradeoff between increased specificity (Fig. 2-2D) and decreased sensitivity (Fig. 2-2F). For the five individual site types, the number of selectively maintained sites showed little correlation with the signal-to-background ratio. For example, the signal-to-background ratio for the 6mer (1.2 at branch length 1.0) was far lower than that for the 8mer (2.6 at branch length 1.0), but signal above background for the 6mer (10,970 at branch length 1.0) was at least as high as that of the 8mer (8543 at branch length 1.0). Thus, $3'$ UTRs acquire and maintain marginally effective target sites in similar numbers as they do more highly effective sites. The 7mer-m8 sites appear most important in terms of the number of sites under selection (Fig. 2-2F), whereas 8mers are the most important in terms of the proportion of sequences under selection and, equivalently, the power for prediction of individual targets (Fig. 2-2D).

Summing together the signal and background estimates for the five site types at the most sensitive conservation cutoff (1.0) yielded 46,441 ± 2175 sites conserved above background (Fig. 2-2F), an average of 534 ± 25 per miRNA family (98 ± 2, 128 ± 7, 80 ± 8, 126 ± 22, 101 ± 14 for 8mer, 7mer-m8, 7mer-A1, 6mer, and offset 6mer sites, respectively). This number of sites was nearly three times higher than the most sensitive previous estimate, which had required perfect 6mer conservation in each of human, mouse, rat, and dog to detect 13,044 $3'$ UTR sites conserved above background, or 210 sites conserved per miRNA family (Lewis et al., 2005). Several factors contributed to this large increase in the estimate of selectively maintained miRNA sites, including the improved methodology, larger and more accurate UTR and miRNA data sets, new genomes, and improved genome quality. To determine whether the principal factor was the newly available genomes, we performed the same analysis on subsets of genomes, keeping the UTR and miRNA data sets and methodology constant (Fig. 2-2G). The sensitivity was robust to the removal of a large number of genomes, suggesting that with current methods, the estimate of the number of selectively maintained sites will remain relatively constant with the addition of newly sequenced genomes.

Detection of selectively maintained sites with higher sensitivity implied that the number of conserved miRNA targets is far higher than previously estimated. Starting with all the sites detected at a given conservation cutoff and then randomly removing for each site type the number of sites corresponding to the predicted background in the relevant UTR bin yielded 9909 ± 302 genes targeted at a branch-length cutoff of 1.0. Using this method of sampling conserved sites, only 7% of genes had multiple conserved sites for the same miRNA family. Thus, for each miRNA family, the number of conserved targets (497 ± 49) approached the number of conserved sites (534 ± 25). Although more sites above background were predicted at the conservation cutoff of 1.0, the number of genes targeted reached a maximum of 10,739 ± 564 at a branch-length cutoff of 0.6, which corresponded to 57.8% ± 3.0% of the human RefSeq data set. This percentage is about twice that of the most sensitive previous estimate (Lewis et al., 2005). Again, the number of targets per miRNA family (438 ± 60) approached the number of sites conserved above background per miRNA family (462 ± 28). Nonetheless, 72% of the 10,739 targeted messages had sites to multiple miRNA families, with an average of 4.2 sites per targeted $3'$ UTR. Indeed, the observed twofold increase in targeted UTRs from a threefold increase in site detection meant that our analysis added many additional newly predicted sites to previously predicted targets, thereby increasing not only the number of predicted targets but also the density of predicted targeting.

2.2.3 Sites with seed bulges and mismatches are rarely under selection

Having found a large and statistically significant number of conserved 6mer sites (Fig. 2-2F), despite their marginal efficacy (Fig. 2-2C), we investigated the possibility of selective conservation of imperfect seed matches, which also display severely compromised efficacy. Reasoning that if any mismatched sites were selectively maintained they would include those with the least disruptive mismatches, we focused on 8mer matches containing either a single mismatch, a G:U wobble, a bulged nucleotide within the site, or a bulged nucleotide within the miRNA. In contrast to the canonical seed-matched types, these imperfect sites displayed little enrichment of conservation (Fig. 2-3A). For all four mismatched classes, signal-to-background ratio hovered near 1.0, rarely exceeding 1.1 at any branch-length cutoff, indicating that the number of sites under selection was at most a small fraction of the total (Fig. 2-3A). The 8mer with a bulge in the site was the only class for which the 5% confidence limit on the ratio consistently exceeded 1.0. This class of sites appeared to have a few hundred sites conserved above background, a number 10 times less than that of even the weakest seedmatched class (Fig. 2-3B). We cannot exclude the possibility that a very small fraction of other mismatched sites might also have been selectively maintained. However, because of the low signal-to-background ratio and low 5% confidence estimate for the number of sites under selection, we conclude that seed-mismatched sites are hardly ever selectively maintained and that including a substantial number of such sites when predicting targets would greatly compromise prediction specificity. These conclusions are supported by recent proteomic experiments demonstrating poor efficacy of targets predicted by methods that allow sites with seed mismatches (Baek et al., 2008; Selbach et al., 2008).

Selective maintenance of sites with a bulge in the site but not those with a bulge in the miRNA corresponds well with previous analyses of plant miRNA targeting (Mallory et al., 2004). This constraint observed in both plant and animal lineages can be explained by the idea that the Argonaute protein binds the miRNA backbone, preorganizing the miRNA seed region such that the Watson-Crick face is poised for pairing to the message (Bartel, 2004). These contacts to the backbone in the seed region, presumably present before and after binding, would constrain the seed backbone, spacing each seed nucleotide such that a bulge in the miRNA would impose a gap in the site that would be difficult to span without disrupting adjacent pairs. In contrast, a bulged nucleotide in the site would be extruded into solvent and therefore more readily accommodated.

2.2.4 Pairing to the $3'$ end of miRNAs displays small but measurable excess conservation

Although pairing to the $3'$ region of the miRNA has long been thought to be consequential, evidence that such pairing enhances the efficacy of mammalian seed-matched sites has been obtained only recently (Grimson et al., 2007). Such sites in which $3'$ pairing productively augments seed pairing are called $3'$-supplementary sites. Productive $3'$ pairing optimally centers on miRNA nucleotides 13-16 and the UTR region directly opposite this miRNA segment (Fig. 2-4A, top). Like seed pairing, $3'$ pairing appears relatively insensitive to predicted thermostability and instead quite sensitive to pairing geometry, preferring contiguous Watson-Crick pairs uninterrupted by bulges, mismatches, or G:U wobbles. These features are captured in a $3'$-pairing score, which awards one point for each contiguous Watson-Crick pair matching miRNA nucleotides 13-16 and a half point for each contiguous pair extending the pairing in either direction. Pairing segments offset from the miRNA are then penalized by subtracting a half point for each nucleotide of offset beyond ± 2 nucleotides from the register directly opposite the miRNA, and then sites are assigned the score of the highest scoring pairing segment (Grimson et al., 2007). For example, the site shown in Figure 2-4A (top), which has seven contiguous, well-positioned pairs would be assigned a score of 5.5. Sites with scores ≥ 3 display modestly increased efficacy and conservation (Grimson et al., 2007).

We set out to determine for each site type the selective maintenance of $3'$-supplementary pairing. At specified cutoffs for branch length and 3′-pairing score we determined the number of sites with supplementary 3′ pairing, estimating the background by repeating the analysis with a chimeric miRNA set created by swapping all possible 5′ and 3′ ends for miRNAs within our 87 miRNA families. For each site, the 3′ pairing score used was the maximum over all members of the miRNA family. For each of the four seed-matched types, and especially for the 7mer-m8 site, selective maintenance of 3′-supplementary pairing was confidently observed (Fig. 2-4A). As expected for a biological signal, specificity increased with greater conservation and with a greater 3′-pairing score. Sensitivity peaked at a pairing score cutoff of 3.0, indicating that as few as 3-4 well-positioned supplementary pairs were selectively maintained (Fig. 2-4A). However, even at this sensitive cutoff, only 2281 ± 537 seed-matched sites had preferentially conserved 3′ pairing. Assuming that sites with selectively maintained 3′ pairing were also drawn from the pool of 44,000 sites with selectively maintained matches to the seed region, we estimate that only 4.9% ± 1.1% of all preferentially conserved sites have preferentially conserved 3′ pairing. Nonetheless, for those rare sites with high 3′ pairing scores, consideration of supplemental pairing provided a useful boost to the overall signal-to-background ratio. For example, for the 49 8mer sites with 3′-pairing scores $\geq$5.0 and branch-length values $\geq$2.0, the aggregate signal-to-background ratio was estimated to be 13:1 (calculated as 6.3 × 2.1, using values from Figs. 2-2D and 2-4A, respectively), implying that the conservation of these individual sites was confidently attributed to miRNA targeting. For the remaining 95.1% of selectively maintained seed matches, which do not have preferential conservation of pairing to the 3′ end of miRNAs, the 3′ region of the miRNA might still interact with the message, but in a way that does not favor matches over mismatches and therefore does not add detectably to targeting specificity.

2.2.5 Selective maintenance of 3′-compensatory sites

Pairing to the 3′ portion of the miRNA can not only supplement a 7-8mer match, it can also compensate for a single-nucleotide bulge or mismatch in the seed region, as illustrated by the *let-7* miRNA sites in *lin-41* and the miR-196 site in *HOXB8*

(Vella et al., 2004; Yekta et al., 2004). Such sites are called 3′-compensatory sites (Fig. 2-4B, top). Previous analyses of mRNA array data failed to detect efficacy of 3′-compensatory sites, even when considering the principles that uncovered consequential 3′-supplementary pairing, which are embodied in the 3′-pairing score (Grimson et al., 2007). This failure is attributed to the idea that 3′-compensatory sites are exceedingly rare, presumably because the amount of pairing needed to compensate for seed mismatches is greater than that needed to supplement seed-matched sites, and as a result, a significant association is difficult to detect (Grimson et al., 2007). Supporting this idea, all experimentally validated examples of metazoan 3′-compensatory sites involve pairing that centers on miRNA nucleotides 13-17 and extends to at least 9 contiguous Watson-Crick pairs, which corresponds to a 3′-pairing score $\geq$6.5 (Vella et al., 2004; Yekta et al., 2004).

An analysis of preferential site conservation could be more sensitive than that of array data, both because site-conservation analysis captures sites mediating translational repression without detectable mRNA destabilization and because site-conservation analysis can simultaneously consider many more miRNAs. With this possibility in mind, we examined the prevalence of 3′-compensatory sites, applying the same methodology as used for 3′-supplementary sites. Sites with single-nucleotide mismatches or bulges had no enrichment in conserved or nonconserved 3′ pairing exceeding the 5% confidence threshold, regardless of the cutoffs (Fig. 2-4B). This result indicated that most of the bulged sites conserved above background (Fig. 2-3) are not conserved because they are 3′-compensatory sites. However, 7-8mer sites with a single G:U wobble had confidently enriched pairing at the relatively high 3′-pairing-score cutoff of 4.0. At a branch-length cutoff of 1.0, the number of sites within the 5% confidence interval numbered only 399, an average of only 4.5 per miRNA family. Thus, our results support previous assertions that mismatched seed sites with 3′-compensatory pairing are only rarely under selective pressure to be conserved (Brennecke et al., 2005; Lewis et al., 2005). Perhaps because such sites with extensive pairing to the 3′ portion of the miRNA possess much more informational complexity than do the 7-8mer perfect matches and therefore emerge much less frequently and are harder to maintain in evolution, these 3′-compensatory sites appear to be used only rarely for biological targeting, comprising 1.5% ± 0.7% of the conserved sites in mammals. The somewhat higher number of 3′-supplementary sites implies that 3′ pairing, when consequential, is principally a supplementary feature of canonical seed sites and more rarely plays a role in conferring activity to imperfect seed sites.

The paucity of 3′-compensatory sites poses special challenges for confidently detecting conserved biological sites above background. Our use of the 3′-pairing score and our observation that the G:U class of mismatches was more frequently compensated by conserved 3′ pairing both represented important inroads into meeting this challenge. The 25 conserved G:U sites with 3′-pairing scores ≥ 6 include the miR-196 site in the *HOXB8* 3′ UTR, which has an off-scale 3′-pairing score of 9.0, and a similar miR-196 site in the *HOXC8* 3′ UTR (Yekta et al., 2004). These 25 sites, together with the seven mismatched sites with scores ≥ 7, are listed in Table 1 and will be included in the next release of TargetScan predictions (targetscan.org). Bulged sites with high 3′-pairing scores (≥ 6) did not appear preferentially conserved and thus are not included in the list.

2.2.6 Mammalian-specific miRNAs have few selectively maintained seed matches

An early study found that sites matching broadly conserved vertebrate miRNAs were more likely to be maintained than those matching mammalian-specific miRNAs (Lewis et al., 2003). Since then, target-prediction specificity has been estimated using only those miRNAs conserved to fish or chicken (Krek et al., 2005; Lewis et al., 2005; Gaidatzis et al., 2007), raising the question of whether the more recently emerged miRNAs have acquired enough conserved targets to detect any conservation signal above background. To address this question, we assembled a set of 53 miRNAs that were present in diverse placental mammals but absent in all sequenced chicken, lizard, and fish genomes (Supplemental Table 2). Examining the placental mammal subset of the phylogeny, we found little preferential conservation for sites matching these mammalian-specific miRNAs (Fig. 2-5A). In contrast to sites matching broadly conserved miRNAs (Fig. 2-2D), the full 8mer was the only seed-match type with signal-to-background ratio consistently and confidently above 1.0 (Fig. 2-5A), and its ratio was no higher than that of offset 6mers matching broadly conserved miRNAs.

The performance of the mammalian-only set also differed from that of the broadly conserved set when considering signal above background, with far fewer 8mers conserved above background (Fig. 2-5B). These differences could be due either to inherent differences in the miRNA sets, such as the level and breadth of expression, or to differential evolutionary time available for beneficial site emergence. To help differentiate between these possibilities, we screened matches to the broadly conserved miRNAs, removing all sites with seed matches conserved beyond mammals, thereby limiting the set of 8mer sites to those more likely to have arisen in mammals after the divergence of mammals and other vertebrates. This removed more than half of the conserved sites matching the broadly conserved miRNAs, showing that part of the reason for the higher number of sites is the much greater time available for beneficial site emergence. However, even when considering the more restricted set of sites, and after normalizing for the numbers of miRNAs in the two sets, the broadly conserved miRNAs had more than four times as many selectively maintained 8mer matches per miRNA than did the mammalian-specific miRNAs (Fig. 2-5B), suggesting that the level and breadth of miRNA expression are also important factors. Combining the signal above background for all site types, the differential appeared far greater, with the mammalian miRNAs averaging only 10.9 ± 3.0 selectively maintained sites.

To determine whether a few of the mammalian-only miRNAs might have conservation patterns resembling those of the broadly conserved set, we examined the conservation of 8mer seed matches corresponding to individual miRNAs. As expected, the signal-to-background ratios of the broadly conserved miRNAs fell mostly outside the control distribution (Fig. 2-5C). The observation that most fell outside the control distribution illustrated that the high signals observed for these broadly conserved miRNAs in aggregate also applied to most of them individually and could not be attributed to chance overlap of a few seed matches to a few highly conserved regulatory sequences. In contrast, the mammalian-specific miRNAs had a distribution only slightly shifted from that of the controls, with an excess of ≈5-10 miRNAs in the 1.5-2.5 signal-to-background range. We conclude that only a small subset of mammalian-specific miRNAs (including most prominently miR-487b, miR-127-3p, miR-379, and miR-543; Supplemental Table 2) have a measurable number of selectively maintained sites and caution that for the majority of mammalian-specific miRNAs, the observation that a site is conserved provides no evidence that it has biological function.

2.2.7 Estimating confidence for selective maintenance of individual sites

The widespread scope of conserved targeting brings to the fore the question of which of these many miRNA target interactions can be predicted with confidence. One raw indicator is the conservation branch length of the site – clearly, more highly conserved sites are more likely to be under selection, particularly when controlling for differential UTR conservation, as in our method. However, our branch-length values did not account for the type of site and its sequence features. For example, a branch length of 1.0 for an 8mer is far more compelling evidence for selective maintenance than a branch length of 3.0 for an offset 6mer (Fig. 2-2D). To control for site type and sequence features, we used our previously described controls to calculate a signal-to-background ratio (S/B) for each site at each branch length. For the purpose of evaluating individual sites, assessing controls at each branch length instead of at each branch-length cutoff is necessary to avoid crediting poorly conserved sites for having the same sequence as many highly conserved sites. We then converted this S/B to a probability of conserved targeting (P_{CT}), which is approximately equal to (S/B - 1)/(S/B) (or near zero, for sites with S/B $<$ 1). This score reflected the Bayesian estimate of the probability that a site is conserved due to selective maintenance of miRNA targeting rather than by chance or any other reason not pertinent to miRNA targeting, allowing for uncertainty in the S/B ratio. For predicting biologically conserved target sites, the score represented 1 - FDR, where FDR was our estimate of the false-discovery rate. This deceptively simple value incorporated knowledge of the conservation level of a particular site, the seed-match type, the number of selectively maintained seed matches for the particular miRNA, the background conservation level for the k-mer, and the UTR conservation context.

The P_{CT} provides a useful criterion for assessing the biological relevance of predicted miRNA-target interactions. Selectively maintained sites are more likely to have detectable biological function, are more likely to have functions in experimental animals that are pertinent to humans, and tend to be more effective (Grimson et al., 2007; Nielsen et al., 2007). To illustrate the ability of this measure to predict site efficacy, we turned to experimental data examining mRNA destabilization after introducing miRNAs (Grimson et al., 2007). As expected, site P_{CT} correlated with the mean level of mRNA destabilization (Fig. 2-6A). In addition, we tested a subset of miRNA sites that were previously considered nonconserved because they were not found in mouse, rat, or dog alignments (Lewis et al., 2005). Those sites newly recognized as selectively maintained were substantially more responsive to the transfected miRNAs than were those still lacking a signal for conserved targeting (Fig. 2-6B).

Having established a more sensitive and precise tool for evaluating site conservation, we overhauled the TargetScan web resource, separating conserved and highly conserved 3′ UTR targets of each miRNA based on the P_{CT} values of their 7-8mer sites, compiling an aggregate probability of conserved targeting (aP_{CT}) for those targets with multiple sites for that miRNA, calculated as $aP_{CT} = 1 - (FDR_{site1} \times FDR_{site2} \times FDR_{site3}...)$. To capture sites that were missing in the annotated human 3′ UTRs but present in the mouse annotations, a mouse-centric version of the analyses was performed in parallel. This improved web resource, called TargetScan version 5.0, also lists the P_{CT} for each 7-8mer site, with the option of sorting the predicted targets of each miRNA by their aP_{CT}. Retained in TargetScan 5.0 are site context scores, which consider features such as the AU content in the vicinity of the site and the position of the site within the message, to predict the function and quantitative efficacy of each site (Grimson et al., 2007). Because the context scores are determined based on information completely orthogonal to site conservation, the P_{CT} values and the context scores provide independent and complementary information useful for predicting the biological relevance and efficacy of each site.

2.2.8 The widespread scope of conserved targeting

One of the key results of applying our new methods for quantitatively evaluating site conservation to the study of miRNA targeting was the sheer number of preferentially conserved sites. When considering only the 6-8mer perfect 3′ UTR matches to the seed regions of broadly conserved miRNAs, 46,441 ± 2175 sites were conserved above background, implying that 57.8% ± 3.0% of the human genes are conserved miRNA targets. Considering also imperfectly matched sites added another 625 ± 239 preferentially conserved sites to the total (Fig. 2-3B); 3′-compensatory sites added another 714 ± 315 (Fig. 2-4B), sites matching more narrowly conserved miRNAs added another 578 ± 158 (Fig. 2-5B), altogether modestly raising the estimated number of preferentially conserved 3′ UTR sites to 48,360. Not considered are the human miRNAs that have escaped confident annotation. However, because these miRNAs have remained unannotated, in part because they are more narrowly conserved and have more restrictive expression domains, our results for the mammalian-specific miRNAs suggest that eventual consideration of these unannotated miRNAs will add only modestly to the current picture of conserved targeting. We suspect that a more significant increase will come by considering targeting in ORFs, which is thought to be widespread, although less than that in 3′ UTRs (Lewis et al., 2005; Stark et al., 2007b). Further increases will come with improved UTR annotations and the consideration of alternative 3′ UTRs. Anticipating these additional sites, we can say with confidence that over 60% of human protein-coding genes are conserved targets of miRNAs, with our best estimate of the actual fraction of human protein-coding genes under pressure to maintain pairing to miRNAs falling above this, at about two thirds of all protein-coding genes.

One surprise of our analysis was the substantial number of selectively maintained 6mer and offset 6mer sites, numbering 10,970 ± 1909 and 8803 ± 1276, respectively.

Indeed, when excluding these sites and considering only the 7-8mer sites, our estimate of the number of conserved targets per broadly conserved miRNA dropped to 292 ± 18, which corresponded to 44.5 ± 3.4% of the RefSeq genes. With the exception of an offset 6mer site for let-7 miRNA in the human *LIN28* 3′ UTR (Wu and Belasco, 2005), 6mer sites typically have very poor efficacy when examined experimentally, as if the majority are inert, and nearly all the rest have very marginal influence on protein output (Fig. 2-2C) (Grimson et al., 2007; Nielsen et al., 2007; Baek et al., 2008; Selbach et al., 2008). Yet 3′ UTRs selectively maintain these 6mer target sites in similar numbers as they do 7 and 8mer sites. Of course, 6mers are much easier to access by mutation and then preserve from mutation, but what would be the selection pressure to preserve such minor effects? One conundrum in the field of miRNA-mediated gene regulation is why so many 7-8mer sites would be conserved so broadly in metazoan 3′ UTRs, when each down-regulates protein output by very little – nearly always less than 50% and usually less than 30%, which would only very rarely produce observable phenotypic consequences (Baek et al., 2008). But the 7-8mer conundrum pales when considering the selective maintenance of 6mers, which appear to tweak gene expression so finely that the effects are difficult to detect at the molecular level, let alone the phenotypic level.

One way to reconcile the high number of preferentially conserved 6-nt sites with the modest efficacy of these sites is to propose that these conserved 6mers are not preferentially conserved based on their activity as 6-nt sites, but instead, represent the inactive (or less active) decay products of preferentially conserved 7-8mer sites. When considering this explanation, two scenarios could contribute the conserved 6mer signals. In one scenario, the extended site has degraded to one of the 6mers in the human lineage. For example, a 7mer site that functions in most animals but has decayed to an inactive 6mer in primates would have contributed to our count of conserved 6mers in human UTRs. In the second scenario, the extended site is conserved in the UTR of human and other species but exceeds the branch-length cutoff only when including additional species in which it has degraded to one of the 6mers. For example, if the 8mer site conserved in the *SPRED1* UTR had degraded to an inactive

6mer in hedgehog, then the site would have contributed to our count of conserved 8mers at branch-length cutoff of 0.9, but not at 1.0; at a cutoff of 1.0 it would have contributed to our count of conserved 6mers. When performing the analyses so as to exclude these two scenarios, our estimate of the number of preferentially conserved sites per highly conserved miRNA family dropped 35%, to an average of 77 ± 22 for 6mer sites and 69 ± 15 for offset 6mer sites (Supplemental Discussion). Thus, some of the preferential conservation of 6-nt matches to the seed region can be explained by their relationship with conserved 7mers, but thousands of 6mers and offset 6mers are selectively maintained without the aid of 7mer conservation. Moreover, because these decreases in preferentially conserved 6-nt sites imply corresponding increases in preferentially conserved 7-8mer sites, consideration of these scenarios does not change our estimates of the total number of preferentially conserved sites and the total number of conserved mammalian targets.

Our observation that these 6mer sites have been selectively maintained so frequently implies the existence of widespread pressure for finely adjusted protein output with surprisingly narrow tolerances for optimal expression in different cell types. The observation that this selective pressure persists over such long branch lengths indicates that these optimal expression levels and narrow tolerances persist in the face of many other changes over surprisingly long evolutionary distances. These results become somewhat easier to reconcile when considering the possibility that 6mers might impart more robust repression in particular cells or conditions that have yet to be accessed experimentally. Hinting at this possibility is the more readily detected activity of 6mers matching miR-430 during the maternal-to-zygotic transition, a developmental stage at which the miR-430 family comprises most of the miRNA molecules in the animal (Giraldez et al., 2006). An alternative possibility is that the 6mer and offset 6mer impart some function other than repressing protein output. For example, if transient miRNA association played a widespread role in mRNA subcellular localization, then many 6mer sites could be conserved without imparting any repression. In either scenario – widespread and persistently narrow gene-expression tolerances or expanded miRNA function – the discovery that so many 6mers and offset 6mers are selectively maintained sets the stage for important future insights into miRNA biology.

2.3 Methods

2.3.1 Alignments and phylogeny

Sequences (3′ UTR) were defined as the longest 3′ UTR for each human RefSeq gene (Pruitt et al., 2005), resulting in 18,577 unique UTRs for 18,577 genes. Human UTR boundaries were used to acquire orthologous UTRs from the 28-genome vertebrate sequence alignments from the UCSC genome browser (Karolchik et al., 2008). The average phylogeny was determined using the branching structure provided by UCSC, and branch lengths were estimated by running DNAML, part of the PHYLIP package (Felsenstein, 1989), on the UTR alignments. To define UTR conservation bins, human UTRs were ranked using the median branch length over single bases. Then, DNAML was run on the sequences in each UTR bin, creating phylogenies with the same branching structure but modified branch lengths scaled for each UTR bin.

2.3.2 miRNA sequences

The broadly conserved miRNA set comprised all 79 families of miRBase entries (release 10.0) with both a human and zebrafish miRNA sharing the same seed sequence (Supplemental Table 1). Additionally, we included 8 human miRNA families with mouse miRBase entries and a conserved foldback in lizard, chicken, frog, or another fish genome, yielding a total of 87 miRNA families. To generate the mammalian-specific miRNA set, miRbase entries for human miRNAs conserved to mouse were manually inspected for aligned sequences from most of the sequenced mammals, perfect conservation of the seed sequence within placental mammals that had aligned sequence, no conservation in species other than placental mammals, and good predicted folding characteristics in most placental mammals. Seven of the remaining miRNAs whose seed matches had strong excess conservation (>2.0 S/B ratio) in species other than placental mammals were also removed (reasoning that this excess conservation was likely caused by an miRNA ortholog that escaped detection because of imperfect alignments or sequencing coverage), leaving 53 mammalian-specific miRNAs (Supplemental Table 2). A list of miRNAs left out of the mammalian-specific category for various reasons is found in Supplemental Table 3.

2.3.3 Background estimation

For each seed-match k-mer, possible control k-mers were first filtered to remove seed matches to other miRNAs, and to have the same length, number of G + C bases, and possible matches to PUF proteins (UGU[A/G]) as the seed match. In the case of 8mer and 7mer-A1 matches, the controls were also constrained to have an A in the $3'$-most nucleotide. For each branch-length cutoff, the 50 control k-mers with the closest expected conservation rate to the seed-match k-mer were selected. Expected conservation was calculated using a first-order Markov model with parameters derived from the empirical dinucleotide conservation rate in $3'$ UTRs at a particular branch-length cutoff. In other words, control k-mers were picked to exactly match length, GC content, and PUF binding sites, and to closely match the expected conservation rate based on the dinucleotide content. The background estimate was the number of occurrences of the k-mer multiplied by the average fraction of sites conserved in control k-mers at the same branch length. As done for the signal, background was determined for each UTR bin separately, and then at each branch length, background from the 10 UTR bins was summed to generate the overall background distribution. At each branch length, confidence intervals were calculated using the Gaussian distribution with mean equal to the background estimate and standard deviation calculated using the background estimates given by individual control k-mers as opposed to the average over all 50.

2.3.4 Seed-match types

For the five major seed-match types (8mer, 7mer-m8, 7mer-A1, 6mer, offset 6mer), raw signal and background values were calculated without regard to whether a shorter seed match was nested within a longer one. Then, for each miRNA and for each branch-length cutoff, the signal for the 8mer match was subtracted from the signal of the corresponding 7mer-m8 match, and the 8mer background estimate was subtracted from the 7mer-m8 estimate. Similarly, 8mer signal and background were subtracted from signal and background of the other seed-match types, 7mer-m8 signal and background were subtracted from that of both 6mers and offset 6mers, and 7mer-A1 signal and background was subtracted from that of 6mers. In cases in which a particular seed match could be considered a 7mer-m8 match to one miRNA and a 7mer-A1 to another miRNA, it was considered only as a 7mer-m8 match to avoid double-counting. Likewise, seed matches ambiguous between 6mers and offset 6mers were considered as only 6mers. Seed matches with mismatches, G:U wobbles, or bulges were filtered so that they did not contain perfect seed matches to other miRNAs.

2.3.5 $3'$ pairing

Pairing scores ($3'$) were calculated as previously described (Grimson et al., 2007). For sites matching miRNA families comprising miRNAs with different $3'$ sequences, the $3'$-pairing score used was the maximum score for any member of the miRNA family. The branch length of a site at a particular $3'$-pairing score was calculated for the set of species having both the seed match conserved and a $3'$-pairing score at least that of the human score. Control $3'$-pairing scores and conservation were estimated by swapping every miRNA seed and $3'$-end family, and recalculating $3'$-pairing scores for every combination. Background estimate and confidence intervals were calculated as before, except that there were 86 control $3'$-end families for each swap instead of 50 control k-mers. For compensatory pairing sites in Table 1, orthologous sites were considered conserved if their $3'$-pairing score was greater than 6.0, regardless of the human pairing score. This was done to capture well-conserved human sites that have recently extended already strong $3'$ pairing.

2.3.6 Probability of conserved targeting

For each human seed-match site, the conservation signal and background were calculated using methods analogous to those described above. Rather than using a branch-length cutoff, we used a branch-length window with a size set to meet a minimum of 20 occurrences of the site in human UTRs and calculated the fraction of seed matches conserved within the branch-length window. In the few cases with less than 20 total occurrences of the seed match, all corresponding sites were assigned a P_{CT} of 0. For the remaining sites, the P_{CT} was defined as $\mathrm{E}[(S - B)/S]$ where B, the background estimate, is a constant, and S is a random variable. S is defined as $\mathrm{max} S_{obs}/N, B$, where S_{obs} is the observed signal and N is the total number of occurrences, distributed around the observed total N_{obs} as Poisson(N_{obs}). Because the background estimate is based on a mean of many measurements, but the signal is based on a single observation, we fix the number of conserved occurrences (S_{obs}) and allow the total number of occurrences (N) to vary. Thus, the P_{CT}, which ranges between 0 and 1, corresponds to a Bayesian estimate of the probability that a site conserved to a particular branch length is conserved due to miRNA targeting. For some miRNAs, we observed high variability of P_{CT} values for sites with close branch-length values; therefore, we implemented a smoothing procedure when deriving P_{CT} values reported at the TargetScan site (Supplemental Discussion).

2.4 Acknowledgments

We thank members of the Bartel and Burge laboratories for helpful discussions, and the Broad Institute of MIT and Harvard, the Human Genome Sequencing Center at the Baylor College of Medicine, the Genome Sequencing Center at Washington University, and the Department of Energy's (DOE's) Joint Genome Institute for providing access to unpublished vertebrate genome sequences. We also thank George Bell, Katherine Gurdziel, Prathapan Thiru, Bingbing Yuan, and Fran Lewitter for implementing our improved methods at the TargetScan website. This work was supported by a Krell Institute/Department of Energy Computational Sciences Graduate Fellowship (to R.C.F.) and by grants from the NIH to C.B.B. and D.P.B.

2.5 References

D. Baek, J. Villén, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, Sep 2008. doi: 10.1038/nature07242.

D. P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116 (2):281–97, Jan 2004.

M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, and W. Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, 14(4):708–15, Apr 2004. doi: 10.1101/gr.1933104.

J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen. Principles of microRNA-target recognition. *PLoS Biol*, 3(3):e85, Mar 2005. doi: 10.1371/journal.pbio.0030085.

K. K.-H. Farh, A. Grimson, C. Jan, B. P. Lewis, W. K. Johnston, L. P. Lim, C. B. Burge, and D. P. Bartel. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, 310(5755):1817–21, Dec 2005. doi: 10.1126/science.1121158.

J. Felsenstein. PHYLIP: phylogenetic inference package. *Cladistics*, Jan 1989.

D. Gaidatzis, E. van Nimwegen, J. Hausser, and M. Zavolan. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, 8:69, Jan 2007. doi: 10.1186/1471-2105-8-69.

A. J. Giraldez, Y. Mishima, J. Rihel, R. J. Grocock, S. V. Dongen, K. Inoue, A. J. Enright, and A. F. Schier. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science*, 312(5770):75–9, Apr 2006. doi: 10.1126/science.1122689.

A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell*, 27(1):91–105, Jul 2007. doi: 10.1016/j.molcel.2007.06.017.

D. Grün, Y.-L. Wang, D. Langenberger, K. C. Gunsalus, and N. Rajewsky. microRNA target predictions across seven Drosophila species and comparison to mammalian targets. *PLoS Comput Biol*, 1(1):e13, Jun 2005. doi: 10.1371/journal.pcbi.0010013.

D. G. Hwang and P. Green. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA*, 101(39):13994–4001, Sep 2004. doi: 10.1073/pnas.0404142101.

D. Karolchik, R. M. Kuhn, R. Baertsch, G. P. Barber, H. Clawson, M. Diekhans, B. Giardine, R. A. Harte, A. S. Hinrichs, F. Hsu, K. M. Kober, W. Miller, J. S. Pedersen, A. Pohl, B. J. Raney, B. Rhead, K. R. Rosenbloom, K. E. Smith, M. Stanke, A. Thakkapallayil, H. Trumbower, T. Wang, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res*, 36 (Database issue):D773–9, Jan 2008. doi: 10.1093/nar/gkm966.

L. Kauppi, A. J. Jeffreys, and S. Keeney. Where the crossovers are: recombination distributions in mammals. *Nat Rev Genet*, 5(6):413–24, Jun 2004. doi: 10.1038/nrg1346.

P. Kheradpour, A. Stark, S. Roy, and M. Kellis. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res*, 17(12):1919–31, Dec 2007. doi: 10.1101/gr.7090407.

A. Krek, D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky. Combinatorial microRNA target predictions. *Nat Genet*, 37(5):495–500, May 2005. doi: 10.1038/ng1536.

S. Lall, D. Grün, A. Krek, K. Chen, Y.-L. Wang, C. N. Dewey, P. Sood, T. Colombo, N. Bray, P. MacMenamin, H.-L. Kao, K. C. Gunsalus, L. Pachter, F. Piano, and N. Rajewsky. A genome-wide map of conserved microRNA targets in C. elegans. *Curr Biol*, 16(5):460–71, Mar 2006. doi: 10.1016/j.cub.2006.01.050.

P. Landgraf, M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A. O. Kamphorst, M. Landthaler, C. Lin, N. D. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Foà, J. Schliwka, U. Fuchs, A. Novosel, R.-U. Müller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D. B. Weir, R. Choksi, G. D. Vita, D. Frezzetti, H.-I. Trompeter, V. Hornung, G. Teng, G. Hartmann, M. Palkovits, R. D. Lauro, P. Wernet, G. Macino, C. E. Rogler, J. W. Nagle, J. Ju, F. N. Papavasiliou, T. Benzing, P. Lichter, W. Tam, M. J. Brownstein, A. Bosio, A. Borkhardt, J. J. Russo, C. Sander, M. Zavolan, and T. Tuschl. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129(7): 1401–14, Jun 2007. doi: 10.1016/j.cell.2007.04.040.

B. P. Lewis, I. hung Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–98, Dec 2003.

B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, Jan 2005. doi: 10.1016/j.cell.2004.12.035.

L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027): 769–73, Feb 2005. doi: 10.1038/nature03315.

W. H. Majoros and U. Ohler. Spatial preferences of microRNA targets in 3' untranslated regions. *BMC Genomics*, 8:152, Jan 2007. doi: 10.1186/1471-2164-8-152.

A. C. Mallory, B. J. Reinhart, M. W. Jones-Rhoades, G. Tang, P. D. Zamore, M. K. Barton, and D. P. Bartel. MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region. *EMBO J*, 23(16):3356–64, Aug 2004. doi: 10.1038/sj.emboj.7600340.

W. Miller, K. Rosenbloom, R. C. Hardison, M. Hou, J. Taylor, B. Raney, R. Burhans, D. C. King, R. Baertsch, D. Blankenberg, S. L. K. Pond, A. Nekrutenko, B. Giardine, R. S. Harris, S. Tyekucheva, M. Diekhans, T. H. Pringle, W. J. Murphy, A. Lesk, G. M. Weinstock, K. Lindblad-Toh, R. A. Gibbs, E. S. Lander, A. Siepel, D. Haussler, and W. J. Kent. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res*, 17(12):1797–808, Dec 2007. doi: 10.1101/gr.6761107.

C. B. Nielsen, N. Shomron, R. Sandberg, E. Hornstein, J. Kitzman, and C. B. Burge. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*, 13(11):1894–910, Nov 2007. doi: 10.1261/rna.768207.

K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 33(Database issue):D501–4, Jan 2005. doi: 10.1093/nar/gki025.

A. Rodriguez, E. Vigorito, S. Clare, M. V. Warren, P. Couttet, D. R. Soond, S. V. Dongen, R. J. Grocock, P. P. Das, E. A. Miska, D. Vetrie, K. Okkenhaug, A. J. Enright, G. Dougan, M. Turner, and A. Bradley. Requirement of bic/microRNA-155 for normal immune function. *Science*, 316(5824):608–11, Apr 2007. doi: 10.1126/science.1139253.

J. G. Ruby, C. Jan, C. Player, M. J. Axtell, W. Lee, C. Nusbaum, H. Ge, and D. P. Bartel. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell*, 127(6):1193–207, Dec 2006. doi: 10.1016/j.cell.2006.10.040.

J. G. Ruby, A. Stark, W. K. Johnston, M. Kellis, D. P. Bartel, and E. C. Lai. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res*, 17(12):1850–64, Dec 2007. doi: 10.1101/gr.6597907.

M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455 (7209):58–63, Sep 2008. doi: 10.1038/nature07228.

A. Stark, J. Brennecke, N. Bushati, R. B. Russell, and S. M. Cohen. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123(6):1133–46, Dec 2005. doi: 10.1016/j.cell.2005.11.023.

A. Stark, P. Kheradpour, L. Parts, J. Brennecke, E. Hodges, G. J. Hannon, and M. Kellis. Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Res*, 17(12):1865–79, Dec 2007a. doi: 10.1101/gr.6593807.

A. Stark, M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, J. G. Ruby, J. Brennecke, H. F. curators, B. D. G. Project, E. Hodges, A. S. Hinrichs, A. Caspi, B. Paten, S.-W. Park, M. V. Han, M. L. Maeder, B. J. Polansky, B. E. Robson, S. Aerts, J. van Helden, B. Hassan, D. G. Gilbert, D. A. Eastman, M. Rice, M. Weir, M. W. Hahn, Y. Park, C. N. Dewey, L. Pachter, W. J. Kent, D. Haussler, E. C. Lai, D. P. Bartel, G. J. Hannon, T. C. Kaufman, M. B. Eisen, A. G. Clark, D. Smith, S. E. Celniker, W. M. Gelbart, and M. Kellis. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, 450(7167):219–32, Nov 2007b. doi: 10.1038/nature06340.

M. C. Vella, E.-Y. Choi, S.-Y. Lin, K. Reinert, and F. J. Slack. The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR. *Genes Dev*, 18(2):132–7, Jan 2004. doi: 10.1101/gad.1165404.

K. H. Wolfe, P. M. Sharp, and W. H. Li. Mutation rates differ among regions of the mammalian genome. *Nature*, 337(6204):283–5, Jan 1989. doi: 10.1038/337283a0.

L. Wu and J. G. Belasco. Micro-RNA regulation of the mammalian lin-28 gene during neuronal differentiation of embryonal carcinoma cells. *Mol Cell Biol*, 25(21):9198–208, Nov 2005. doi: 10.1128/MCB.25.21.9198-9208.2005.

X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–45, Mar 2005. doi: 10.1038/nature03441.

Z. Yang and J. Bielawski. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol (Amst)*, 15(12):496–503, Dec 2000.

S. Yekta, I. hung Shih, and D. P. Bartel. MicroRNA-directed cleavage of HOXB8 mRNA. *Science*, 304(5670):594–6, Apr 2004. doi: 10.1126/science.1097434.

2.6 Figures

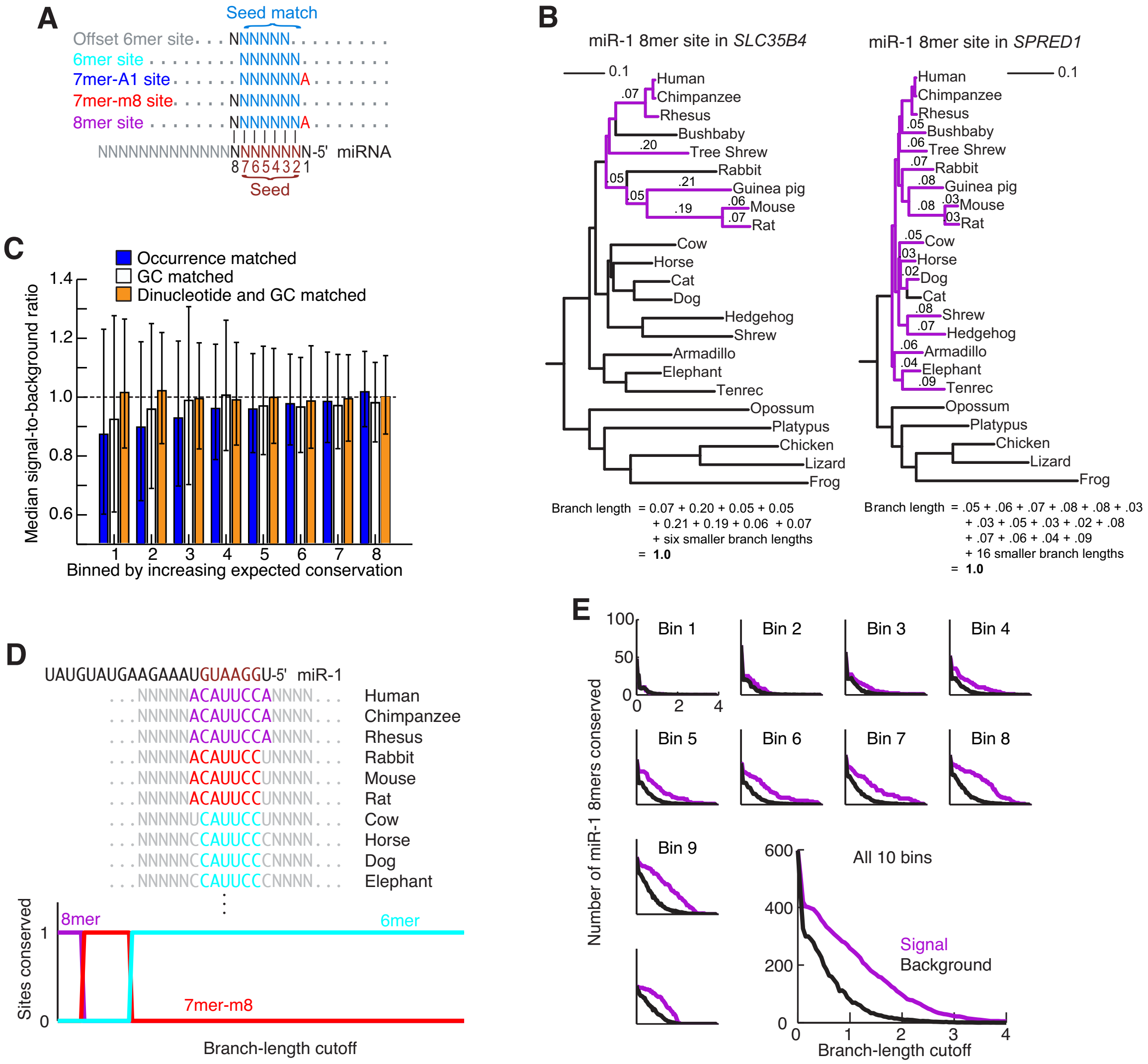


Figure 2-1: Method for detecting preferential conservation of miRNA sites. (A) Sites matching the miRNA seed region. All four canonical sites (colored) share six contiguous Watson-Crick matches to the miRNA seed (nucleotides 2-7); the offset 6mer contains six contiguous matches to nucleotides 3-8. (B) Phylogenetic conservation of individual sites. Each panel represents a miR-1 8mer site conserved to a branch length of 1.0. The 3′ UTR of *SLC35B4* falls into the fourth-most poorly conserved bin, which has a phylogeny with relatively long branch lengths (left), whereas the 3′ UTR of *SPRED1* falls into the most well-conserved bin, which has a phylogeny with shorter branch lengths (right). Branch segments connecting the species with sites are colored purple, with numbers indicating the lengths of the longer segments. The lengths of the segments connecting the species are summed to yield the branch-length score (equation). (C) Performance of controls matched for number of occurrences in human 3′ UTRs, GC content, or expected conservation as calculated by a first-order Markov model. For each possible RNA 7mer, conservation rates (signal) and average conservation rates for 50 control 7mers (background) were calculated for all 3′ UTRs at a branch-length cutoff of 1.0. Error bars indicate 25th and 75th percentiles. Because most 7mer motifs are not selectively maintained, well-performing controls should yield median signal-to-background ratios near 1.0, with low variability for every bin. (D) Nested site conservation. Because the seed-matched sites are interrelated, we subtract conserved instances of extended sites from those of shorter sites. This hypothetical site is an 8mer match to miR-1 in a human 3′ UTR that is more broadly conserved as a 7mer-m8 than as an 8mer site, and as a 6mer than as a 7mer-m8 site. As a result of nested subtraction, our method considers this site an 8mer at low branch-length cutoffs but not a 6mer or 7mer. At moderate branch-length cutoffs, it switches to a conserved 7mer-m8 site, and at high branch-length cutoffs it switches to a conserved 6mer site but not a 7mer or 8mer. (E) Signal and background for the miR-1 8mer site. For each UTR bin 1 through 10, with bin 1 having the least conserved UTRs and bin 10 the most conserved, the number of miR-1 sites conserved at the indicated branch-length cutoff is plotted with estimated background (small plots). Results for all 10 bins were then combined to represent the aggregate signal and background for this site (large plot).

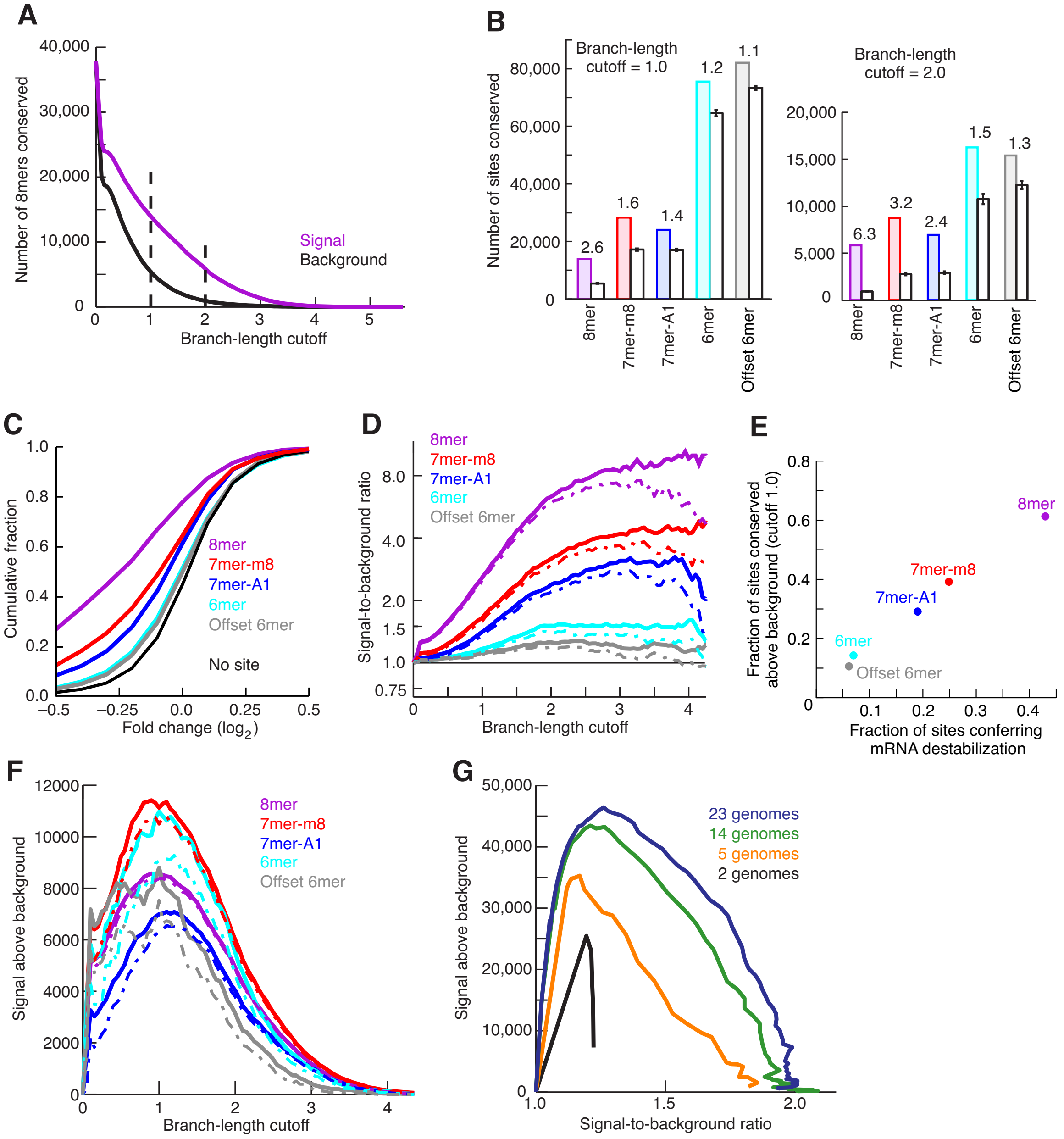


Figure 2-2: Conservation of major seed-match types. (A) Conservation of 8mer sites for 87 broadly conserved miRNA families. High-sensitivity and high-specificity cutoffs are highlighted with broken lines at 1.0 and 2.0, respectively. (B) Conservation and background estimate for mutually exclusive site types at high sensitivity (left) and high specificity (right). The signal-to-background ratio is indicated above the pair of bars. Error bars indicate one standard deviation in the estimated background, based on subsampling of individual control *k*-mers. (C) Efficacy of offset 6mer sites. Microarray data monitoring mRNA destabilization following transfection of 11 miRNAs was analyzed as described previously (Grimson et al. 2007). Shown is the cumulative distribution of changes for transcripts containing exactly one offset 6mer site and no other canonical sites in their 3′ UTR. For comparison, previously reported analyses of messages with single canonical sites are also shown (Grimson et al. 2007). (D) Signal-to-background ratio for indicated sites at increasing branch-length cutoff. Broken lines indicate 5% lower confidence limit (*z*-test). (E) Correlation of site conservation rate and experimental efficacy. Fraction of sites conserved above background was calculated as ([SignalBackground]/Signal) at a branch-length cutoff of 1.0. The minimal fraction of sites conferring destabilization was determined from the cumulative distributions (C), considering the maximal vertical displacement from the no-site distribution (correcting for the bumpiness of the distributions as described previously (Grimson et al. 2007)). (F) Estimates of signal above background for the major site types. Broken lines indicate 5% lower confidence limit (*z*-test). (G) Aggregate conservation above background for all major site types when using using subsets of genomes. To facilitate overlay of the plots, the X-axis is signal-to-background ratio rather than branch-length cutoff. The 14-genome subset represents the non-fish species originally available in the UCSC 17-way alignments. The five-genome subset contains human, mouse, rat, dog, and chicken, and the two-genome subset contains only human and mouse.

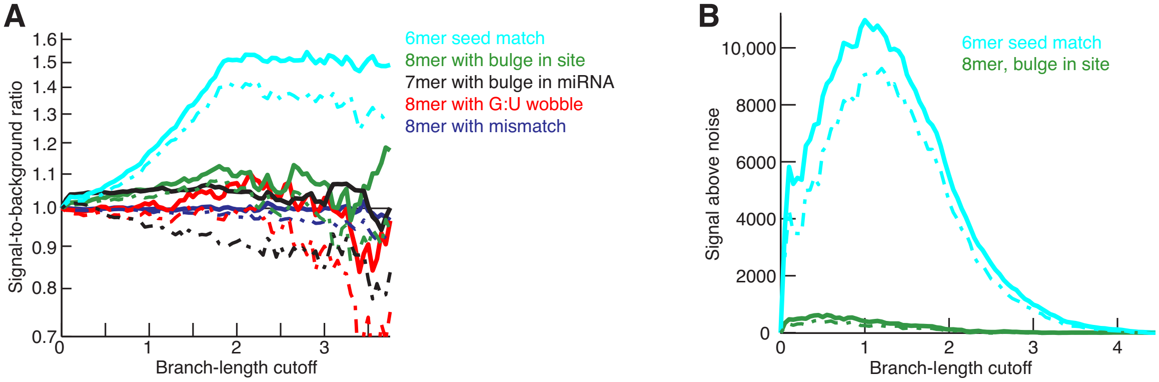


Figure 2-3: Occasional preferential conservation of imperfect sites. (A) Signal-to-background ratio for sites with the indicated single-nucleotide mismatches and bulges. A mismatch or G:U wobble must occur opposite miRNA seed nucleotides 2-7. A bulge in the site must occur between bases that pair to consecutive seed nucleotides 2-7, and a site creating a bulge must involve a 7mer match that skips one of the seed nucleotides 2-7. Results for the canonical 6mer site (Fig. 2-2D) are included for comparison. Broken lines indicate 5% lower confidence limit (z-test). (B) Weak signal above background for a class of imperfect sites found to have a significantly positive signal-to-background ratio. Results for the canonical 6mer site (Fig. 2-2F) are included for comparison. Broken lines indicate 5% lower confidence limit (z-test).

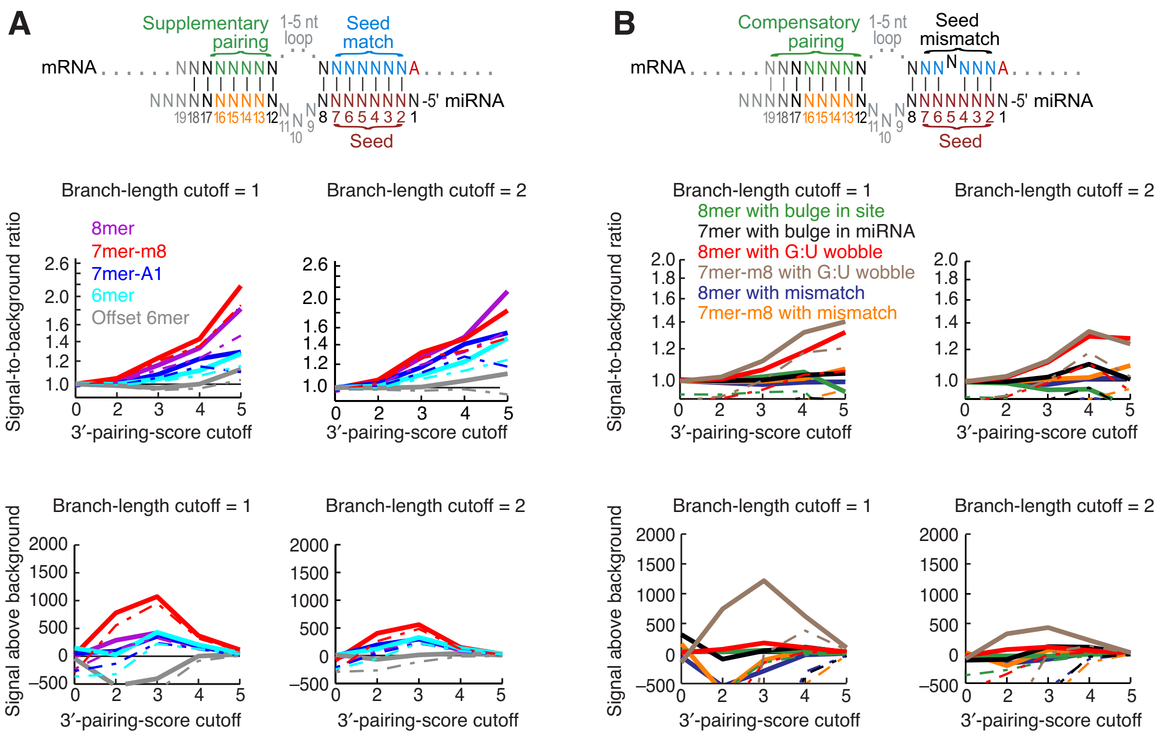


Figure 2-4: Conserved pairing to the 3′ ends of miRNAs. (A) Preferential occurrence of pairing to the 3′ region of the miRNA, which can supplement canonical sites (diagrammed at top). Signal-to-background ratio (top two graphs) and signal above background (bottom two graphs) are plotted at conservation branch-length cutoffs of 1.0 (left) and 2.0 (right). Five percent confidence limits are shown as dashed lines. (B) Preferential occurrence of pairing to the 3′ region of the miRNA, which can compensate for mismatches or bulges in seed pairing (diagrammed at top). As in A, except that seed matches were replaced as indicated with sites containing a mismatched, G:U wobble, or bulged nucleotide.

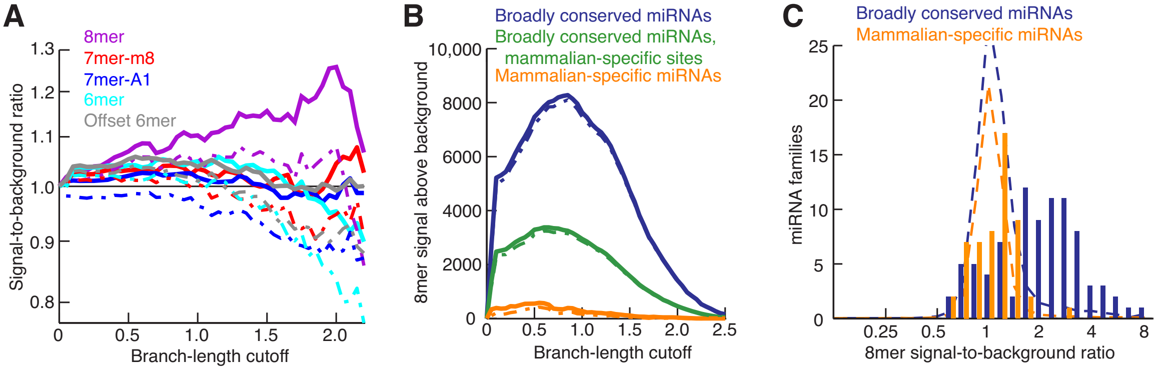


Figure 2-5: Conservation of sites matching mammalian-specific miRNAs. (A) Signal-to-background ratio for sites matching 53 mammalian-specific miRNA families in 18 placental mammals, otherwise as in Figure 2-2D. (B) Signal above background for 8mer sites matching either broadly conserved or mammalian-specific miRNAs in 18 placental mammals, otherwise as in Figure 2-2F. Analysis of 8mer sites matching broadly conserved miRNAs considers either all sites (blue) or excludes those sites conserved beyond placental mammals (green). (C) Signal-to-background ratios for 8mer sites matching individual miRNAs in orthologous 3′ UTRs of placental mammals at optimal sensitivity (branch-length cutoff of 0.85). For the broadly conserved miRNA set, conservation signal excludes sites conserved beyond placental mammals. Distributions expected if miRNA targeting conferred no preferential conservation were estimated using the average signal-to-background ratio of 8mer controls selected for each site, considering GC content and dinucleotide-based conservation (broken lines). Expectations differed between the two sets because of different miRNA numbers and different dinucleotide compositions.

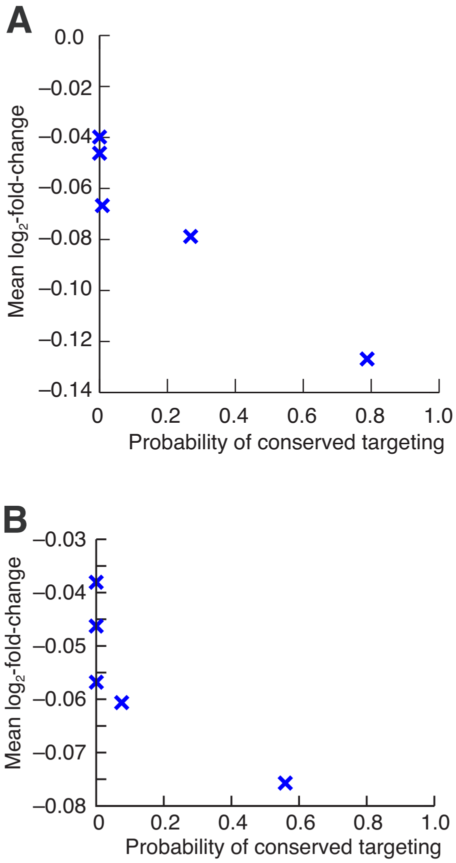


Figure 2-6: Correlation of P_{CT} with mRNA destabilization. (A) Destabilization of human messages with exactly one 7mer-m8 3′ UTR site to a transfected miRNA (Grimson et al. 2007). Messages were grouped into six equal bins based on the site P_{CT}. (B) Destabilization of human messages with exactly one 7mer-m8 3′ UTR site to a transfected miRNA (Grimson et al. 2007), considering only those sites that were not conserved in either mouse, rat, or dog.

2.7 Tables

Table 2.1: Conserved sites with imperfect seed pairing and high $3'$ -pairing scores. Listed are all 7mer and 8mer G:U wobble sites with $3'$ -pairing score $\geq$ 6.0 and branch length $\geq$ 1.0. Also listed are all 7mer and 8mer mismatched sites that meet the above criteria and have human 3'-pairing score $\geq$ 7.0.

$3'$ -Pairing score	Branch length	miRNA	Refseq ID	Gene Name	Seed type
9.0	1.85	miR-196a	NM_024016	HOXB8	8mer GU wobble
7.5	1.2	miR-145	NM_030809	FAM130A1	8mer Mismatch
7.0	3.25	miR-365	NM_002398	MEIS1	8mer Mismatch
7.0	2.6	miR-519d	NM_153020	RBM24	8mer Mismatch
7.0	1.85	miR-153	NM_032521	PARD6B	7mer Mismatch
7.0	1.6	miR-590-5p	NM_033656	BRWD1	7mer Mismatch
7.0	1.35	miR-29b	NM_024834	C10orf119	7mer Mismatch
7.0	1.15	miR-222	NM_002855	PVRL1	8mer Mismatch
6.5	1.7	miR-19b	NM_017637	BNC2	8mer GU wobble
6.5	1.5	miR-613	NM_014903	NAV3	7mer GU wobble
6.5	1.35	miR-191	NM_134265	WSB1	7mer GU wobble
6.5	1.3	miR-15b	NM_001039590	USP9X	7mer GU wobble
6.5	1.15	miR-145	NM_001039457	ATP6V0B	7mer GU wobble
6.0	3.5	miR-301a	NM_022893	BCL11A	7mer GU wobble
6.0	3.45	miR-196a	NM_022658	HOXC8	8mer GU wobble
6.0	3.4	miR-19a	NM_016396	CTDSPL2	7mer GU wobble
6.0	2.95	miR-20a	NM_015215	CAMTA1	8mer GU wobble
6.0	2.55	miR-130a	NM_004721	MAP3K13	7mer GU wobble
6.0	2.15	miR-106b	NM_020814	MARCH4	7mer GU wobble
6.0	2.15	miR-424	NM_001418	EIF4G2	8mer GU wobble
6.0	2.1	miR-302d	NM_002024	FMR1	7mer GU wobble
6.0	1.8	miR-520e	NM_014494	TNRC6A	7mer GU wobble
6.0	1.75	miR-190	NM_001003652	SMAD2	8mer GU wobble
6.0	1.55	miR-424	NM_007374	SIX6	7mer GU wobble
6.0	1.3	miR-130a	NM_012308	FBXL11	7mer GU wobble
6.0	1.25	miR-519d	NM_005808	CTDSPL	8mer GU wobble
6.0	1.2	miR-190	NM_201572	CACNB2	8mer GU wobble
6.0	1.2	miR-519d	NM_001012393	OPCML	7mer GU wobble
6.0	1.15	miR-15b	NM_152277	UBTD2	7mer GU wobble
6.0	1.1	miR-129-5p	NM_020801	ARRDC3	7mer GU wobble
6.0	1.1	miR-33a	NM_178826	TMEM16D	7mer GU wobble
6.0	1.05	miR-302c	NM_015215	CAMTA1	7mer GU wobble

Chapter 3

The evolution of microRNA targeting

This chapter was made possible by Calvin Jan, who generously shared his new annotations of *C. elegans* 3′ UTRs and contributed ideas to the analysis.

Chapter 3

The evolution of microRNA targeting

3.1 Introduction

In recent years, the principles of miRNA targeting have become increasingly clear. The importance of seed matches, their relative efficacy, and the extent of pairing to the $3'$ end of the miRNA have all been verified by a number of methods. Conservation analysis, global analysis of gene expression changes following miRNA or siRNA transfection, and knockdown of individual miRNAs are among the most prominent (Lewis et al., 2005; Grimson et al., 2007; Nielsen et al., 2007; Baek et al., 2008; Selbach et al., 2008). The majority of this information comes from experiments in mammals, with a minority in *Drosophila* (Bartel, 2009). Despite the crucial role that *C. elegans* genetics has had in elucidating miRNAs and their functions (Lee et al., 1993; Wightman et al., 1993; Reinhart et al., 2000; Pasquinelli et al., 2000; Lagos-Quintana et al., 2001; Lau et al., 2001), there is little known about the role of seed match types, target pairing rules, or the relative efficacy of miRNA targets in *C. elegans*. This is in part because the lack of cell lines and *in vitro* extract systems make the necessary biochemistry more difficult in nematodes compared to other model organisms.

Conservation is an attractive approach for assessing similarities and differences in miRNA targeting rules between clades, especially in model systems lacking other experimental tools. This has been difficult in *C. elegans* until now because of the extremely poor annotations of $3'$ UTRs. However, a recent technique for the genome-wide experimental determination of $3'$ UTRs has recently been applied to *C. elegans*, yielding nearly ten thousand high-quality 3′ UTR annotations (C. Jan, in preparation). Interestingly, *C. elegans* 3′ UTRs are far shorter than those in mammals or insects, more closely resembling the UTRs of *S. cerevisiae* than of other metazoans. Given the widespread impact of miRNAs on 3′ UTR evolution in mammals and fruitflies (Farh et al., 2005; Stark et al., 2005), this raised the obvious question of whether the mean 3′ UTR length in a species correlated with any miRNA targeting phenomena. Intriguingly, proliferating mammalian cells express messenger RNAs with shorter 3′ UTRs on average, which can play a role in oncogenesis (Sandberg et al., 2008; Mayr and Bartel, 2009). We set out to exploit the wide variation in 3′ UTR lengths between species to learn underlying principles of the co-evolution of 3′ UTR length and miRNA targeting, which might apply also to cell types in the same species expressing messages with different 3′ UTR lengths.

Having recently determined the extent of conserved miRNA targeting in vertebrates (Friedman et al., 2009), here we apply similar methods to verify the importance of seed matches in *C. elegans*. We discover new seed match types that are preferentially conserved in *C. elegans* but not in vertebrates or flies. Despite several genetically-identified 3′ compensatory targets in nematodes (Reinhart et al., 2000; Abrahante et al., 2003; Vella et al., 2004), we find that these form an extremely small class of preferentially conserved nematode miRNA targets. We find genome-wide evidence for nearly a thousand lineage-specific targets for *C. elegans* that are not conserved in other nematodes. Finally, we discover that miRNA target efficacy and density vary as a function of 3′ UTR length and present a model for the co-evolution of these parameters.

3.2 Results

3.2.1 Widespread conservation of miRNA seed matches in nematodes

Having determined the set of *C. elegans* $3'$ UTRs (C. Jan, personal communication), we investigated the impact of miRNA targeting on the evolution of those $3'$ UTRs. We applied an algorithm for detecting miRNA seed match conservation described previously (Friedman et al., 2009, chapter 2) to multiple alignments of six nematode genomes (Figure 3-1A). Although there are far fewer nematode genomes available than vertebrate genomes, the method was not strongly sensitive to the number of genomes, and the evolutionary time covered by the phylogeny was comparable to that for the vertebrates. Briefly, we quantified the extent to which any k-mer was conserved using a branch-length score over phylogenies controlled for local conservation rates. Because a sequence can be conserved for many reasons other than microRNA targeting, we interpreted the conservation scores by comparing them to background conservation estimated from cohorts of control k-mers, selected for similar expected conservation based on their dinucleotide content. Thus, after controlling for local conservation rates, sequence composition, seed match type, and phylogenetic structure, any difference between the conservation of a miRNA seed match and its background conservation can be attributed to selective maintenance of miRNA targeting by natural selection.

We applied this method to sequences complementary to the 58 *C. elegans* miRNA families that are conserved to *C. briggsae* (Supplemental table 3.3). Examining the conservation above background of hexamers complementary to any region of the miRNAs revealed strong, statistically significant, and specific conservation for sequences matching the seed of the miRNA, nucleotides 2 through 7 (Supplemental figure 3-4). We noted that matches to position 1 through 6 of the miRNA were also significantly conserved above background, even when excluding matches to position 7 (full seed matches). In contrast, matches to positions 1 to 6 are not conserved above background levels in vertebrates (Friedman et al., 2009, Appendix A). As a result, we investigated the role of the nucleotide opposite position 1 of the miRNA. In vertebrates, miRNA seed matches are more often conserved and confer more transcript downregulation if they have an adenosine in this position than if they have a Watson-Crick match (Lewis et al., 2005; Nielsen et al., 2007; Grimson et al., 2007). Surprisingly, we found that in worms matches to positions 2-8 were nearly equally conserved when flanked by an adenosine or a uridine, whereas we observed a strong preference for an adenosine in vertebrates (Figure 3-1C). Interestingly, recent expression profiling in miR-124 knockout worms showed a targeting preference for uridines opposite position 1 (Clark et al., 2010). In contrast to the slight preference for adenosines and uridines flanking matches to positions 2-8, matches to positions 2-7 or 2-6 only showed a clear preference for an adenosine opposite position 1 (Supplemental figure 3-5A). Because most *C. elegans* miRNAs have a 5′ uridine, the preference for adenosines flanking 2-6 or 2-7 matches could be due to Watson-Crick interactions. However, when examining the set of seven conserved miRNAs that do not have a 5′ uridine, we observed a statistically significant preference for an adenosine rather than a Watson-Crick match opposite position 1 of the miRNA (Supplemental figure 3-5B).

In total, we found evidence for substantial and statistically significant conservation above background of six seed match types (Figure 3-1B), including two that have not been observed in genome-wide analysis of vertebrate targeting: the 8mer-U1 and the 6mer-A1. We observed statistically significant conservation for hexamer matches to nucleotides 3-8 or 4-9 as well, although the signal-to-background ratios for these shifted hexamer seed matches were low enough to have little use for target prediction (Supplemental figure 3-4). Because the 8mer-U1 and 6mer-A1 have no genome-wide experimental support for efficacy, we tested these new seed match types using two sources of data: an alg-1 cross-linking immunoprecipitation (CLIP) experiment, representing the genome-wide targeting preferences of miRNAs in *C. elegans* (Zisoulis et al., 2010), and a miR-124 knockout experiment (Clark et al., 2010). We found significant enrichment of known seed match types as well as of 8mer-U1 seed matches in alg-1 CLIP tag clusters, but not of 7mer-U1 or 6mer-U1 matches (Table 3.1).

The 6mer-A1 matches failed to achieve statistical significance, but were still clearly enriched more than matches to nucleotides 2-5 flanked by other nucleotides opposite position 1 of the microRNA. In miR-124 knockout cells, mRNAs containing any seed match type including the 8mer-U1 and 6mer-A1 were significantly de-repressed (Figure 3-1D). Therefore, we performed further analyses using the set of seven seed match types with strong evidence for preferential conservation as well as experimental evidence for in vivo targeting (Figure 3-1B).

We next addressed the relative importance of the seed match types in nematode miRNA targeting. The signal-to-background ratios, or fold-enrichment of conservation, of the seed matches form a natural hierarchy at moderate branch-length cutoffs, with 8mers conserved at a higher rate than 7mers, which were conserved more than 6mers (Figure 3-2A). In vertebrates, a similar ranking was observed, corresponding precisely to the experimentally measured efficacy of the seed match types in human cells (Friedman et al., 2009, Figure 2-2E). Therefore, we predict that the hierarchy of seed match type efficacy is qualitatively similar between nematodes and vertebrates.

With the seed match types confidently annotated, we examined the scope of selectively maintained miRNA targeting. Each seed match type has over 600 sites confidently conserved above background levels (Figure 3-2B). Combining the signal and background from each seed match type at a sensitive branch-length cutoff of 0.5, we find 8,993 ± 278 more seed matches conserved than background controls, representing our estimate for the number of selectively maintained target sites. This corresponds to an average of 0.55 ± 0.02 target sites per *C. elegans* UTR. In order to estimate the number of genes targeted by selectively maintained seed matches, for each seed match type at each UTR conservation level, we randomly selected conserved sites totaling the signal above background and asked how many genes were represented by the target sites. This yielded 4,934 ± 670 genes with conserved seed matches, or 29.9% ± 4.1% of the dataset. While similar methods had found 57.8% ± 3.0% human genes targeted, we can predict that the estimate for the number of nematode genes targeted will rise slightly when more genomes become available for comparison.

3.2.2 Detectable but weak conservation of pairing to the 3′ end of microRNAs

In light of strong genetic evidence for some miRNA targets with mismatches or bulges disrupting seed pairing in *C. elegans*, compensated by strong pairing to the 3′ end of the miRNA (Reinhart et al., 2000; Vella et al., 2004), it is possible that these target types are both broadly effective and well conserved in nematodes. Therefore we next searched for preferential conservation of imperfect seed matches. As is the case in vertebrates (Friedman et al., 2009), there were small amounts of conservation for seed matches with G:U wobbles or bulges in the target, but less than 400 of these imperfect seed match sites are conserved above background levels(Figure 3-6). There was no significant conservation above background for sites with other mismatches or bulges on the miRNA side of the duplex (data not shown). These analyses only required that the position of the bulged or mismatched nucleotide be conserved, not that the nucleotide itself be conserved.

In vertebrates, pairing to the 3′ end of the miRNA has been shown to supplement seed matches or compensate for imperfect seed matches, increasing the efficacy of miRNA targets (Grimson et al., 2007). We next examined whether this could also be a targeting determinant in nematodes. Querying the flanking sequences of conserved seed matches using a model for conserved 3′ pairing previously developed (Friedman et al., 2009), we see statistically significant conservation totaling over 50 sites at a 95% confidence level (Supplemental figure 3-7). This supplementary targeting remains an extremely small class compared to the number of seed match target sites. Imperfect seed matches are also flanked by a small amount of conserved 3′ pairing, although there was no pairing cutoff or conservation cutoff that yielded more than 15 sites of compensatory pairing conserved above background at 95% confidence (Supplemental figure 3-8).

Even if there are a small number of 3′ compensatory targets in nematodes, perhaps the most significant members of the class are both conserved above background and biologically relevant. Indeed, we found seven conserved instances with a 3′ pairing score of at least six, compared to zero instances for shuffled control cohorts (Table 3.2). Notably, the top three of these targets are the two let-7 target sites in lin-41 and the let-7 target site in hbl-1, all of which have been genetically identified (Reinhart et al., 2000; Abrahante et al., 2003; Vella et al., 2004). We therefore predict that the remaining four sites could have important regulatory roles as well. In addition, we found seven sites with bulges (whether on the miRNA or target side) with a 3′ pairing score of at least five and a branch length score of at least 1.0, compared with about two expected by chance. Interestingly, we predict that miR-45 targets lin-14 at the exact 3′ end of its 3′ UTR via a compensatory site. Therefore we conclude that strong 3′ compensatory sites form a tiny but interesting class of conserved miRNA targets in *C. elegans*.

3.2.3 Short 3′ UTRs are associated with a higher seed match density and greater relative usage of weak seed matches

Because 3′ UTRs contain important regulatory information both in nematodes and vertebrates, one expects differences in UTR length to have substantial functional and evolutionary consequences. *C. elegans* UTRs have a median length of 115 nucleotides, compared to 224 in *D. melanogaster* and 733 in *H. sapiens*. We therefore used miRNA seed matches as a case study for the functional impact of varying 3′ UTR length. Although *C. elegans* 3′ UTRs contain fewer selectively conserved miRNA seed matches than human 3′ UTRs (149 compared with 534 per conserved miRNA family), the density is far higher in *C. elegans* (50.8 compared with 26.2 per conserved miRNA family per megabase 3′ UTR, Figure 3-3A, top). Given that conserved miRNA targets are rare at the extreme ends of human 3′ UTRs, one possible explanation is that this space is made available for miRNA targeting in *C. elegans*. However, there is no increase in seed match density in the first 15 nucleotides of the UTR, known as the "ribosomal shadow" (Supplemental figure 3-9A), or after the polyadenylation signal (Supplemental figure 3-9B). Because the polyadenylation signal occurs closer to the 3′ end of *C. elegans* UTRs than human UTRs, there is an increase of effective UTR of about 5 nucleotides per gene, but only enough to explain a 3% increase in target density. Thus, we conclude that the increase in conserved seed match density is due to selective pressure to maintain miRNA targeting despite the small 3′ UTRs.

Because our miRNA target prediction approach can only find substantially conserved targets, we also considered the evolutionary footprint of non-conserved seed matches. We observed that miRNA seed matches, particularly 8mers, occurred at a higher density than expected by chance in *C. elegans* 3′ UTRs but not human 3′ UTRs (Figure 3-3A, bottom). This increase was specific to 3′ UTRs, and was highly significant relative to a first-order Markov model or to control 8mers with similar expected occurrence rates (Supplemental figure 3-10). The high rate of occurrence of miRNA seed matches in *C. elegans* can be explained by decreased mutation away from seed matches. In humans, both conserved and non-conserved miRNA seed matches have a widespread impact on mRNAs and protein levels (Farh et al., 2005; Baek et al., 2008). As a result, there is a strong evolutionary pressure for many 3′ UTRs to mutate seed matches, creating “anti-targets”. The selective pressure to maintain beneficial seed matches is balanced in humans by the selective pressure to avoid detrimental seed matches, resulting in no net enrichment of miRNA seed matches in 3′ UTRs (Figure 3-3B, 3-10). In *C. elegans*, beneficial seed matches still occur at a high density in a high percentage of genes. However, detrimental seed matches are expected to occur at a lower rate, simply because the 3′ UTRs are shorter. Hence, the evolutionary impact of anti-targets is smaller and the resulting selective pressure in balance is to maintain seed matches, leading to their overall enrichment (Figure 3-3B). If this model is true, then miRNA seed match enrichment should be a general property of short 3′ UTRs in any context, assuming comparable numbers of selectively maintained miRNA targets. We tested this hypothesis by assembling a set of validated Refseq annotations and conserved miRNA families for various vertebrate species (Supplemental table 3.4). Using this dataset, we found that the enrichment of 8mer seed matches correlated with the mean 3′ UTR length of the species (Figure 3-3C). The property of enrichment should also occur for short 3′ UTRs within a species, again assuming comparable numbers of selectively maintained seed matches in short and long $3'$ UTRs. Indeed, binning *C. elegans*, human, or *D. melanogaster* $3'$ UTRs by their length, we consistently observe the greatest enrichment of miRNA seed matches in the shortest UTRs (Figure 3-3D).

There are 1,155 occurrences of 8mer-A1 seed matches in excess of those expected to occur by chance in *C. elegans* $3'$ UTRs, presumably due to preferential maintenance of conserved beneficial miRNA targets. However, we found evidence for only 784 8mer seed matches conserved above background levels by our metrics (Figure 3-2B). This left a balance of 371 seed matches with evidence of selective maintenance but without evidence for conservation in the whole-genome alignments. Noting that the closest relatives of *C. elegans* in our phylogeny are separated by an evolutionary distance comparable to the distance between human and mouse, it is likely that many such targets are conserved between *C. elegans* and closely related nematodes that are not represented in the phylogeny. We therefore predict that at least 371 seed matches are preferentially conserved in a lineage-specific manner, i.e. within the *C. elegans* lineage as represented in our phylogeny.

Despite the high density of seed matches within *C. elegans* $3'$ UTRs, the total number of seed matches is still roughly four-fold less than in vertebrates. Because miRNAs are expressed in nematodes at a similar copy number per cell as in humans (Lim et al., 2003), this signifies a greater number of miRNA molecules per seed match in nematode cells compared to human cells. As a result, one might speculate that the miRNA silencing complex begins to saturate the strongest mRNA binding sites, freeing more of the silencing complex to target weaker binding sites. Under this model, when UTRs are shorter, the differences in efficacy between seed match types become smaller, because weaker seed match types are targeted more often. Indeed, we observe that 6mer and 7mer seed matches are relatively more conserved in nematodes than in vertebrates (Figure 3-3E). As a further test, we applied our method for finding conserved seed matches to a set of 16 drosophila species using *D. melanogaster* Refseq $3'$ UTR annotations and 51 conserved drosophila miRNA families. *Drosophila* $3'$ UTRs have a median length of 224, intermediate between human and *C. elegans* $3'$ UTRs (733 and 115, respectively). Likewise, the drosophila 7mer and 6mer seed match types have more preferential conservation than in vertebrates, but less than in nematodes. Thus, we propose that increased relative efficacy of marginal seed match types is a general property of cells containing short 3′ UTRs.

3.3 Discussion

Taking advantage of experimentally determined 3′ UTRs, we used a conservation approach to examine the specificity of miRNA targeting in *C. elegans*. We found that the rules for miRNA targeting elucidated in vertebrates generally hold true in nematodes, including the hierarchy in which 8mer seed matches are more conserved than 7mers or 6mers. However, there were some notable differences in specificity, including strong conservation for two new seed match types, the 8mer-U1 and the 6mer-A1. The *in vivo* efficacy of these new seed matches was verified by derepression of targets in a miRNA knockout context (Clark et al., 2010) and by enriched binding in a cross-linking experiment (Zisoulis et al., 2010). We also show that despite their prominance in the *C. elegans* miRNA targeting literature, 3′ compensatory targets represent a tiny portion of all conserved nematode miRNA targeting. We find evidence for widespread conserved miRNA targeting in nematodes, with nearly 150 seed matches conserved above background per miRNA family, and over 40% of *C. elegans* mRNAs as conserved targets of miRNAs. In addition, we find the first genome-wide evidence for selective maintenance of lineage-specific targeting, totalling over 350 8mer seed matches in *C. elegans*. Presumably, this lineage-specific targeting extends to many more 7mer and 6mer seed matches, despite the fact that we cannot observe their enrichment in 3′ UTRs due to a higher level of neutral occurrences and anti-targeting. Therefore, we expect that lineage-specific selective maintenance of seed matches forms a large class of miRNA targeting in *C. elegans* rivaling the extent of broadly conserved seed matches. It is tempting to speculate that this is true in other lineages, such as primates.

We find two properties that consistently correlate with the mean 3′ UTR length of a species. First, the density of seed matches, whether conserved or non-conserved, is increased in species with short 3′ UTRs. We proposed a model in which UTR shortening contributes to avoidance of deleterious seed matches, as opposed to the more familiar mechanism of avoidance by mutation. This model predicts that seed matches will be enriched in all contexts when 3′ UTRs are short but miRNAs are still relatively highly expressed. Indeed, the property holds when comparing between several species and also when comparing genes with varying UTR lengths within a species. Stark et al. (2005) previously observed that housekeeping genes have short 3′ UTRs and further avoid miRNA seed matches by mutation, presumably due to evolutionary pressure to maintain high levels of expression. In the case of *C. elegans*, the short 3′ UTRs may be due to selection to maintain a small genome size, especially given the extremely small size of intergenic regions. However, this does not affect our model because it predicts enrichment of seed matches as a consequence, not a cause, of short 3′ UTRs. One crucial assumption of this model is that the number of beneficial miRNA seed matches is not linearly proportional to 3′ UTR length. The fact that the 3′ UTR length is correlated with both non-conserved and conserved seed match density (Figure 3-3A) suggests that this assumption is reasonable.

The second property correlated with 3′ UTR length was the conservation signal-to-background ratio of weaker seed matches such as 6mers and 7mers (Figure 3-3E). Given that the efficacy of seed match types correlates well with their conservation above background (Friedman et al., 2009), the higher conservation of marginal seed matches could be explained by an increased number of miRNA molecules relative to the amount of expressed 3′ UTR sequence. Although there are far fewer seed matches in *C. elegans* 3′ UTRs than in human 3′ UTRs, miRNAs are expressed at similar levels between these species (Lim et al., 2003). As a result, miRNAs may begin to saturate the strongest targets (8mer-A1 sites), leaving a higher effective concentration to target 7mers and 6mers. One interesting question is the extent to which this phenomenon causes new seed match types such as the 6mer-A1 to become effective, relative to differences in Argonaute proteins that enable the efficacy of these new target types.

The principles learned here have implications both for the use of *C. elegans* as a model system for miRNA targeting, as well as for the evolution of 3′ UTRs in general. Although miRNA targeting in nematodes largely follows the same rules as in vertebrates, the differences highlighted here should be taken into account when generalizing interactions and mechanisms across species. In contrast, the conservation of seed matches we found in the *Drosophila* clade qualitatively mirrored the results in vertebrates more closely, suggesting that results from fruitfly miRNA experiments may generalize slightly better to mammals. The co-evolution of miRNAs and 3′ UTR length has become more interesting in light of recent evidence that widespread shortening of 3′ UTRs is involved in proliferation and oncogenesis (Sandberg et al., 2008; Mayr and Bartel, 2009). Evolutionary pressures to maintain or avoid miRNA targeting act along a continuum of UTR lengths expressed under various physiological conditions in humans. Our model predicts that under any conditions in which 3′ UTRs are shorter, there may be a higher density of miRNA seed matches and weaker seed matches may be more effective. As cancers decrease their 3′ UTR lengths, they may experience an increase in the efficacy of weak miRNA seed matches, causing widespread misregulation of gene expression. This may be a partial explanation for the observation that overall miRNA expression levels are generally lower in cancers (Lu et al., 2005).

3.4 Methods

3.4.1 Datasets and conservation

3′ UTR annotations were based on a dataset of experimentally determined 3′ ends for *C. elegans* (C. Jan, in preparation), or RefSeq annotations for *D. melanogaster*. 3′ UTR alignments were extracted from Multi-Z alignments, (6-way for nematodes, 15 way for drosophila) from the UCSC genome browser. Conservation analysis was done exactly as in Friedman et al. (2009) (Chapter 2), except that 5 UTR conservation bins were used for *D. melanogaster*, and 4 UTR bins for *C. elegans*. This was to compensate for the smaller total sequence space of 3′ UTRs in these species.

3.4.2 Enrichment of *k*-mers

For each set of seed matches, 1000 cohorts of control k-mers were chosen to match the seed match length, number of G+C nucleotides, and number of CpG dinucleotides. An enrichment is calculated by taking the ratio of the number of seed match occurrences in a given region to the mean number of occurrences for the controls. The p-value is generated by counting the fraction of control cohorts with more extreme observed / expected occurrence ratios (based on a first-order Markov model) than the seed matches.

3.4.3 Microarray analysis

Microarrays following miR-124 knockout (Clark et al., 2010) were analyzed by selecting sets of genes for each seed match. Set consists of genes with exactly one $3'$ UTR seed match of a given type and no other matches of any type. The background was based on genes with no seed matches of any type in their $3'$ UTRs. Only the top 50% of genes in terms of expression are analyzed.

3.5 References

J. E. Abrahante, A. L. Daul, M. Li, M. L. Volk, J. M. Tennessen, E. A. Miller, and A. E. Rougvie. The Caenorhabditis elegans hunchback-like gene lin-57/hbl-1 controls developmental time and is regulated by microRNAs. *Dev Cell*, 4(5):625–37, May 2003.

D. Baek, J. Villén, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, Sep 2008. doi: 10.1038/nature07242.

D. P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2): 215–33, Jan 2009. doi: 10.1016/j.cell.2009.01.002.

A. M. Clark, L. D. Goldstein, M. Tevlin, S. Tavaré, S. Shaham, and E. A. Miska. The microRNA miR-124 controls gene expression in the sensory nervous system of Caenorhabditis elegans. *Nucleic acids research*, Feb 2010. doi: 10.1093/nar/gkq083.

K. K.-H. Farh, A. Grimson, C. Jan, B. P. Lewis, W. K. Johnston, L. P. Lim, C. B. Burge, and D. P. Bartel. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, 310(5755):1817–21, Dec 2005. doi: 10.1126/science.1121158.

R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19(1):92–105, Jan 2009. doi: 10.1101/gr.082701.108.

A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell*, 27(1):91–105, Jul 2007. doi: 10.1016/j.molcel.2007.06.017.

M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 294(5543):853–8, Oct 2001. doi: 10.1126/science.1064921.

N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294 (5543):858–62, Oct 2001. doi: 10.1126/science.1065062.

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

B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, Jan 2005. doi: 10.1016/j.cell.2004.12.035.

L. P. Lim, N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, and D. P. Bartel. The microRNAs of Caenorhabditis elegans. *Genes Dev*, 17(8):991–1008, Apr 2003. doi: 10.1101/gad.1074403.

J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz, and T. R. Golub. MicroRNA expression profiles classify human cancers. *Nature*, 435(7043):834–8, Jun 2005. doi: 10.1038/nature03702.

C. Mayr and D. P. Bartel. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138(4):673–84, Aug 2009. doi: 10.1016/j.cell.2009.06.016.

C. B. Nielsen, N. Shomron, R. Sandberg, E. Hornstein, J. Kitzman, and C. B. Burge. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*, 13(11):1894–910, Nov 2007. doi: 10.1261/rna.768207.

A. E. Pasquinelli, B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Müller, J. Spring, A. Srinivasan, M. Fishman, J. Finnerty, J. Corbo, M. Levine, P. Leahy, E. Davidson, and G. Ruvkun.

Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–9, Nov 2000. doi: 10.1038/35040556.

B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772):901–6, Feb 2000. doi: 10.1038/35002607.

R. Sandberg, J. R. Neilson, A. Sarma, P. A. Sharp, and C. B. Burge. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science*, 320(5883):1643–7, Jun 2008. doi: 10.1126/science.1155390.

M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455 (7209):58–63, Sep 2008. doi: 10.1038/nature07228.

A. Stark, J. Brennecke, N. Bushati, R. B. Russell, and S. M. Cohen. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123(6):1133–46, Dec 2005. doi: 10.1016/j.cell.2005.11.023.

M. C. Vella, E.-Y. Choi, S.-Y. Lin, K. Reinert, and F. J. Slack. The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR. *Genes Dev*, 18(2):132–7, Jan 2004. doi: 10.1101/gad.1165404.

B. Wightman, I. Ha, and G. Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5):855–62, Dec 1993.

D. G. Zisoulis, M. T. Lovci, M. L. Wilbert, K. R. Hutt, T. Y. Liang, A. E. Pasquinelli, and G. W. Yeo. Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. *Nat Struct Mol Biol*, 17(2):173–9, Feb 2010. doi: 10.1038/nsmb.1745.

3.6 Figures

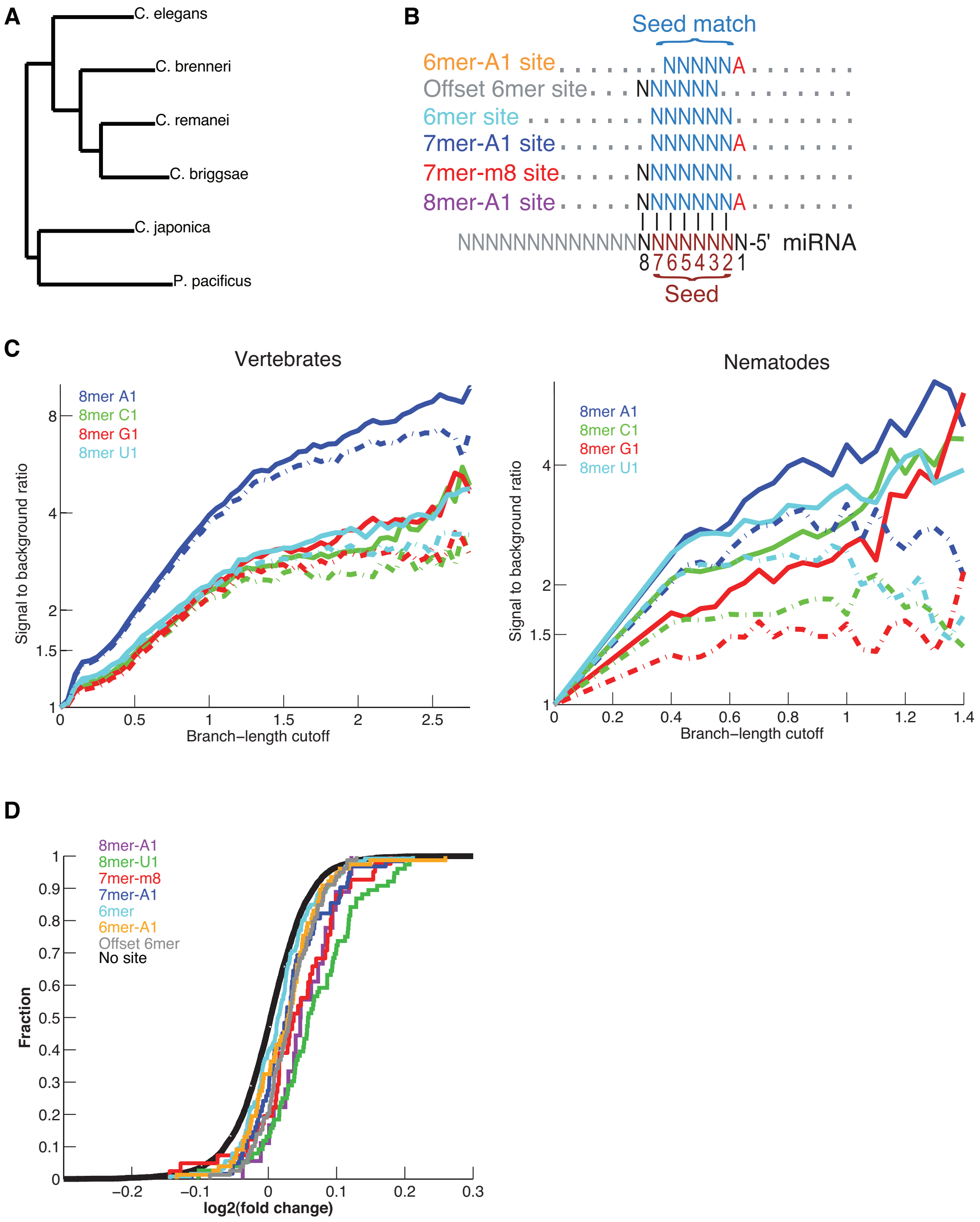


Figure 3-1: A) Phylogeny of six nematode species with sequenced genomes. B) Schematic of selected seed match types with conservation evidence in *C. elegans*. Watson-Crick (WC) matches to the miRNA seed can be flanked by adenosines at position 1 or a WC match at position 8. There are also two off-register match types, the offset 6mer and the 6mer-A1. C) Assessment of "t1A" effect in vertebrates and nematodes. Conservation signal-to-background ratio of matches to nucleotides 2-8 flanked by specific nucleotides opposite position 1 is plotted as a function of conservation stringency. Broken lines indicate 5% lower confidence limit (z-test). D) Cumulative distribution functions for gene expression differences after miR-124 knockout in *C. elegans* (Clark et al. 2010). Each gene in a set has exactly one 3′ UTR seed match of that type and no other matches of any type. All seed match types plotted are statistically significant versus messages with no miRNA seed match (Kolmogorov-Smirnov test).

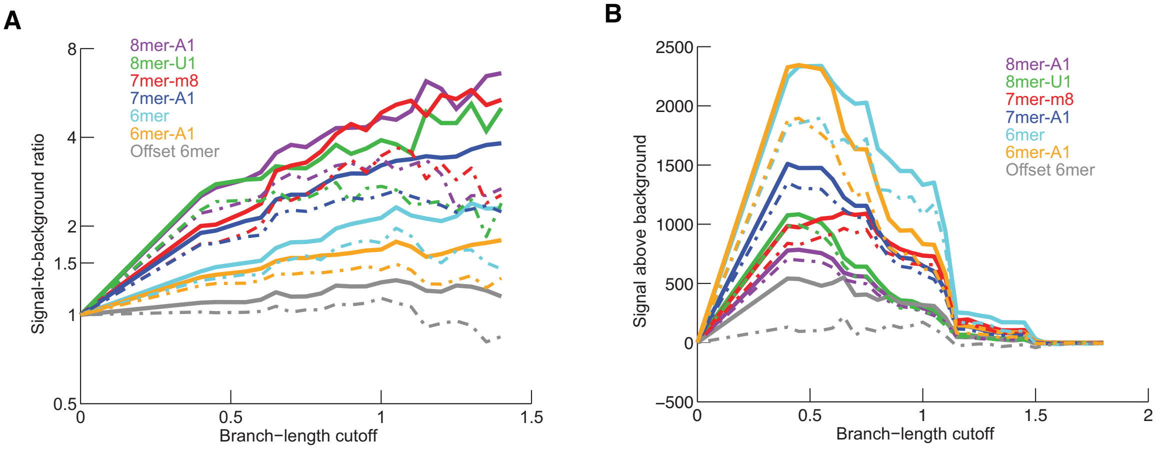


Figure 3-2: Conservation of *C. elegans* seed matches in nematodes. A) Conservation signal-to-background ratio of various seed match types is plotted as a function of conservation stringency (branch-length cutoff). Broken lines indicate 5% lower confidence limit (z-test). B) Conservation signal above background of seed match types is plotted as a function of conservation stringency, otherwise as in part A.

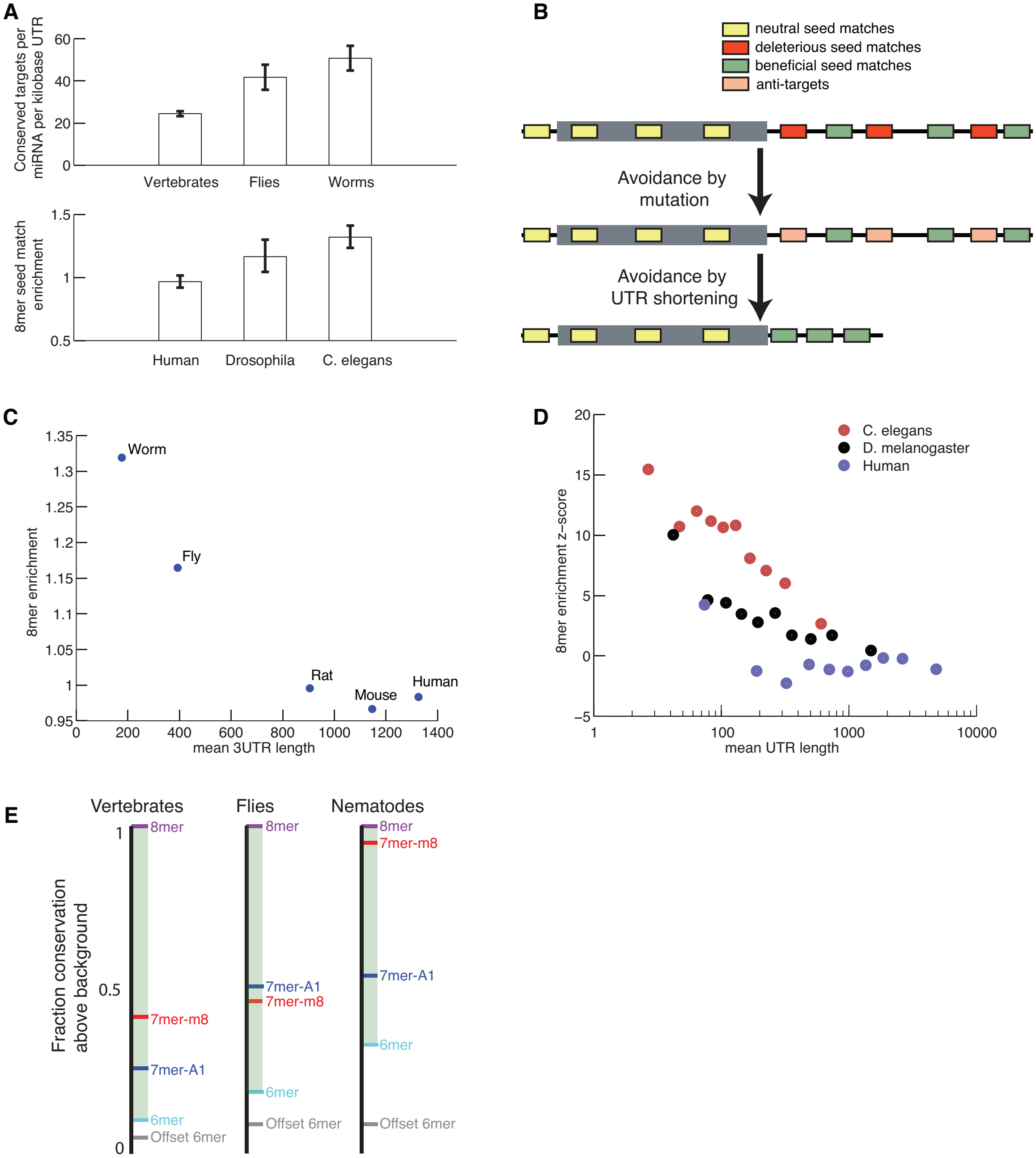


Figure 3-3: Relationship between UTR length and miRNA targeting. A) Density of seed matches in three clades. Top panel: Number of conserved miRNA seed matches above background at a maximally sensitive branch length cutoff per miRNA family per kilobase of 3′ UTR. Error bars represent one standard deviation. Bottom panel: enrichment of seed matches in 3′ UTRs above expectation based on dinucleotide content. Error bars represent one standard deviation. B) Model for relationship between 3′ UTR and miRNA target evolution. At top, seed matches are placed randomly in a gene in the 5′ UTR (left), ORF (grey box), and 3′ UTR (right). Deleterious seed matches can be avoided by either mutating them, leaving a signature of anti-targets (center), or by shortening 3′ UTRs and re-arranging some beneficial seed matches (bottom). C) Enrichment of 8mer-A1 seed matches in 3′ UTRs above expectation based on dinucleotide content for five species with validated UTR annotations. D) Enrichment of 8mer-A1 seed matches in 3′ UTRs within species. For each species, 3′ UTRs are sorted by length into ten bins of equal size, and the z-score for enrichment in that bin is plotted. E) Relative strength of seed match types across clades. Fraction of signal above background (signal-to-background ratio minus 1) for each seed match type is taken at a branch length of 1.0 and normalized to the 8mer-A1. The separation between 6mers and 8mers is highlighted.

3.7 Tables

Table 3.1: Enrichment of potential seed match types in ALG-1 CLIP tags from (Zisoulis et al., 2010). For each set of seed matches, 1000 cohorts of control k-mers were chosen to match the number of G+C nucleotides and the number of CpG dinucleotides. The observed/expected ratio compares the number of seed match occurrences to the mean of the controls, and the p-value represents the fraction of control cohorts with more extreme observed / expected occurrence ratios. Seed match types with conservation support are highlighted in bold.

Potential Seed Match type	Observed / Expected Ratio	P-value
8mer A1	**1.46**	**< 0.001**
8mer C1	1.11	0.16
8mer G1	1.13	0.12
8mer U1	**1.16**	**0.03**
7mer A1	**1.19**	**< 0.001**
7mer C1	1.05	0.30
7mer G1	1.07	0.21
7mer U1	1.05	0.22
6mer A1	**1.07**	**0.07**
6mer C1	0.99	0.57
6mer G1	0.99	0.61
6mer U1	1.03	0.29

Table 3.2: Highly conserved 3′ compensatory targets. Sites with mismatches or bulges in a seed match but meeting conservation cutoffs providing significant signal above background are shown. Sites with any mismatch or bulge type having a 3′ pairing score of at least 6 and a conservation branch length of at least 0.5 are shown. Additionally, sites with a bulge, a 3′ pairing score of at least five, and a branch length score of at least 1.0 are shown.

miRNA	Target mRNA	Mismatch type	Branch length	3′ pairing score
let-7	lin-41	Target bulge	0.9	6.5
let-7	lin-41	7mer G:U wobble	0.9	6.5
let-7	hbl-1	miRNA bulge	0.9	6.5
miR-255	unc-83	Mismatch	1.25	6
miR-356	stn-1	miRNA bulge	1.1	6
miR-75	ZK1127.10	Mismatch	0.9	6
miR-266	ceh-14	G:U wobble	0.5	6
miR-87	pde-4	miRNA bulge	1.25	5.5
miR-1834	ram-2	Target bulge	1.25	5.5
miR-45	lin-14	miRNA bulge	1.1	5.5
miR-90	H28G03.1	miRNA bulge	1.25	5
miR-239b	icd-1	miRNA bulge	1.25	5
miR-228	acn-1	miRNA bulge	1.25	5
miR-72	M02B1.2	miRNA bulge	1.05	5

3.8 Supplemental Figures

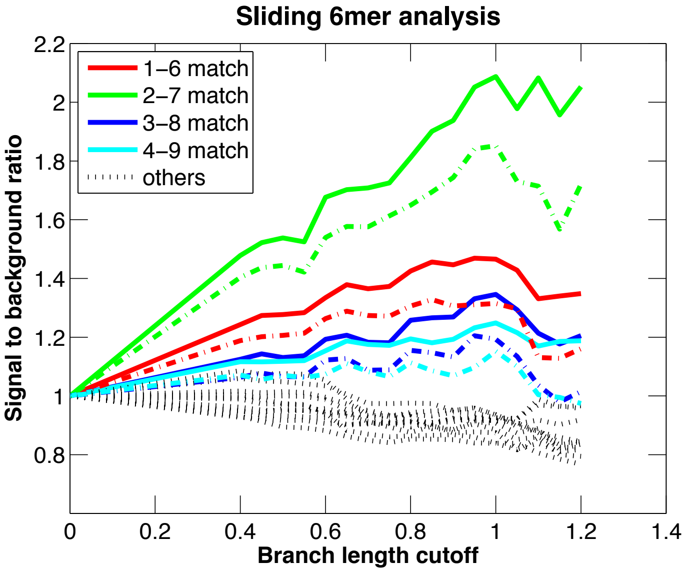


Figure 3-4: Signal-to-background ratios for hexamers complementary to contiguous portions of the microRNA are plotted as a function of branch-length cutoff. All 7mer matches to the miRNA are excluded from this analysis. Colored broken lines indicate 5% confidence lower bounds. Matches to the canonical seed have the highest conservation above background levels, while matches to positions 1-6 and 4-9 surprisingly have statistically significant conservation above background. These site types have signal-to-background ratios rivaling or exceeding that of the 3-8 match, which has experimental support in vertebrates.

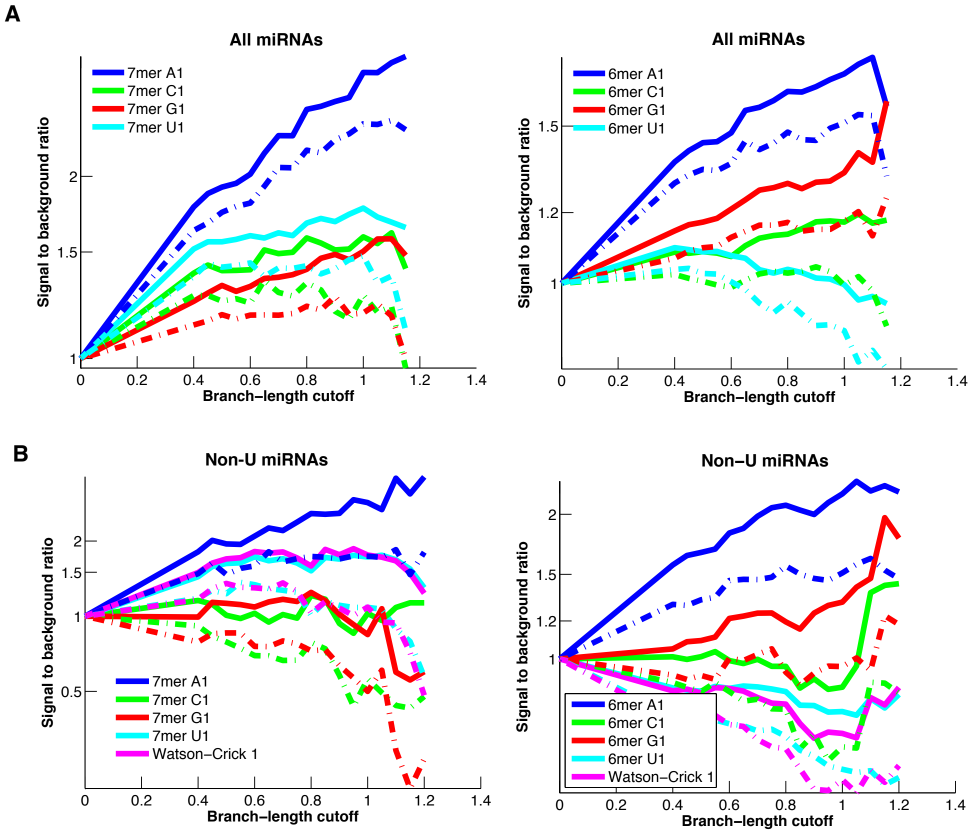


Figure 3-5: A) Signal-to-background ratio for matches to nucleotides 2-7 of the miRNA (left) or 2-6 of the miRNA (right), flanked by specific nucleotides opposite position 1. Broken lines indicate 5% confidence lower bounds. Matches flanked by an adenosine are significantly more conserved in both cases. B) As in part A but only for the seven conserved nematode miRNA families that do not start with a U. The signal-to-background ratio for Watson-Crick matches to position 1 are less conserved than adenosines at position 1 for both 7mers and 6mers.

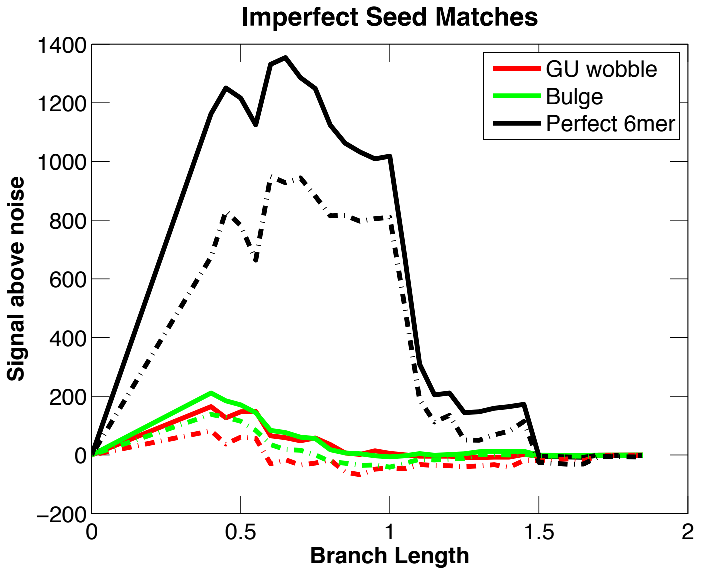


Figure 3-6: Signal above background for 8mer seed matches with a single nucleotide target bulge or a single G:U wobble pairing. Each type has statistically significant conservation above background (broken lines above zero), but a small amount of conservation relative to a perfect 2-7 6mer match, included for comparison.

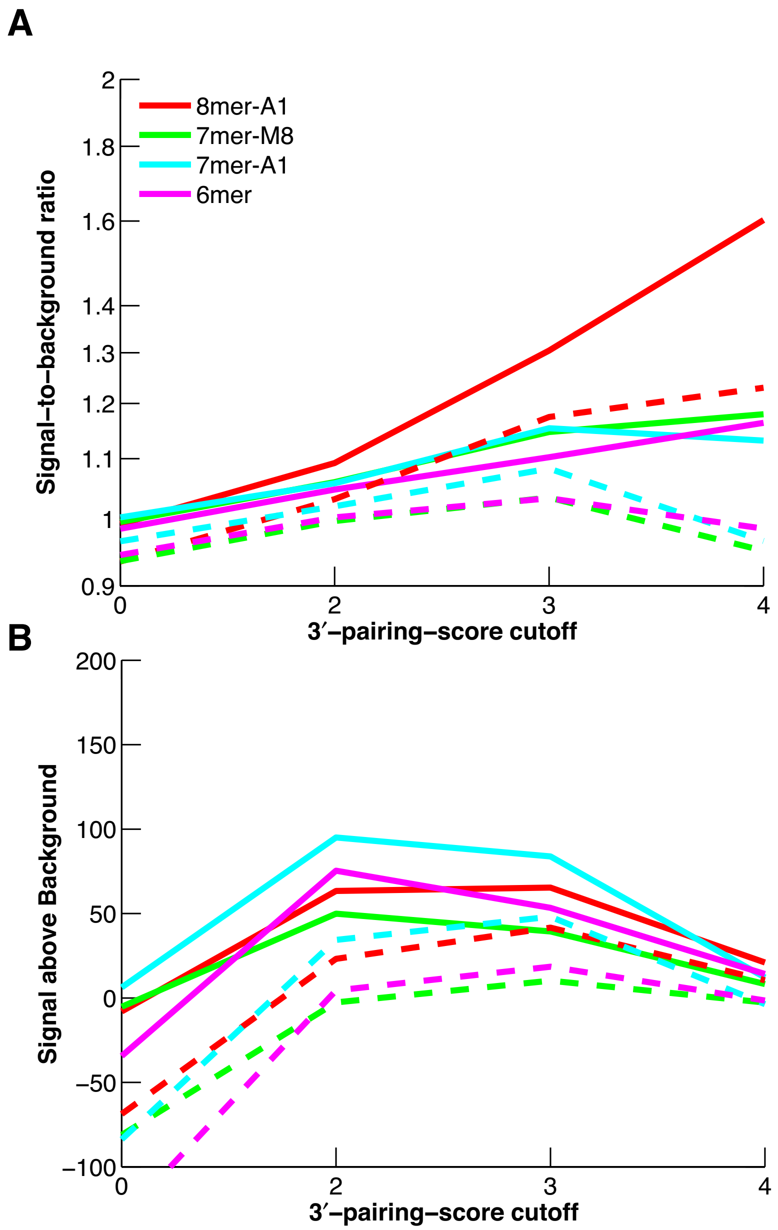


Figure 3-7: Supplemental 3′ pairing conservation. For selected seed match types, signal-to-background ratio (A) or signal above background (B) is plotted versus the 3′ pairing score cutoff. The branch-length cutoff used was 0.5, which yielded the maximum statistically-significant signal above background. There are too few sites to be plotted with a 3′ pairing score of 5 or greater. Broken lines indicate 5% confidence lower bounds.

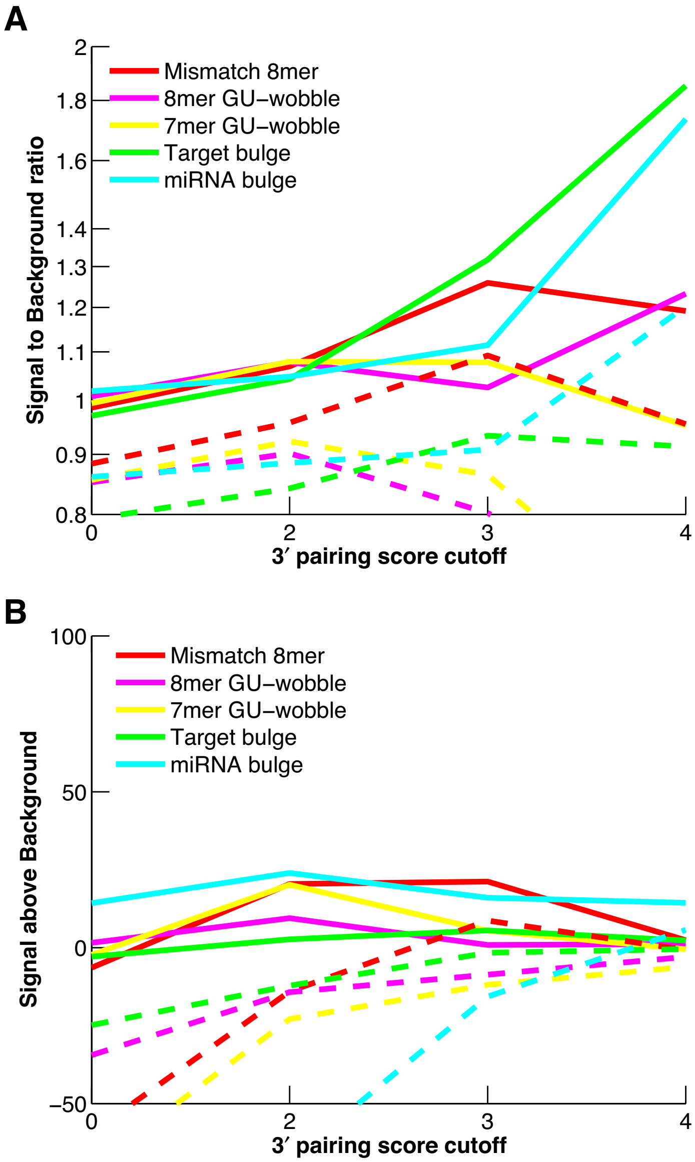


Figure 3-8: Compensatory 3′ pairing conservation. As in figure 3-7, but 3′ pairing flanks 8mer seed matches with a single mismatch, G:U-wobble, bulge on the target or miRNA side, or 7mer-m8 seed match with G:U-wobble pair. The branch-length cutoff with maximal statistically-significant signal above background was 1.05, shown here. Again, there were too few sites meeting a 3′ pairing score cutoff of 5 to plot.

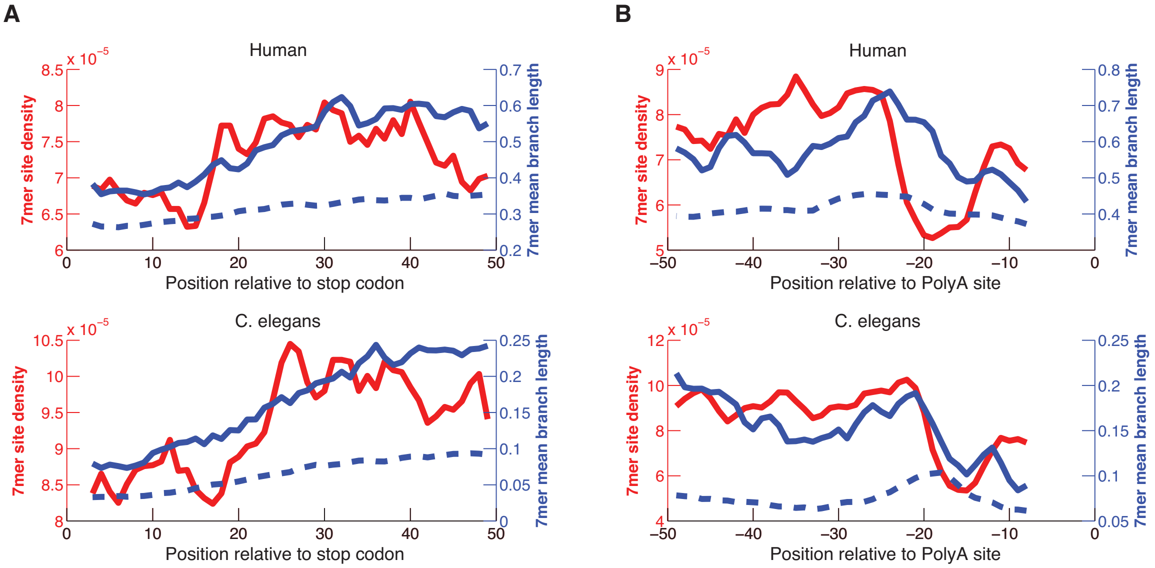


Figure 3-9: Positional density and conservation of miRNA seed matches. A) 7mer-M8 seed match occurrence and conservation near the stop codon. Seed matches per base of sequence per miRNA family are plotted in red (left axis). Mean conservation branch length of seed matches is plotted in blue, with dinucleotide-matched background mean branch length as blue dashed line (right axis). The “ribosome shadow” in the first 15 nucleotides of 3′ UTR affects both seed match density and conservation in both humans and nematodes. B) As in part A, except that position is relative to the site of cleavage and polyadenylation. *C. elegans* have slightly more effective UTR space due to a shorter distance from polyadenylation signal to the site of cleavage.

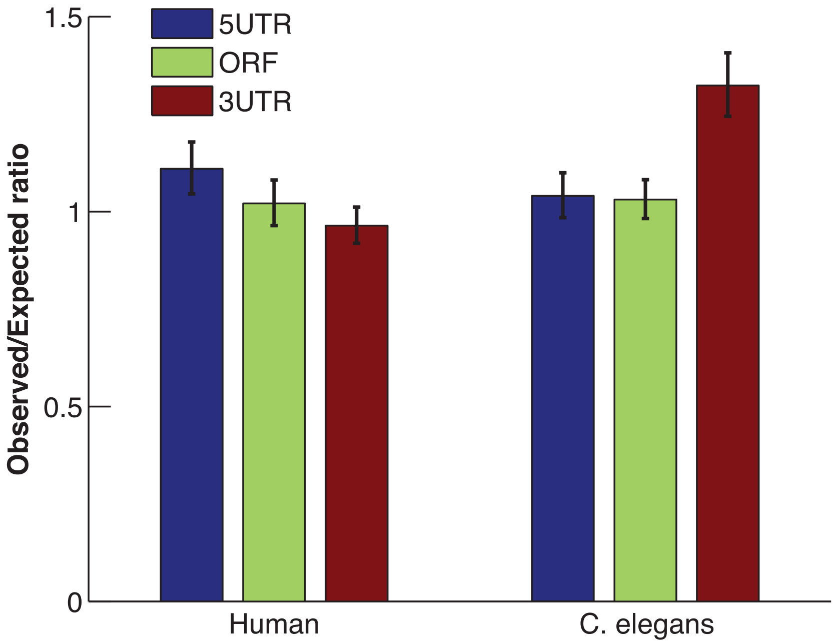


Figure 3-10: Enrichment of 8mer-A1 seed matches in human and *C. elegans* mRNA regions. Background expectation is based on 1000 control cohorts of k-mers with matching G+C content and CpG dinucleotides. Error bars represent one standard deviation. The only region with statistically significant enrichment of seed matches is *C. elegans* $3'$ UTRs.

3.9 Supplemental Tables

Table 3.3: Conserved nematode miRNA families examined in this chapter. The common miRNA seed and a list of *C. elegans* miRNAs comprising the family are shown.

Seed + nt 8	miRNAs in family
AAAUGCA	miR-232
AAAUGCC	miR-357
AACUGAA	miR-255
AAGCUCG	miR-231;miR-787
AAGGCAC	miR-124
AAGUGAA	miR-86;miR-785
AAUACGU	miR-70
AAUACUG	miR-236
AAUCUCA	miR-259
AAUGCCC	miR-786
ACAAAGU	miR-85
ACCCGUA	miR-51;miR-52;miR-53;miR-54;miR-55;miR-56
ACCCUGU	miR-57
ACUGGCC	miR-240
AGCACCA	miR-49;miR-83
AUCACAG	miR-2;miR-43;miR-250;miR-797
AUCAUCG	miR-392
AUGACAC	miR-63;miR-64;miR-65;miR-66;miR-229
AUGGCAC	miR-228
AUUAUGC	miR-60
AUUGCAC	miR-235
CACAACC	miR-67
CACAGGA	miR-249
CACCGGG	miR-35;miR-36;miR-37;miR-38;miR-39;miR-40;miR-41;miR-42
CACUGGU	miR-359
CCCUGAG	lin-4;miR-237
CCCUGCC	miR-789
CCGCUUC	miR-788
CCUUGUU	miR-354
CUUUGGU	miR-244
GAAAGAC	miR-71
GACCGUA	miR-360
GACUAGA	miR-44;miR-45;miR-61;miR-247
GAGAUCA	miR-80;miR-81;miR-82;miR-1834
GAGAUCG	miR-58
GAGGUAG	let-7;miR-48;miR-84;miR-241;miR-795
GAUAUGU	miR-50;miR-62;miR-90
GCAAAUC	miR-254
GCAAGAA	miR-268
GGAAUGU	miR-1;miR-796
GGCAAGA	miR-72;miR-73;miR-74;miR-266
GGCACAA	miR-784
GGCAGUG	miR-34;miR-1824
GUCAUGG	miR-46;miR-47
UAAAGCU	miR-75;miR-79
UAAGUAG	miR-251;miR-252
UACAUGU	miR-246
UAUUAGU	miR-230
UAUUGCU	miR-234
UCAUCAG	miR-77
UGAGCAA	miR-87;miR-233;miR-356
UGCGUAG	miR-242
UUGGCAC	miR-790;miR-791
UUGGUCC	miR-245
UUGUACU	miR-238;miR-239a;miR-239b
UUGUUUU	miR-355
UUUGUAU	lsy-6
ACAGAAG	miR-ruby

Table 3.4: UTR dataset used in figure 3-3C. For *C. elegans*, UTRs were annotated using experimentally-determined 3′ ends (C. Jan, personal communication). For other species, only Refseq annotations with "validated" status were used. MicroRNA families used were as in (Friedman et al., 2009), table 3.3, or in the case of *Drosophila*, all miRNAs with seeds conserved to *D. pseudoobscura*

Species	Number of 3′ UTRs	miRNA families
H. sapiens	18,383	87
M. musculus	10,866	87
R. norvegicus	1,542	87
D. melanogaster	9,259	51
C. elegans	8,143	58

Chapter 4

High-throughput quantitative measurement of the DNA binding specificity of the *GCN4* transcription factor

Robin C. Friedman, Razvan Nutiu, Shujun Luo, Irina Khrebtukova, Gary P. Schroth, Christopher B. Burge

This chapter was a collaboration with Razvan Nutiu in the Burge laboratory and Shujun Luo and Irina Khrebtukova at Illumina corporation. The HiTS-FLIP assay was conceived by R.N. and C.B.B., designed by R.N., R.C.F., S.L., G.P.S, and C.B.B., and performed by S.L. and I.K. Other experiments were performed by R.N., and all analysis was performed by R.C.F.

Chapter 4

High-throughput quantitative measurement of the DNA binding specificity of the *GCN4* transcription factor

4.1 Abstract

In recent years, progress in determining the binding sites and binding affinities of transcription factors has accelerated. High-throughput sequencing and microarray technology have been applied to chromatin immunoprecipitation (ChIP), *in vitro* selection, or direct protein binding (protein binding microarrays, or PBMs) to determine transcription factor binding specificity *in vivo* and *in vitro*. However, no method yet exists that is high-throughput, comprehensive, extremely quantitative, and can measure biophysical parameters of binding directly. Here we present High-Throughput Sequencing - Fluorescent Ligand Interaction Profiling (HiTS-FLIP), a new technique that couples high-throughput sequencing with direct visualization of *in vitro* transcription factor binding. We apply HiTS-FLIP to *S. cerevisiae GCN4*, a transcription factor that acts as a master regulator of the response to amino acid starvation. With a single flow cell and single experiment, we collect over 440,000,000 direct measurements of Gcn4p binding over a range of concentrations. The measured intensities predict regions bound in an *in vivo* ChIP-chip assay better than PBMs or

other models. Using this data, we reconstruct all known Gcn4p binding preferences and describe several new features of Gcn4p binding, including an extended consensus sequence of binding (TATGACTCATA), the global importance of half-site binding for Gcn4p binding, and interdependencies between individual positions of the binding motif. We utilize the concentration-dependent binding curves to estimate dissociation constants for Gcn4p binding to all 10mers. Finally, we show the *in vivo* relevance of these intensities, explaining the activation timing of genes following *GCN4* induction using the expected binding affinity of Gcn4p to promoters.

4.2 Introduction

A key goal of systems biology is to gain a global understanding of the regulation of gene expression by cataloging molecular interactions. Many techniques applied to this end to date have been high-throughput but not quantitative (e.g. yeast two-hybrid, affinity co-purification), or quantitative but not high throughput (e.g. surface plasmon resonance, gel shift assays). In the case of protein-DNA interactions, methods such as ChIP-Chip (Ren et al., 2000) and ChIP-Seq (Johnson et al., 2007) enumerate global *in vivo* interactions, but often reflect the binding of multi-protein complexes rather than the direct binding (Gordân et al., 2009). Additionally, ChIP is subject to biases due to fragmentation, antibody specificity, and differential affinity for protein confirmations that make quantification difficult. Techniques based on *in vitro* selection (Klug and Famulok, 1994) have proven useful for defining consensus binding preferences, and protein binding microarrays (PBMs) (Mukherjee et al., 2004) for enumerating all short sequences bound by a protein. Although semi-quantitative, these methods do not directly observe binding in multiple conditions and therefore do not provide direct measurement of biophysical parameters. The fully quantitative and comprehensive binding preferences of transcription factors would be useful for separating direct from indirect binding (Gordân et al., 2009), for distinguishing context effects such as chromatin structure from binding affinity (Wasson and Hartemink, 2009), and for constraining computational models of transcriptional systems (Endy and Brent, 2001).

Here we address this problem by taking advantage of the ease of generating vast amounts of data using second-generation sequencing technologies such as the Illumina Genome Analyzer. We describe High-Throughput Sequencing - Fluorescent Ligand Interaction Profiling (HiTS-FLIP), a general method for the direct and quantitative measurement of protein-DNA interaction affinities on a scale of hundreds of millions of datapoints in a single experiment. For the proof of principle experiment, we profile the binding of *GCN4*, a yeast transcription factor with a large body of knowledge about its binding preferences built over three decades of research. *GCN4*, a basic leucine zipper protein (bZIP), binds as a dimer to a relatively simple sequence, is itself a classic example of a translationally regulated mRNA, and is a master regulator of the response to amino acid starvation with conserved function in mammalian cells (Hinnebusch, 2005). In a single experiment, we reconstruct all known aspects of *GCN4* binding preferences, learn new subtleties of its specificity, and quantify its binding affinity to hundreds of thousands of sequences.

4.3 Results

4.3.1 HiTS-FLIP: High-Throughput Sequencing Fluorescent Ligand Interaction Profiling

The Illumina Genome Analyzer provides well over a hundred million sequences per run by building clusters of DNA containing the same sequence on a flow cell and sequencing by synthesis *in situ*. Nucleotides tagged with individual fluorophores are added one at a time, and visualized using a charge-coupled device (CCD) camera. The sequences of each cluster of DNA are then assembled *in silico* by matching the fluorescence of each cluster cycle-by-cycle. We reasoned that in analogy to fluorescent nucleotides, we could add fluorescently tagged proteins to the flow cell, visualize their binding over every DNA cluster, and match the bound clusters to their sequences based on their position in the flow cell. Assuming an even coverage over sequence space, one could then observe the unbiased *in vitro* binding preferences of the tagged protein. Our procedure was therefore conceptually simple: 1) Build and sequence over 100 million clusters of random synthetic DNA; 2) Wash away the sequenced second strand and rebuild double-stranded DNA; 3) Flow on various concentrations of fluorescently-tagged proteins; 4) Quantify the binding to each cluster by visualizing fluorescence; and 5) Combine the bound sequences into a comprehensive, quantitative landscape of binding preferences (Figure 4-1).

4.3.2 Gcn4p binds its canonical motif on flow cells *in vitro*

We chose to demonstrate the HiTS-FLIP method using *S. cerevisiae GCN4*, a well-characterized dimeric bZIP transcription factor and a master regulator of yeast amino acid biosynthesis (Hinnebusch, 2005). We expressed and purified the *GCN4* protein, Gcn4p, with an mOrange fluorescent tag and applied the HiTS-FLIP method to a single flow cell of random synthetic 25-mers. We imaged five concentrations of Gcn4p-mOrange: 1, 5, 25, 125, and 625 nM. After matching the fluorescent intensity of Gcn4p-mOrange binding to sequencing clusters and normalizing for cluster size and flow cell position (see supplemental discussion), we were able to quantify the binding of Gcn4p to over 88 million clusters. Selecting the top 0.5% of clusters by raw intensity with Gcn4p at a 125nM concentration, the known binding motif of Gcn4p, TGACTCA (Oliphant et al., 1989), was enriched by 40-fold over its rate of occurrence in all clusters. An unbiased search for enriched 7mers and by MEME (Bailey and Elkan, 1994) both found TGACTCA as the top enriched motif (data not shown).

Because each heptamer occurred roughly 124,000 times on our flow cell, we reasoned that subtle preferences for individual sequences would be quantifiable. The binding intensity of clusters with the Gcn4p binding heptamer did not depend on the location of the heptamer within the sequence, so we considered the binding intensity on all clusters containing a given k-mer as equivalent and independent measurements (see supplemental discussion). Therefore we devised a simple algorithm for quantifying Gcn4p binding preferences for sequences of length k: we searched for the k-mer with the highest mean binding intensity, removed the sequences containing that motif, and repeated to yield a ranked list of k-mers by binding preference. After normalizing for cluster size and background fluorescence (see supplemental discussion), we quantified the binding of the top 10 mutually-exclusive heptamers as well as four randomly-selected control k-mers with no expected binding to Gcn4p (Figure 4-2). For the canonical heptamer, we observed concentration-dependent binding ranging from virtually zero to saturating binding. Other heptamers did not reach saturation, but exhibited clear concentration dependent binding with a consistent hierarchy of binding strength regardless of the concentration. Treating each of seven lanes used in a flow cell as technical replicates, we observed little variability (Figure 4-2, error bars). Thus, we observed reproducible and consistent rankings of all 8,192 heptamers (treating reverse complements as equivalent).

4.3.3 The Gcn4p binding motif is surprisingly complex

Previous studies using ChIP-Chip data or protein binding microarrays (PBMs) have found that the flanking nucleotides of the Gcn4p motif can also affect binding (Hill et al., 1986; Oliphant et al., 1989; Zhu et al., 2009). Given the vast amount of data available, we asked to what extent the sequence flanking the canonical heptamer influenced binding. Because the ranking of k-mers was consistent between concentrations, we considered only binding intensities at a single concentration, 125nM, for this portion of the analysis. We enumerated all single nucleotide extensions of the canonical motif TGACTCA and asked whether any had a statistically different mean binding intensity. Four 8mers, representing single-base extensions of the 7mer, had significantly higher binding intensity. Repeating the procedure with the top scoring 8mers revealed that the near-palindromic 9mer sequence ATGACTCAT had significantly higher binding intensity than all 8mers, confirming previously known *in vivo* preferences (Hill et al., 1986). Little is known about the binding preferences of Gcn4p to sequences longer than 9 nucleotides *in vitro*. We surprisingly found two 10mers, ATGACTCATA and TATGACTCAT, occurring 1462 and 2241 times on our flow cell respectively, having significantly higher binding intensity than the highest-scoring

9mer. Repeating the procedure, we did not find any 11mer motifs with significantly higher binding than either of the top 10mers. This suggests that noise precludes us from determining the binding preferences of all 11mers with statistical confidence. However, we combined the two top 10mers to define a new near-palindromic 11mer binding consensus for Gcn4p, TATGACTCATA.

We next asked whether sequences that are not canonical GCN4 binding motifs could nevertheless be bound *in vitro* and *in vivo*. We compared the mean binding intensity of all 8mers with their expected intensity based on a position weight matrix (PWM) developed from ChIP-Chip binding and evolutionary conservation data (MacIsaac et al., 2006), (Figure 4-3A). The highest ranked 8mer sequences were similar for the PWM and for the HiTS-FLIP binding intensity, but the two methods agreed poorly for hundreds of sequences. The differences between binding on the flow cell and the expectation based on the PWM could be due to differing conditions *in vitro* and *in vivo*. However, sequences that bind surprisingly well in the HiTS-FLIP assay, but are expected to bind poorly based on the PWM (Figure 4-3A, red), are highly enriched in regions recovered in a GCN4 ChIP-Chip experiment following amino acid starvation (Harbison et al., 2004), (Figure 4-3B, red). In contrast, sequences that scored well in the PWM model but had low binding intensity (Figure 4-3A, green) were slightly but not significantly less enriched in ChIP-Chip binding regions (Figure 4-3B, green). Similar results were found when comparing to a PWM based on *in vitro* binding data (Zhu et al., 2009), suggesting that the deficiency was inherent to the PWM representation and not to the specific method used.

To extend these results, we undertook a systematic comparison between multiple predictions for GCN4 binding affinity. We ranked all 8mers by their expected binding strength based on 125nM HiTS-FLIP, the PWM based on ChIP-Chip and conservation data (MacIsaac et al., 2006), a PWM based on *in vitro* binding using a protein binding microarray (PBM) (Zhu et al., 2009), and raw intensities for PBM binding. The top ranked 8mers for all methods were significantly enriched in the ChIP-Chip bound regions (Harbison et al., 2004) (Figure 4-3C). HiTS-FLIP found statistically significant enrichment in the ChIP-Chip binding data for roughly 1,100 8mers, suggesting that our method was sensitive for weak-binding motifs that are, at least in aggregate, relevant *in vivo*. The top 1,000 8mers by HiTS-FLIP had a significantly higher enrichment than those for PBMs, and equaled the enrichment for the PWM, despite the fact that the PWM was trained on this same dataset and the HiTS-FLIP method is completely independent from ChIP-chip.

The fact that both HiTS-FLIP and PBM-based binding intensities for 8mers predicted *in vivo* binding enrichment much better than PWMs, even when the PWM was based on k-mer binding intensities (Zhu et al., 2009), suggested an inherent limitation of the PWM representation for the Gcn4p binding motif, such as interdependency between positions of the motif. We next systematically determined the effect of pairwise mismatches from the consensus 7mer on binding affinity (Figure 4-3D). If each position within the consensus sequence were recognized independently of other positions, then mismatch would reduce the binding affinity by a constant amount regardless of the identity of the rest of the 7mer. In visual terms, each column in figure 4-3D would be uniform, whereas rows would be identical. Instead, each half of the 7mer is nearly independent of the other half, consistent with a model in which a half-site is the crucial unit of recognition. Gcn4p binds as a dimer in which each subunit prefers the sequence ATGAC, introducing inherent asymmetry into the binding motif (Sellers et al., 1990). We confirm that asymmetry with our binding intensities (Supplemental figure 4-6). Our pairwise interdependencies make sense in light of the binding of a dimer to two half-sites. Given that a mismatch already exists in the right half of the sequence, successive mismatches in the right half have a small effect, compared to mismatches in the left half that would destroy an intact half-site. Corroborating the model of half-site binding, we found that cumulative mismatches within either half-site have successively weaker effects (Supplemental figure 4-6). In fact, Hollenbeck and Oakley (2000) previously showed that Gcn4p dimers can bind to half-sites in isolation. Clearly this introduces an inherent dependence between positions in the binding model, since the ATGAC half-sites are nearly independent of each other. Importantly, this effect cannot easily be captured by considering pairwise or tertiary dependencies, since each nucleotide of the ATGAC half-site will be dependent on the others.

The Gcn4p dimer is also flexible enough to accommodate variable spacing of the half sites, binding both ATGACTCAT and ATGACGTCAT (Sellers et al., 1990). HiTS-FLIP confirms these known binding preferences and finds that Gcn4p binds with surprisingly high affinity to a sequence with an extra nucleotide spacer, yielding ATGACNGTCAT (Supplemental figure 4-7). To determine whether this binding is due to simultaneous recognition of a site with a central spacer or two separate half-sites, we compared the affinity of the sequence with the half-sites in the same orientation (precluding simultaneous binding by a dimer) to the sequence with half-sites in reverse-complement orientation (potentially permitting simultaneous binding). In fact, the sequence with the tandem orientation had a significantly higher affinity than the sequence with reverse-complement orientation, suggesting that the binding to ATGACNGTCAT may be due to independent binding to two half-sites (Supplemental figure 4-7).

4.3.4 Direct measurement of hundreds of thousands of equilibrium binding constants

A key advantage of the HiTS-FLIP assay is the ability to adjust conditions on the flow cell and re-image binding. In this work, we demonstrate this ability by altering the concentration of Gcn4p on the flow cell in order to determine equilibrium binding constants. By varying the concentration from 1nM to 625nM, we observed the full range from no appreciable Gcn4p binding to saturation of binding for the canonical 7mer motif, TGACTCA (Figure 4-2). Other 7mers bound more weakly but approached the canonical motif's binding strength at high concentrations. Gcn4p is a coiled-coil bZIP transcription factor, binding as a dimer. Previous assays have modeled Gcn4p binding as a cooperative process with a Hill coefficient of 2 (Hollenbeck and Oakley, 2000). To verify this assumption, we fit a curve to the binding intensities for TGACTCA using the standard Hill equation. The resulting Hill coefficient (h) of 2.1 was quite close to the 2.0 expected for cooperative binding by a dimer. We therefore fixed h to 2.0 and fit curves to data for all 7mers, 8mers, 9mers, and 10mers, calculating K_D values for each. These K_D values will be published online separately. This compendium of dissociation constants represents a quantitative landscape of the thermodynamic equilibrium of Gcn4p binding on an unprecedented scale. Calculating these constants for 10mers, for which we had enough statistical power to observe subtle differences in affinity, we have compiled dissociation constants for Gcn4p binding to 524,800 sequences in a single experiment. We suspect that this represents as many dissociation constants for pairs of protein and DNA sequence as have been previously quantified.

4.3.5 Quantitative binding constants reflect *in vivo* function

While we have shown that the HiTS-FLIP binding intensities reflect *in vivo* binding, binding does not necessarily lead to functional relevance. To verify whether HiTS-FLIP binding intensities are functionally significant, we turned to timecourse microarray assays following induction of transgenic *GCN4* (Chua et al., 2006), and following amino acid starvation, which induces endogenous *GCN4* (Gasch et al., 2000). Most genes that were activated following *GCN4* induction should contain binding sites to Gcn4p in their promoters, and indeed both sets are highly enriched for strong binding sites (Supplemental figure 4-8). However, transcription factor binding is concentration-dependent (Figure 4-2). Therefore one would expect that if the levels of Gcn4p are monotonically increasing over a timecourse, then genes with strong binding sites in their promoters would be activated early, while genes with increasingly weak binding sites would be activated later. In both the case of *GCN4* driven by a *GAL* promoter, and in the more physiological case of amino acid starvation, we see exactly that trend (Figure 4-4). This suggests that differences in the affinity of Gcn4p binding to promoters are not random, but rather reflect a gradient of functionally distinct responses.

4.4 Discussion

In a single flow cell and a single experiment, we have taken roughly 440 million measurements of protein binding to particular DNA sequences (88 million clusters times five concentration points). If one decided to build flow cells with a higher density of clusters and take measurements for more concentrations, we expect that one could easily take more than a billion such measurements in a single experiment. While each data point is fairly noisy, the large amount of data provides enough statistical power to determine the binding preferences of a factor with remarkable accuracy. Using this data, we were able not only to confirm the large body of knowledge about Gcn4p binding preferences (Hill et al., 1986; Oliphant et al., 1989; Sellers et al., 1990; Hollenbeck and Oakley, 2000; Zhu et al., 2009), but to learn several new nuances in Gcn4p specificity. First, we determined global quantitative affinities in the form of equilibrium dissociation constants for all 10mers. Second, we identified with statistical significance a longer Gcn4p optimal binding site, which extends to an 11mer sequence, TATGACTCATA. Third, we underscored the importance of taking into account the underlying biophysics of Gcn4p binding, in which each of the two dimers binds to a half-site with at least moderate affinity by itself.

We observed a correlation between the Gcn4p binding affinities determined by HiTS-FLIP and *in vivo* binding (Figure 4-3), but more importantly, we observed a correlation between binding affinity and the timing of gene activation following the induction of *GCN4* expression (Figure 4-4). This phenomenon is reminiscent of other systems in which the timing of responses is controlled by binding motif strength. For example, during pharynx development in *C. elegans*, *PHA-4* binding sites in the promoter of pharyngeal genes control not only the activation of their expression but also the timing of their activation by the relative affinity of the binding motif (Gaudet and Mango, 2002). This confirms the importance of weak binding motifs in two ways: first, it shows that sequences with low affinity are bound *in vivo* when their regulator is at a high concentration; but it also shows that sequences with low affinity are functionally distinct and therefore may be evolutionarily adaptive.

We expect that HiTS-FLIP will be a broadly useful approach for quantitative profiling of the sequence-specific binding affinity of any DNA binding molecule. This method is complementary to others that are highly quantitative but have lower throughput (Maerkl and Quake, 2007), or are effective but provide less data and are less quantitative (Mukherjee et al., 2004). However, HiTS-FLIP provides several advantages that cannot be matched by other methods, including protein binding microarrays. First, it provides hundreds of millions of data points, more than is possible using any other technology. Second, it allows for easy manipulation of conditions on the flow cell, such as temperature, pH, and salt concentrations. We have varied the simplest parameter, protein concentration, but one could also obtain informative data about protein folding, electrostatics, enthalpy, heat capacity, and other interesting biophysical parameters of binding. Third, it is flexible in terms of the length and sequence of DNA placed on the flow cell, allowing proteins with much more complex motifs to be profiled. Fourth, the ability to measure multiple fluorescent wavelengths allows hetero- and homo-dimerization of protein bound to the flow cell to be measured. Finally, one could perform multiple runs of HiTS-FLIP on a flow cell after a sequencing run, reducing its cost per run. We expect that the availability of quantitative and accurate biophysical catalogs of weak binding affinities will enable new types of modeling for synthetic biology and systems biology applications, and will open the door for a new and more nuanced understanding of the regulation of gene expression.

4.5 Methods

4.5.1 HiTS-FLIP

Flow cells were built using random 25mer oligonucleotides using the single-end adapter kit and sequenced on a Genome Analyzer IIX instrument. DNA was denatured and resynthesized using a fluorescently labeled primer. The *GCN4* fusion with mOrange was cloned and expressed in *E. coli* and purified using nickel columns to a concentration of 3.25 μM. Gcn4p was serially diluted in PBS and sequentially applied at 1, 5, 25, 125, and 625 nM with 300 μg BSA. Before imaging we washed with 600 μg BSA in 2mL PBS. Mapping of protein intensities to clusters was performed using Firecrest, part of the Genome Analyzer software package.

4.5.2 Normalization

Becuase the raw intensity varies from cluster to cluster and based on position within the flow cell during sequencing, we normalize the protein binding intensity to the average intensity of the cluster measured during sequencing cycles (see supplemental discussion). The intensity from the fluorescently-labeled primer is not used to normalize, but rather assists with mapping. The background intensity is calculated as the median normalized intensity over all clusters and is then subtracted from the normalized intensity of each clusters to yield the normalized intensity above background. Finally, we added a correction for photobleaching for concentrations greater than 1nM (see supplemental discussion). This normalized intensity above background is the intensity reported in all figures.

4.5.3 Dissociation constants

Dissociation constants were calculated by least squares fit with the hill coefficient fixed at 2 and the maximum intensity fixed at the median measured intensity for the canonical 9mer motif, using the equation:

$$\phi_B = \frac{[\text{Gcn4p}]^h}{[\text{Gcn4p}]^h + K_D^h}$$

Where ϕ_B is the fraction of DNA bound, [Gcn4p] is the Gcn4p concentration, K_D is the dissociation constant of Gcn4p to the sequence, and h is the Hill coefficient of binding. Constants were determined up to 1 μM due to noise in the estimates after this point.

4.5.4 Enrichment of *k*-mers

Enrichment in ChIP-Chip bound regions (Harbison et al., 2004) or promoters of activated genes (Chua et al., 2006; Gasch et al., 2000) was calculated with a simple counting procedure. The frequency of the occurrence of each k-mer was calculated in all yeast promoters (500 nucleotides upstream of each ORF, or to the next annotated gene if closer than 500 nucleotides upstream). The expected occurrence of a k-mer in the region of interest was then the expected frequency of that k-mer times the number of k-mers in the region. Enrichment was simply the number of occurrences minus the expected number of occurrences. 8mers were scored against the PWM by selecting the highest score over any orientation and register having at least four nucleotides of overlap.

4.6 References

T. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, Jan 1994.

G. Chua, Q. D. Morris, R. Sopko, M. D. Robinson, O. Ryan, E. T. Chan, B. J. Frey, B. J. Andrews, C. Boone, and T. R. Hughes. Identifying transcription factor functions and targets by phenotypic activation. *Proceedings of the National Academy of Sciences of the United States of America*, 103(32):12045–50, Aug 2006. doi: 10.1073/pnas.0605140103.

D. Endy and R. Brent. Modelling cellular behaviour. *Nature*, 409(6818):391–5, Jan 2001. doi: 10.1038/35053181.

A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11(12):4241–57, Dec 2000.

J. Gaudet and S. E. Mango. Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4. *Science*, 295(5556):821–5, Feb 2002. doi: 10.1126/science.1065175.

R. Gordân, A. Hartemink, and M. Bulyk. Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res*, Sep 2009. doi: 10.1101/gr.094144.109.

C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. MacIsaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, Sep 2004. doi: 10.1038/nature02800.

D. E. Hill, I. A. Hope, J. P. Macke, and K. Struhl. Saturation mutagenesis of the yeast his3 regulatory site: requirements for transcriptional induction and for binding by GCN4 activator protein. *Science*, 234(4775):451–7, Oct 1986.

A. G. Hinnebusch. Translational regulation of GCN4 and the general amino acid control of yeast. *Annu Rev Microbiol*, 59:407–50, Jan 2005. doi: 10.1146/annurev.micro.59.031805.133833.

J. J. Hollenbeck and M. G. Oakley. GCN4 binds with high affinity to DNA sequences containing a single consensus half-site. *Biochemistry*, 39(21):6380–9, May 2000.

D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–502, Jun 2007. doi: 10.1126/science.1141319.

S. J. Klug and M. Famulok. All you wanted to know about SELEX. *Mol Biol Rep*, 20(2):97–107, Jan 1994.

K. D. MacIsaac, T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo, and E. Fraenkel. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, 7:113, Jan 2006. doi: 10.1186/1471-2105-7-113.

S. J. Maerkl and S. R. Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–7, Jan 2007. doi: 10.1126/science.1131007.

S. Mukherjee, M. F. Berger, G. Jona, X. S. Wang, D. Muzzey, M. Snyder, R. A. Young, and M. L. Bulyk. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet*, 36(12):1331–9, Dec 2004. doi: 10.1038/ng1473.

A. R. Oliphant, C. J. Brandl, and K. Struhl. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol Cell Biol*, 9(7):2944–9, Jul 1989.

B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290 (5500):2306–9, Dec 2000. doi: 10.1126/science.290.5500.2306.

J. W. Sellers, A. C. Vincent, and K. Struhl. Mutations that define the optimal half-site for binding yeast GCN4 activator protein and identify an ATF/CREB-like repressor that recognizes similar DNA sites. *Mol Cell Biol*, 10(10):5077–86, Oct 1990.

T. Wasson and A. J. Hartemink. An ensemble model of competitive multi-factor binding of the genome. *Genome Res*, 19(11):2101–12, Nov 2009. doi: 10.1101/gr.093450.109.

C. Zhu, K. J. R. P. Byers, R. P. McCord, Z. Shi, M. F. Berger, D. E. Newburger, K. Saulrieta, Z. Smith, M. V. Shah, M. Radhakrishnan, A. A. Philippakis, Y. Hu, F. D. Masi, M. Pacek, A. Rolfs, T. Murthy, J. Labaer, and M. L. Bulyk. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome research*, 19(4):556–66, Apr 2009. doi: 10.1101/gr.090233.108.

4.7 Figures

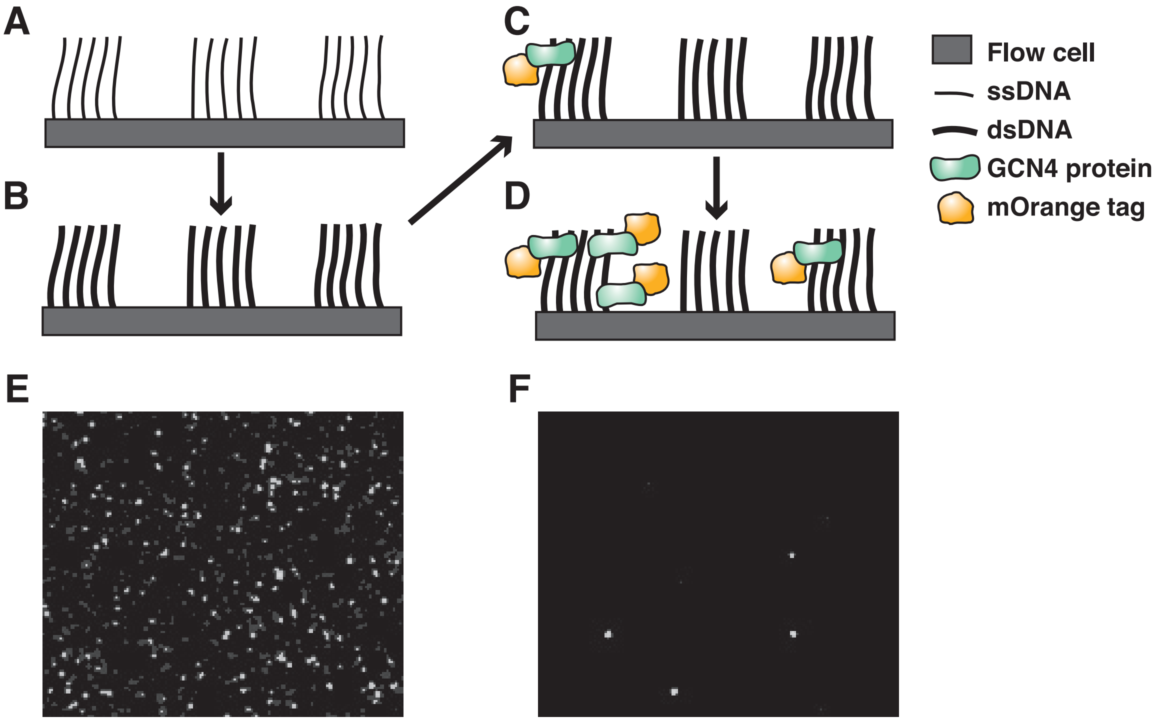


Figure 4-1: Schematic of HiTS-FLIP method. A) A Genome Analyzer flow cell is built and sequenced as usual. The DNA on the chip is organized into clusters having the same sequence, which in this case is random 25mer oligonucleotides. The second strand is then denatured to yield a single strand. B) A primer and polymerase are added to generate complete double-stranded DNA on the flow cell. C) As small amounts of fluorescent fusion protein are added, only high affinity sites are bound after washing. The locations and amounts of binding are imaged by CCD camera. D) With higher concentrations of protein, high affinity sites are bound by more protein, and low affinity sites also begin to be bound. E) A small portion of a sequencing image from a flow cell, showing the roughly one out of four clusters with a fluorescently tagged adenine for that cycle. F) The same portion of the flow cell with fluorescently tagged Gcn4p bound. The small number of clusters with fluorescent signal can be matched with the clusters from sequencing images *in silico*.

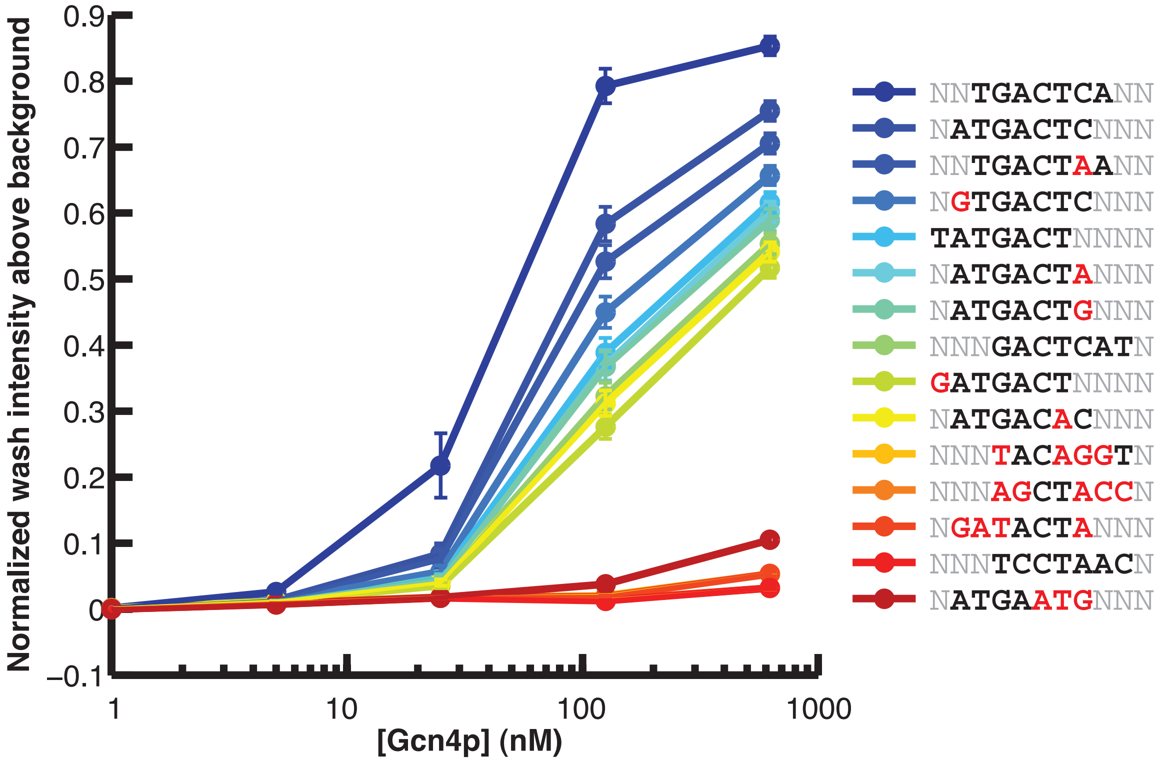


Figure 4-2: HiTS-FLIP binding by concentration. For five concentrations of Gcn4p-mOrange, the median intensity of binding is plotted for selected 7mers. The median was calculated separately for each of seven lanes on the flow cell, and error bars represent one standard error over the seven lanes. Four negative control 7mers are included for comparison. 7mers are aligned to the consensus 11mer of TATGACTCATA with the convention that the 7mer be oriented to match the C in the middle position. Grey "N" letters indicate that any nucleotide could be included in that position; red letters indicate mismatches from the consensus.

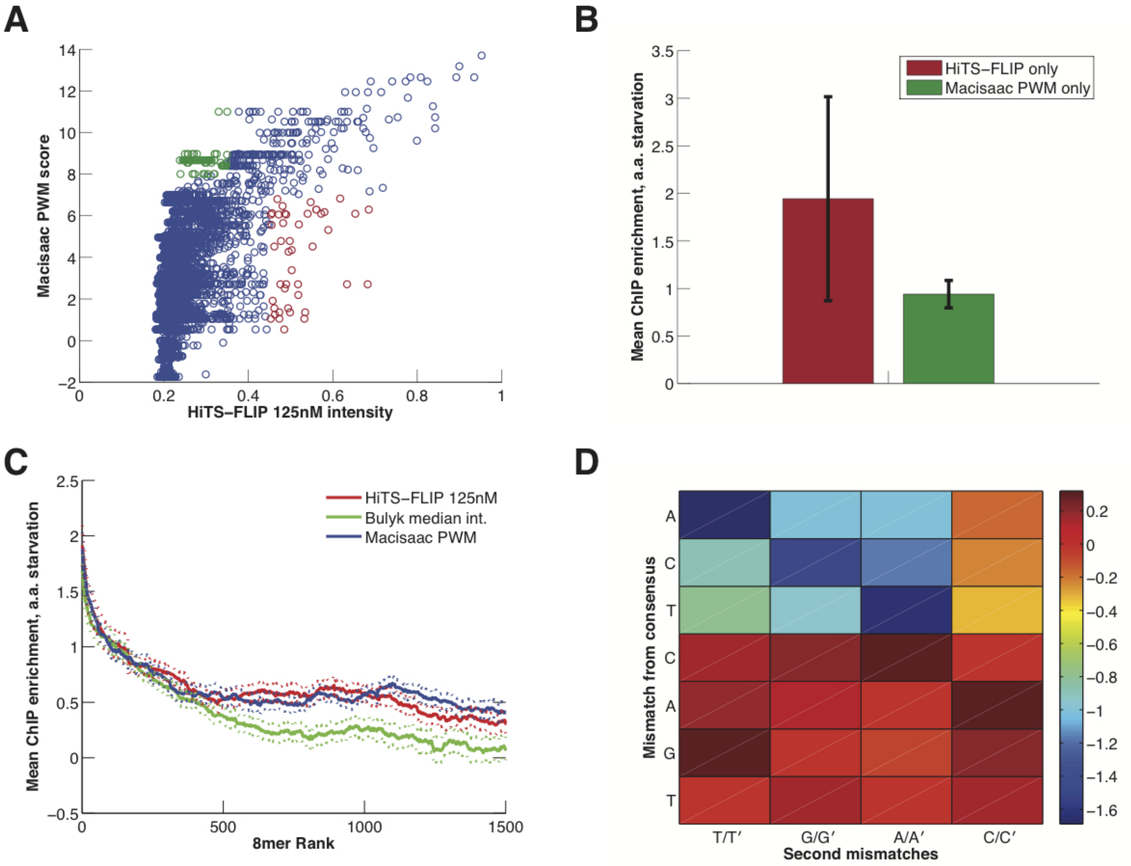


Figure 4-3: HiTS-FLIP intensities correlate well with ChIP-Chip data. A) Scatter plot of all 8mers, comparing the HiTS-FLIP binding intensity at 125nM of Gcn4p-mOrange and the score using a PWM based on ChIP-Chip and conservation data (Macisaac et al. 2006). Points in red are 8mers with high HiTS-FLIP binding intensity but low PWM score, points in green have high PWM score but low HiTS-FLIP intensity. B) 8mers that only scored well in HiTS-FLIP were highly enriched in Gcn4p-bound regions following amino acid starvation (Harbison et al., 2004). In contrast, 8mers that scored well with the PWM but poorly with HiTS-FLIP were marginally enriched in the bound regions. Error bars represent one standard error. C) Comparison of three methods for scoring 8mers: HiTS-FLIP binding intensity with 125nM Gcn4p-mOrange, direct binding intensity from protein binding microarrays (Zhu et al., 2009), and PWM (Macisaac et al. 2006). All 8mers are ranked using each method, and mean enrichment in Gcn4p-bound regions is shown for a running average of 50 8mers. Dotted lines represent ± one standard error. D) Matrix of asymmetrical nucleotide interdependencies. Each square represents the ratio of the binding signal for a 7mer having a mismatch in the left half (e.g. T) to the symmetrical mutation in the right half (e.g. T′). 7mer sequences are consensus in all other positions except for the position indicated by each row. The first four and last three nucleotides have the most correlated effects, signifying the importance of intact half sites (TGAC).

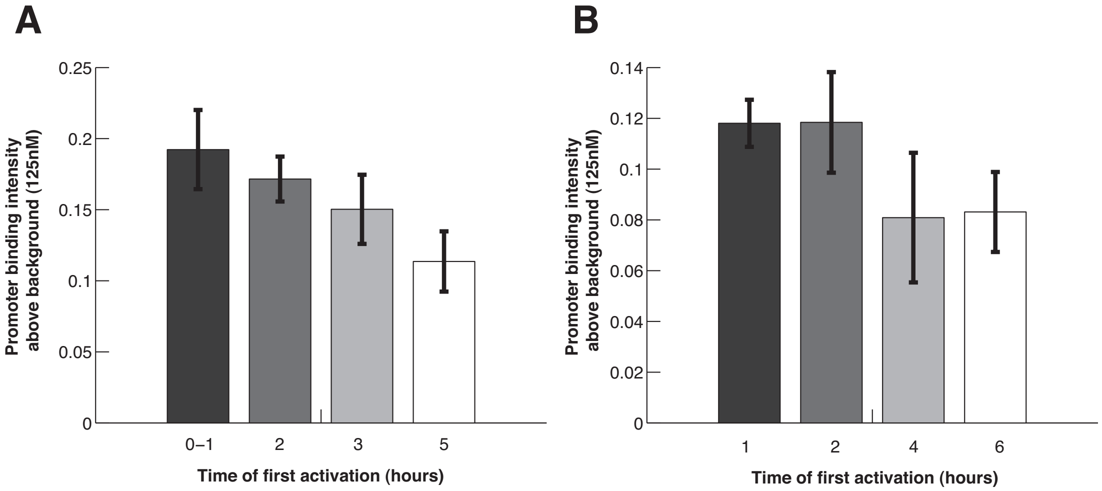


Figure 4-4: Activation by Gcn4p *in vivo*. A) Gene expression was measured by microarray timecourse in yeast cells overexpressing *GCN4* driven by a *GAL* promoter (Chua et al., 2006). For genes significantly activated at each timepoint, we searched their promoter for Gcn4p binding sites and report the highest expected intensity for a 9mer. Genes with weaker binding motifs are activated later, when *GCN4* expression is the highest. Error bars represent one standard error of the mean. B) As in part A, except in a microarray timecourse during amino acid starvation (Gasch et al., 2000). In these conditions, GCN4 is expected to be both translationally and transcriptionally upregulated (Hinnebusch, 2005).

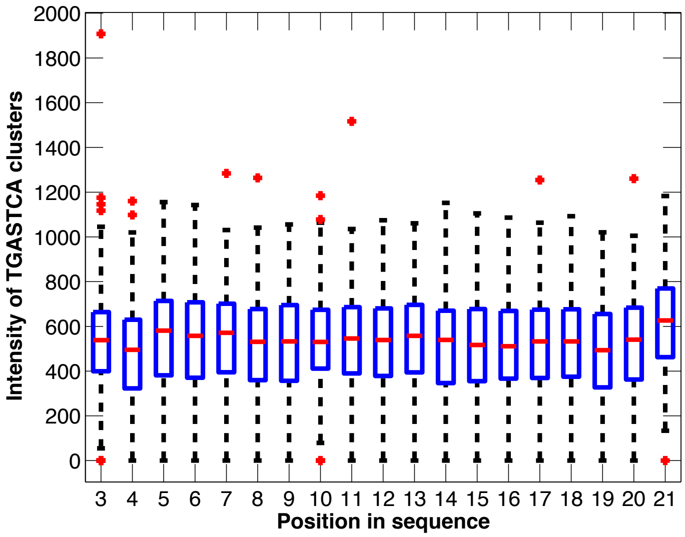
4.8 Supplemental discussion

4.8.1 Technical corrections

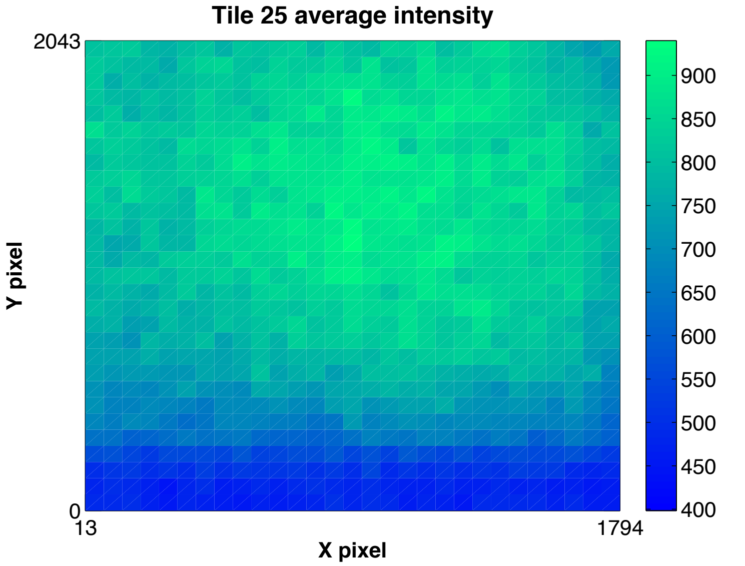
To determine the corrections needed for mapping normalization, photobleaching, and other factors, we performed a set of pilot experiments using various lengths of random oligomers.

We focused on the intensity of sequences containing the canonical GCN4 binding motif (TGASTCA), reasoning that most of the variation in the intensity of these clusters would be due to technical rather than biological reasons. Variables correlating with the intensity of these clusters should therefore be corrected for.

First, we examined whether the motif position in the sequence correlated with intensity due to accessibility of different parts of the DNA. Shown below is a plot of motif-containing-cluster raw intensity for the random 25-mers as a function of position in the sequence. The distribution of raw intensities for motif-containing clusters appears to be independent of the position of the motif in the sequence. Similar results were obtained for 20-mers and 15-mers.



We next examined whether intensity varies as a function of position within the flow cell. Plotted below is the average sequencing intensity for tile 25 of a 15-mer run. Clearly, intensity is strongly influenced by position within a particular tile. For



different tiles, different patterns were observed. Therefore, we investigated normalization based on the average sequencing intensity of a cluster or the labeled primer intensity. The average sequencing intensity correlates better with protein intensity ($R^2 = 0.48$) than the labeled primer intensity does ($R^2 = 0.30$). One interpretation of this data is that the intensity measurements are inherently noisy, and that averaging over dozens of sequencing cycles provides a better measurement of cluster size and brightness than a single measurement. As a result we decided to proceed with analysis based on protein intensity normalized by the average sequencing cycle intensity, reducing technical noise.

To determine the efficiency of washing and photobleaching, we performed an experiment in which we imaged a bound flow cell twice in a row without washing, and in a separate experiment imaged after washing. We found that the second round of imaging without washing reduced fluorescent intensity of TGASTCA clusters by roughly 35%, presumably due to photobleaching. The washing reduced intensity by roughly 65%. When we sequentially flow on increasing concentrations of Gcn4p, we do not completely wash off the protein on the flow cell, and some of what remains will be photobleached. Therefore, we correct the intensity at concentration i (C_i) using the equation:

$$C_{i,corrected} = C_i + C_{i-1} \times (1 - 65\%) \times (35\%)$$

This equation assumes that the fraction of Gcn4p remaining after washing and photobleaching will remain bound in the next set of imaging. We believe this to be a reasonable assumption, given that without this correction we observe decreasing Gcn4p signal for TGASTCA clusters at the highest concentration.

4.9 Supplemental figures

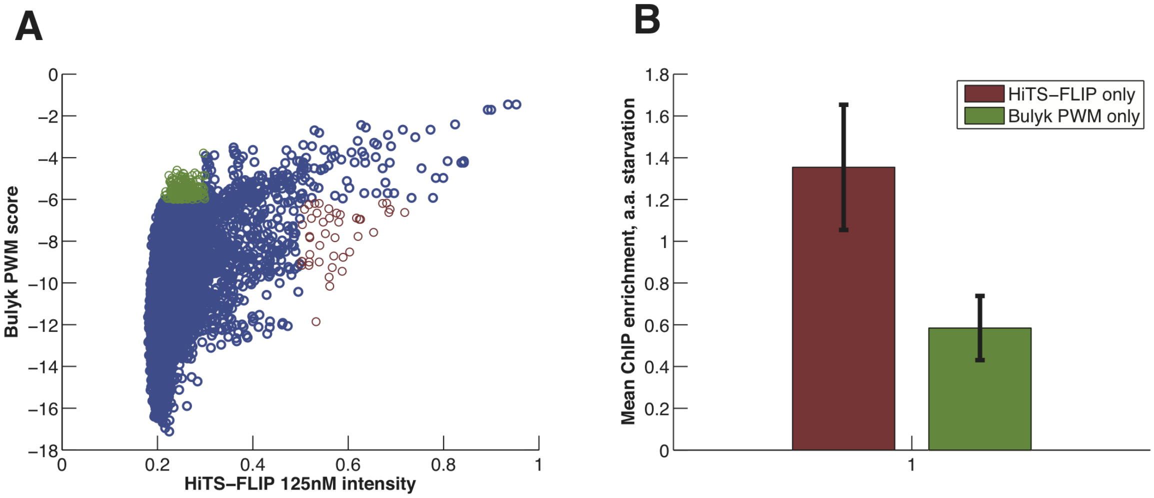


Figure 4-5: HiTS-FLIP intensities correlate better with ChIP-Chip data than PWM predictions. Plot is exactly as in figure 4-3A and B except that the PWM is based on protein binding microarray data (Zhu et al., 2009).

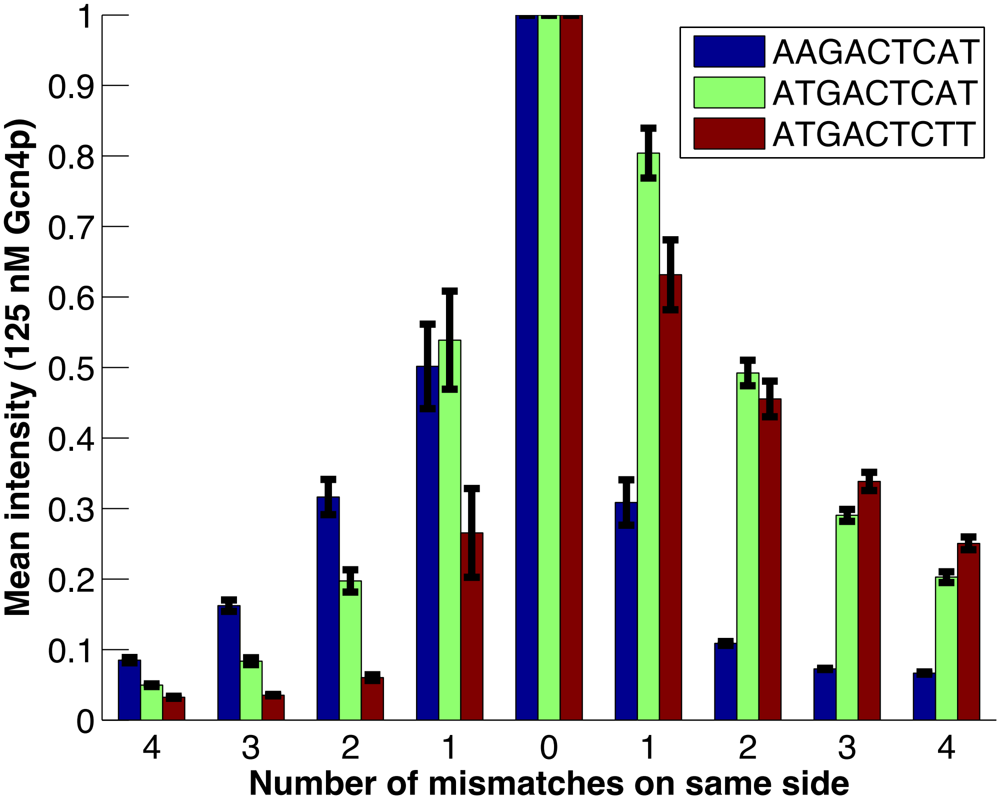


Figure 4-6: Half-site asymmetry and importance of intact half-sites. For three 9mers, the result of changing either the left or right half is plotted. On the right half of the graph, mismatches accumulate in the 3′ end of the sequence; likewise the number of the left half corresponds to the number of mutations in the 5′ half of the sequence. The central "C" nucleotide is always maintained. The intensity is relative to no mismatches, so the bars at zero have a height of exactly one. The green bars show mismatches away from the canonical binding sequence, revealing that mismatches are better tolerated on in the 3′ half of the binding site. This confirms known preferences of Gcn4p (Sellers et al., 1990). When a mismatch already exists in the 3′ half (red bars), the asymmetry is maintained. However, when a symmetrical mismatch exists instead in the 5′ half (blue bars), the asymmetry is reversed. Because the 3′ half is closer to a consensus half-site match, further mismatches are tolerated better in the 5′ half in this case. Taken together, these results confirm that intact half-sites are an important feature in Gcn4p binding.

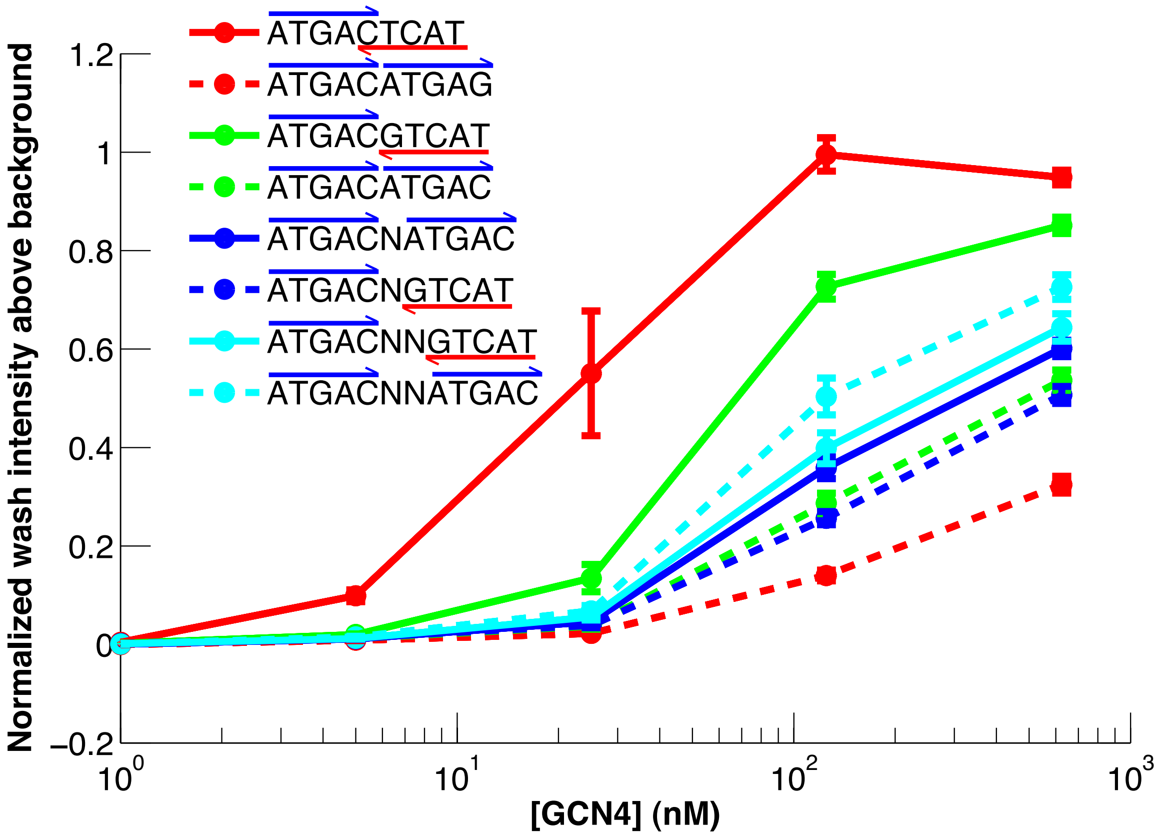


Figure 4-7: Effect of spacing between half-sites on Gcn4p binding. The canonical 9mer motif, a palindromic 10mer consisting of two complete half-sites, and palindromic 10mers with intervening spacers are plotted. For each sequence, a set of two half-sites in the same orientation (as opposed to reverse-complement) and with equivalent spacing is plotted with dashed lines for comparison. Half-sites in forward orientation are marked in blue; those in reverse orientation are marked in red. The dashed lines should represent only binding to two half-sites separately, whereas solid lines represent both independent binding of two half-sites as well as cooperative binding by one Gcn4p molecule.

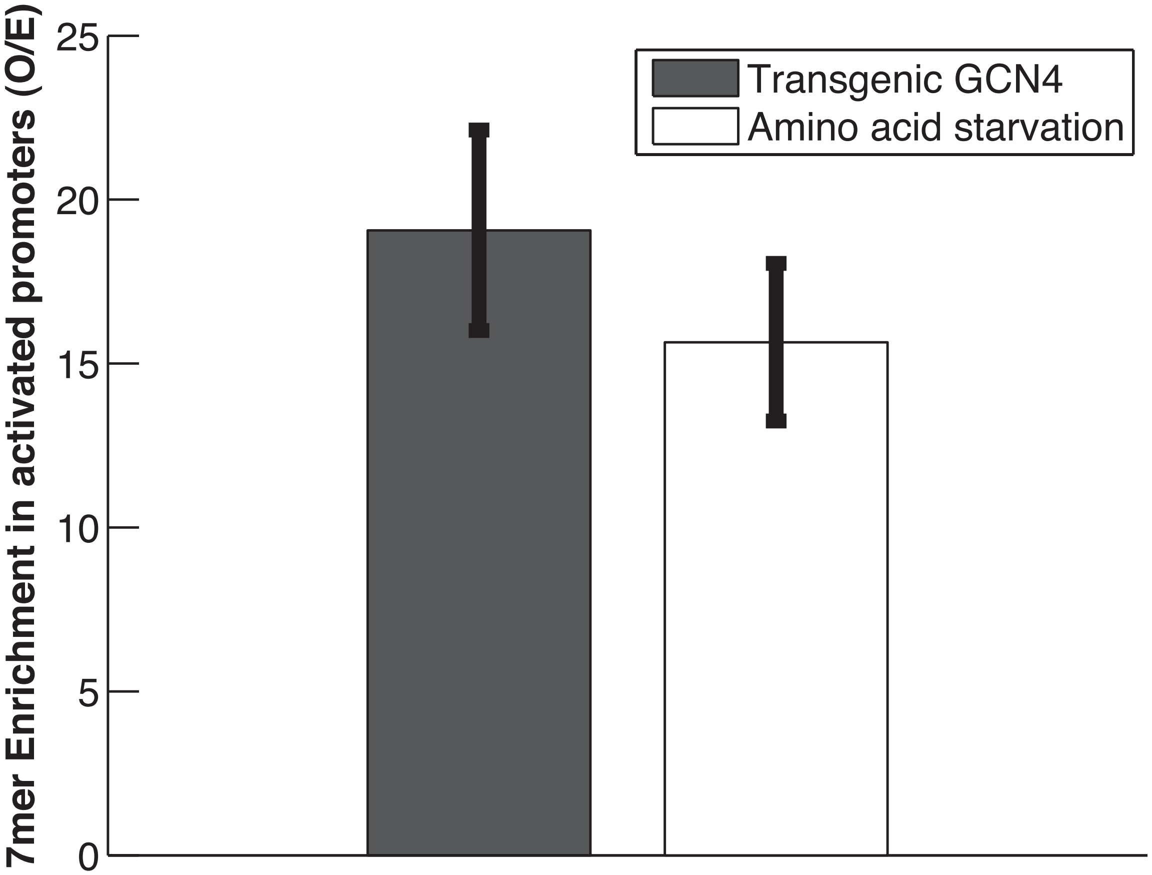


Figure 4-8: Enrichment of Gcn4p binding sequences in promoters of *GCN4*-activated genes. For a transgenic *GCN4* experiment (Chua et al., 2006) and an amino acid starvation time-course (Gasch et al., 2000), the promoters of genes that were significantly activated were examined. For the top ten 7mers by HiTS-FLIP binding intensity, the average enrichment over an expectation based on all yeast promoters was plotted. Error bars represent one standard error of the mean.

Chapter 5

Conclusions

5.1 Summary

This thesis has presented two novel approaches to assaying gene regulatory interactions, both leading to novel biological insights. In chapter 2, I described a method for quantifying the conservation of miRNA target sites above background levels, building on previous work (Lewis et al., 2003, 2005; Brennecke et al., 2005; Kheradpour et al., 2007) but providing substantial improvements that enabled new types of analysis. For the first time, I was able to reliably quantify the relative contribution of different seed match types to conserved mammalian miRNA targeting, discovering a new seed match type (the offset 6mer) and uncovering the surprising prevalence of natural selection acting on seed matches with low experimental efficacy. I quantified the conservation of minor classes of targets having imperfect seed matches or strong $3'$ supplementary or compensatory pairing. I showed that mammalian-specific miRNAs are qualitatively different from more broadly-conserved miRNAs, having far less conservation. In chapter 3, I applied this method to nematodes and flies, finding some surprising differences in miRNA target conservation between clades. Nematodes show surprisingly high conservation of weak seed matches in comparison to vertebrates, as well as two types of seed match with no evidence for efficacy or conservation in vertebrates, the 6mer-A1 and the 8mer-U1. I showed that $3'$ UTR length correlated with the conservation of weak seed match types, as well as the density of miRNA targets. I provided a plausible evolutionary model to explain this correlation based on the avoidance of deleterious seed matches, a well-known phenomenon (Farh et al., 2005; Stark et al., 2005). In chapter 4, I presented HiTS-FLIP, a new method for quantifying transcription factor binding preferences with hundreds of millions of measurements. I applied this method to the yeast transcription factor *GCN4*, confirming all known preferences for *GCN4* binding and discovering new subtleties in its binding preferences, including an extended consensus sequence and an ability to accommodate a spacer between two half binding sites. This assay also provides direct and quantitative estimates for hundreds of thousands of dissociation constants, which should prove invaluable for computational modeling of transcription factor binding.

5.2 Conceptual progress

5.2.1 Technical advances

Although there were countless technical problems solved during the course of this thesis work, there are a few conceptual advances that I hope will stand out and impact a broad community of researchers. Perhaps the most general is the necessity of controlling for local variation in conservation rates during analysis of purifying selection. Most classical methods for measuring natural selection, for example the Ka/Ks test, explicitly control for local conservation rates (Yang and Bielawski, 2000). However, many genome-scale methods for detecting purifying selection on motifs, for example Xie et al. (2005), have not accounted for this variable. I have shown that local variation in conservation rates, due to both alignment artifacts and biological causes, can strongly influence measured patterns of natural selection (appendix A). In particular, predictions for the selection on individual sites in the genome such as the P_{CT} (section 2.2.7) are extremely susceptible to this bias if it is not properly controlled for. A related lesson is that dinucleotide conservation rates and interrelationships between target sequences must be accounted for in order to have accurate prediction of individual target sequences. It was only by carefully disentangling 6mer seed matches from 7mers and 8mers that I could confidently measure the surprising extent of 6mer targeting in mammals (section 2.2.2, appendix A). The P_{CT} itself represents a third technical advance, providing a convenient score for the natural selection of a potential target sequence. The P_{CT} has the advantage of a simple interpretation (the probability of a sequence being under natural selection to maintain miRNA targeting) that hides the complexity of taking into account local conservation rates, alignment artifacts, dinucleotide composition, seed match types, and the global extent of natural selection for seed matches of a miRNA. The major impact of this project will likely be the widespread use of its target predictions (www.targetscan.org), but hopefully the methods behind the target prediction will have an influence on future evolutionary analysis as well.

The HiTS-FLIP project represents mostly a proof of principle, showing that the method recapitulates *in vivo* binding and function and can reveal complex binding preferences in remarkable detail. Because HiTS-FLIP generates remarkable amounts of data, is amenable to the study of proteins with extremely complex binding motifs, and can quantify biophysical properties of binding, I expect this method to be used for a number of applications. Hopefully, my analysis of *GCN4* binding preferences will convince others that HiTS-FLIP is tractable and generates biologically relevant data, especially details about suboptimal binding sites.

5.2.2 Importance of low affinity interactions

A major theme of this thesis is that low-affinity interactions between *trans*-acting factors and *cis*-elements are biologically relevant, both in the context of 6mer seed matches having strong conservation despite their weak efficacy (chapters 2 and 3) and *GCN4* binding affinity determining the timing of gene activation following amino acid starvation (chapter 4). The idea that quantitative affinities are important for gene regulation is hardly a new insight – examples of weak interactions having profound effects are strewn throughout the literature. And yet, I believe this fact is often overlooked, chiefly for two reasons: the reductionist approach to biology necessitates prioritizing only strong effects for followup; and the noise inherent in experimental determination of biological effects makes validation of weak interactions difficult.

One way of circumventing these limitations is to focus on situations in which the *trans*-factor is expressed at a high level relative to its targets, causing formerly weak interactions to become prominent. For example, I found stronger conservation of 6mer seed matches in species with short 3′ UTRs (chapter 3). The zebrafish maternal to zygotic transition provides another example, in which miR-430 is expressed at an extremely high level, repressing many maternal mRNAs having only "weak" 6mer seed matches. The ratio between the expression of a *trans*-factor and a *cis*-element can vary in a spatial dimension as well. For example, in the developing *Drosophila* embryo, Dorsal activates enhancers with weak binding sites (Type 1 enhancers) only in ventral regions of the embryo, which paradoxically have the highest levels of Dorsal expression (Papatsenko and Levine, 2005). This thesis serves as a reminder from a systems biology point of view that low affinity interactions should not be ignored.

5.3 Approaches to studying molecular interactions

I have utilized two disparate approaches to studying regulatory interactions, evolutionary conservation and an *in vitro* binding assay. I have verified the relevance of predictions based on these two approaches by applying them to explain microarray data following the induction of regulators or *in vivo* ChIP-Chip binding data. These different approaches each have strengths and weaknesses that complement each other. Methods to profile *in vivo* binding events, such as ChIP-Seq and CLIP-Seq, have deservedly enjoyed particular attention. However, I would like to emphasize the strengths of an evolutionary conservation approach over other methods. A major challenge in studying genome-wide gene regulation is how to determine whether interactions are biologically relevant. Profiling of gene expression, for example after introducing a transgenic regulator or knocking out an endogenous one, can assign function to binding events. However, small differences in gene expression could simply be noise, whereas the biologically important processes might be subject to canalization that masks the impact of these changes from the phenotype. Given an appropriate null model and statistical framework, a significant signal for evolutionary conservation surely corresponds to function that is important for the survival of the organism in some way, sidestepping any problem of noisy gene expression. On the other hand, knowledge of the *in vitro* binding preferences of *trans*-factors enables a more mechanistic understanding of gene regulation. Using *in vitro* data, one can disentangle the intrinsic binding affinity of sites from co-factor binding and context or chromatin effects (Gordân et al., 2009; Wasson and Hartemink, 2009).

5.4 Extensions and applications

There are several avenues of research that would be fruitful follow-ups to this thesis work. The miRNA target prediction algorithm would be well served by a more thorough determination of $3'$ UTR alignments. So far I have been using UCSC's multi-z alignments due to their accessibility and comprehensive nature, but believe that a more careful consideration of orthologous versus paralagous relationships could substantially expand the estimate for the number of conserved targets. Also, to date I have focused on $3'$ UTRs, the regions in which miRNA targeting is most effective (Grimson et al., 2007). However, miRNAs are known to repress messages with targets in the ORF (Bartel, 2009), leaving an opening for new target predictions to make a substantial contribution. A new conservation model would need to be aware of reading frame and amino acid coding in order to separate conservation signal from background in this context.

The analysis of HiTS-FLIP could benefit greatly from a biophysical model of binding based on the measured dissociation constants. In chapter 4, the analysis of promoter regions and regions bound in ChIP-Chip is based on the maximum binding intensity of any k-mer in the region. A biophysical model could convert dissociation constants into occupation probabilities, allowing an integration of weak binding over a long sequence. As for further experiments, HiTS-FLIP would be quite informative for a number of transcription factors. A gene such as p53, with a complex 20-mer consensus binding sequence and numerous splice variants and polymorphisms that have unknown effects on binding, would be an excellent candidate.

However, there is a different kind of follow-up question that is perhaps personally more interesting: how exactly does one translate catalogs of molecular interactions such as those provided by this thesis into true biological insight? Although I have attempted to address this in small ways already (for example, see figure 3-3B), the answer is not at all obvious. As an illustration, let us consider the innate immune response in the mammalian gut. If one wants to design a drug that reduces inflammation, thereby resolving chronic infections that depend on an inflammatory state, there are several possible approaches to selecting a candidate target gene. For example, one could employ a high-throughput screen of small molecules or siRNAs, take a reductionist approach and use low-throughput techniques to identify individual interactions, or use a systems approach and model the immune response network, simulating the response to different perturbations. For a system as complex as innate immunity, the problem would be daunting even with a combination of all three approaches. However, quantitative catalogs of gene regulatory interactions could assist with all three approaches. For candidates found by a high-throughput siRNA screen, predicted regulatory interactions could point to potential mechanisms of action and prioritize follow-up experiments. In the case of the innate immune response, relationships with known regulatory cytokines and NF-κB family members might help separate false positives from promising candidates. Likewise, predicted interactions could be used to generate hypotheses for low-throughput techniques. For example, interactions found between a miRNA and an NF-κB gene by conservation analysis, but not supported in the literature, might point to fruitful avenues of research. Finally, modeling approaches are hopeless unless they are constrained by large amounts of quantitative data. Models of the transcriptional response mediated by NF-κB might become quantitative enough to be useful if informed by a HiTS-FLIP assay first. As always, the devil is in the details – how exactly global regulatory data is translated into meaningful and practical biological advances will ultimately determine the utility of systems biology.

5.5 References

D. P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2): 215–33, Jan 2009. doi: 10.1016/j.cell.2009.01.002.

J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen. Principles of microRNA-target recognition. *PLoS Biol*, 3(3):e85, Mar 2005. doi: 10.1371/journal.pbio.0030085.

K. K.-H. Farh, A. Grimson, C. Jan, B. P. Lewis, W. K. Johnston, L. P. Lim, C. B. Burge, and D. P. Bartel. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, 310(5755):1817–21, Dec 2005. doi: 10.1126/science.1121158.

R. Gordân, A. Hartemink, and M. Bulyk. Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res*, Sep 2009. doi: 10.1101/gr.094144.109.

A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell*, 27(1):91–105, Jul 2007. doi: 10.1016/j.molcel.2007.06.017.

P. Kheradpour, A. Stark, S. Roy, and M. Kellis. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res*, 17(12):1919–31, Dec 2007. doi: 10.1101/gr.7090407.

B. P. Lewis, I. hung Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–98, Dec 2003.

B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, Jan 2005. doi: 10.1016/j.cell.2004.12.035.

D. Papatsenko and M. Levine. Quantitative analysis of binding motifs mediating diverse spatial readouts of the Dorsal gradient in the Drosophila embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 102(14): 4966–71, Apr 2005. doi: 10.1073/pnas.0409414102.

A. Stark, J. Brennecke, N. Bushati, R. B. Russell, and S. M. Cohen. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123(6):1133–46, Dec 2005. doi: 10.1016/j.cell.2005.11.023.

T. Wasson and A. J. Hartemink. An ensemble model of competitive multi-factor binding of the genome. *Genome Res*, 19(11):2101–12, Nov 2009. doi: 10.1101/gr.093450.109.

X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–45, Mar 2005. doi: 10.1038/nature03441.

Z. Yang and J. Bielawski. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol (Amst)*, 15(12):496–503, Dec 2000.

Appendix A

Supplementary material for chapter 2

Supplemental Discussion

Combining signal and background distributions for the 10 UTR bins

Controlling for the local conservation rate is of vital importance in this type of study. Local conservation rate can affect the variability in the background estimate, thereby decreasing sensitivity and statistical power of the analysis. Moreover, the local conservation rate can dramatically distort estimates of preferential conservation (Lewis et al., 2005). To illustrate this distortion, we estimated the conservation signal and background without controlling for UTR conservation, and then plotted the signal-to-background ratio for conserved sites that fall within ten different levels of local conservation (Supplemental Fig. 3, upper panel). These results illustrate how previous methods, which use a single tree and have a single estimate of the background distribution for all UTRs, overestimate the preferential conservation of sites in highly conserved UTRs and can miss the preferential conservation of sites in poorly conserved UTRs. Indeed, when using a single tree and single background estimate, miRNA seed matches in the poorly conserved UTRs appear to be actively avoiding conservation when compared to the background, an observation that is very unlikely to be biological. Therefore, previous methods, which do not control for local conservation, reliably show that many seed-matched sites are preferentially conserved, but they are not reliable in distinguishing individual sites that are preferentially conserved from those that are conserved by chance.

There are multiple reasonable ways to control for local conservation. Our method of separating the UTRs into bins based on their conservation rates raised the question of how to combine the data from these bins. It was not obvious *a priori* that the bins could be treated in the same way. If the bins were not equivalent with respect to the relevant measurement (in this case, branch-length cutoff), the complete analysis might have to be performed separately for each of the ten UTR bins. However, if the bins were equivalent, then the signal and background values for each bin could be safely combined by simply summing at each cutoff to create the aggregate signal and background distributions.

In order to make the UTR bins more comparable, we recalculated the phylogenetic trees separately for each bin. This approach allowed for fine-tuning of the relative branch-lengths, which may provide additional benefits beyond a uniform rescaling of all branch lengths. We reasoned that after this kind of scaling, the signal-to-background ratio would be close to equivalent for all ten bins at a given branch-length cutoff. Indeed, the signal above background reached a maximum at about the same branch-length cutoff (1.0) for each of the bins, after recalculating the trees for each bin (data not shown). Moreover, the large variability in signal-to-background ratio observed with a single tree was greatly reduced upon recalculating the trees for each bin (Supplemental Fig. 3). Most of the remaining variability could be attributed to edge effects in bin 1 and bin 10. Hence, after recalculating the phylogenetic trees for each bin, we could safely combine the results for the ten UTR bins into one estimate of signal and background, since our confidence in the preferential conservation of a site at any particular branch length was largely independent of the UTR bin to which it was assigned.

Nested seed matches

As schematically depicted in Figure 1D, we nested smaller seed matches within larger ones, which led to a substantial increase in sensitivity. Because many miRNA sites have species-specific differences in seed-match type, a sensitive method was required for determining the largest conserved unit. Our approach was to begin with the largest seed-match class (8mers), and subtract both the signal and the background of the larger seed matches from the signal and background of the smaller ones. Hence, a 6mer conserved to branch length 1.0 contributed to the number of conserved 6mers only if it was the largest functional seed-match unit that was conserved to that branch length; if the 6mer was subsumed in an 8mer conserved to branch length 1.0, it was not counted as a conserved 6mer. In this way, we classified seed match conservation as the longest seed-match type possible, but also allowed for species-specific differences without losing sensitivity.

6mer signal above background

Given the observation of many conserved 6mer seed matches above background, it is natural to ask whether these sites are being selectively maintained or whether there could be other, technical reasons for their preferential conservation. Indeed, because of the methodology discussed above, conserved 6mers may appear in some species as 7mer or 8mer seed matches. This leaves open the possibility that the observed preferential conservation of 6mers could be due to mutation from conserved 7mers, i.e., it could be due to preferential conservation in the 7mer form, with only chance conservation as 6mers.

We have performed two analyses to test whether the 6mer conservation observed could be attributed to decay of conserved 7mers. First, we tested the possibility that 7mer sites in human contribute to 6mer conservation signal through decay in orthologous species by examining only the subset of 6mers that were not part of a 7mer in the human UTR. We found in this subset that there was significant enrichment in conservation for both canonical 6mer seed matches and offset 6mers (Supplemental Figure 2B). The converse possibility is that the seed match is conserved as a 7mer in other species, but is a 6mer in human. Because of technical issues, this possibility was difficult to evaluate directly. But to get a sense of its impact we reasoned that the number of 6mers in human preferentially conserved as a 7mer in other species that include mouse, would mirror the number of 6mers in mouse preferentially conserved as a 7mer in other species that include human. With this in mind, we performed an analysis examining conserved 7mers that have decayed to 6mers in the mouse UTRs. Testing all possible branch-length cutoffs, we counted the number of such sites, scaled by the proportion of 7mer sites conserved above background (given by (S-B)/S). The cutoff yielding the most 6mer decay (i.e., capturing the most sites above background) was 0.9, corresponding to selectively maintained 7mer sites that are conserved to a branch length of 0.9 but would appear as a 6mer conserved to 1.0 in a mouse-centric analysis. For all 87 broadly conserved families combined, there were 888 decayed 6mers and 362 decayed offset 6mers above background. Symmetrically, one would expect that roughly the same number of conserved 6mers we observed in human are conserved because the site is a selectively maintained 7mer or 8mer in other species. Combining these two sources of error in an aggregate estimate, we still predict 77 6mer seed matches conserved above background per miRNA, and 69 per miRNA for offset 6mers.

By eliminating all 6mer conservation when the human site has a 7mer, some sites that are preferentially conserved as 6mers were surely lost, causing the first analysis to overestimate the number of 6mers conserved due to decay of a human 7mer. In fact, when allowing for single mismatches in 7mers that create 6mer seed matches, extra conservation is added equally to the signal and to the background, yielding roughly the same number of predicted targets as a 7mer conservation analysis. This suggests that when our methods detect a preferentially conserved 6mer that is a 7mer in human, the preferential conservation of the site is due to its presence as a 6mer in other vertebrates and not due to its presence as a 7mer. In the second analysis, in all likelihood there are selectively-maintained 7mers conserved to branch-lengths less than 0.9 that are not accounted for because this preferential conservation is difficult to detect with our methods, potentially leading to an underestimate of the number of human 6mers preferentially conserved only because of their activity as 7mers in other species. In balance, we believe that the overestimate of the first analysis outweighs the underestimate of the second analysis, making our aggregate estimate conservative. Thus, we cannot explain the conservation of 6mer seed matches by their relationship to conserved 7mers.

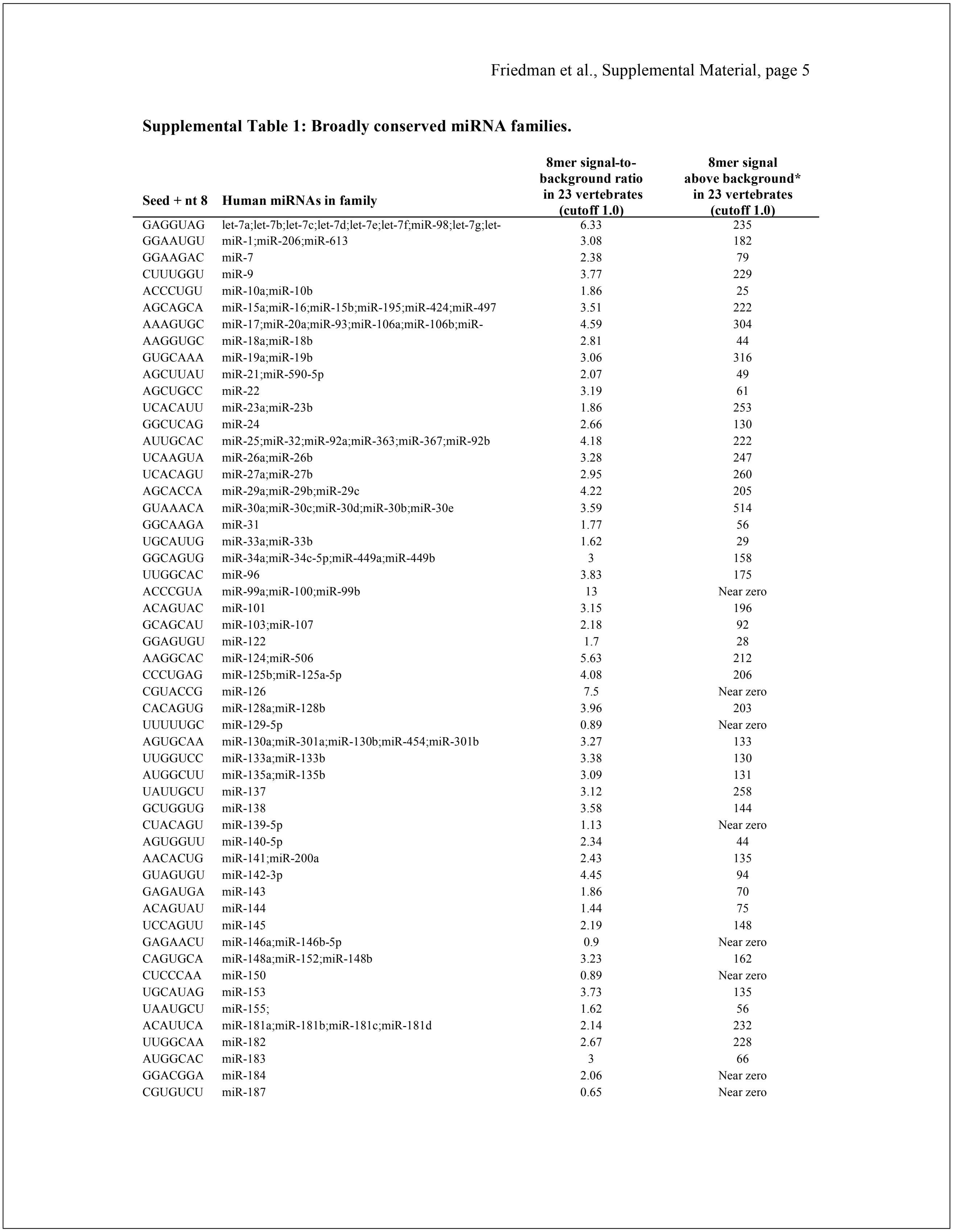
It is worth noting that in cases found by the above analysis, in which the 6mer preferential conservation might be attributable to conserved 7mers, our estimate for the number of sites conserved above background and the number of preferentially conserved miRNA targets remains the same. The difference in our two estimates of 6mer conservation above background merely reflects the difficulty of assigning the preferential conservation we observe to individual seed-match types. In cases in which a seed-match site is broadly conserved, but exists as different seed-match types in different species, it is not obvious which type should be assigned the preferential conservation, and in many cases orthologous sites with different seed match types are sure to be simultaneously selectively maintained. Despite this uncertainty, our methods find such preferential conservation with high sensitivity without double-counting, and we have shown that even the weakly effective 6mer and offset 6mer seed matches have substantial conservation independent of the other types.

***P*$_{\mathrm{CT}}$ values reported on the TargetScan website**

While calculating P_{CT} values for the TargetScan website, we observed a high variability of P_{CT} values for sites with close branch-length values. This variability was observed for only a subset of miRNAs, and even for those in which it was observed, the strong underlying trend of higher PCT at higher branch lengths was clear, which explains the correlation with the experimental data when looking at the P_{CT} scores in aggregate (Figure 6). Nonetheless, we considered it prudent to implement a smoothing procedure when deriving P_{CT} values reported at the TargetScan website. Thus, for each miRNA and for each seed match type, we fit a modified sigmoid function to the P_{CT} scores using a least squares estimator. The function was given by:

$$P_{CT} = \max\left(0, \beta_0 + \frac{\beta_1}{1 + \exp(-\beta_2 x + \beta_3)}\right)$$

where, x is the branch length value for a particular site. In other words, the P_{CT} score reported on the Targetscan website was either the output of a modified sigmoid function given by $\beta_0 + \beta_1/(1 + \exp(-\beta_2 x + \beta_3))$, or zero if that function was negative. The values of β_0, β_1, β_2, and β_3, were fit so that the function would best match the raw P_{CT} values. β_0 and β_1 can be interpreted as offsetting and scaling the P_{CT}, respectively, whereas β_2 and β_3 can be interpreted as scaling and offsetting the influence of the branch-length value, respectively. There is no special significance to the particular form of the function or the value of the parameters — rather, we observed that the P_{CT} values closely followed this modified sigmoid function, and that in all cases the modified sigmoid function closely matched the data but substantially smoothed curves plotting P_{CT} with respect to the branch-length.



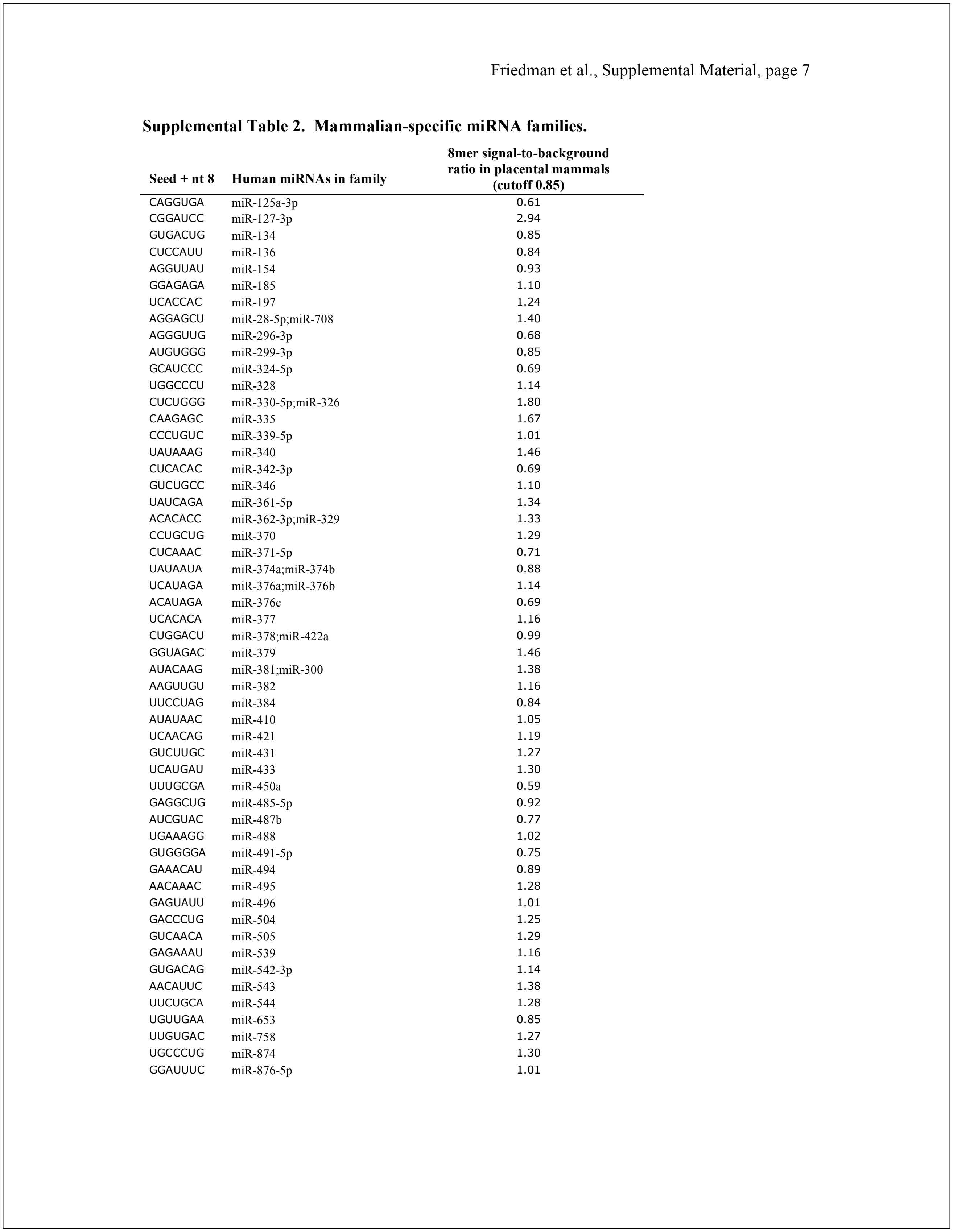
Supplemental Table 1: Broadly conserved miRNA families.

Seed + nt 8	Human miRNAs in family	8mer signal-to-background ratio in 23 vertebrates (cutoff 1.0)	8mer signal above background* in 23 vertebrates (cutoff 1.0)
GAGGUAG	let-7a;let-7b;let-7c;let-7d;let-7e;let-7f;miR-98;let-7g;let-	6.33	235
GGAAUGU	miR-1;miR-206;miR-613	3.08	182
GGAAGAC	miR-7	2.38	79
CUUUGGU	miR-9	3.77	229
ACCCUGU	miR-10a;miR-10b	1.86	25
AGCAGCA	miR-15a;miR-16;miR-15b;miR-195;miR-424;miR-497	3.51	222
AAAGUGC	miR-17;miR-20a;miR-93;miR-106a;miR-106b;miR-	4.59	304
AAGGUGC	miR-18a;miR-18b	2.81	44
GUGCAAA	miR-19a;miR-19b	3.06	316
AGCUUAU	miR-21;miR-590-5p	2.07	49
AGCUGCC	miR-22	3.19	61
UCACAUU	miR-23a;miR-23b	1.86	253
GGCUCAG	miR-24	2.66	130
AUUGCAC	miR-25;miR-32;miR-92a;miR-363;miR-367;miR-92b	4.18	222
UCAAGUA	miR-26a;miR-26b	3.28	247
UCACAGU	miR-27a;miR-27b	2.95	260
AGCACCA	miR-29a;miR-29b;miR-29c	4.22	205
GUAAACA	miR-30a;miR-30c;miR-30d;miR-30b;miR-30e	3.59	514
GGCAAGA	miR-31	1.77	56
UGCAUUG	miR-33a;miR-33b	1.62	29
GGCAGUG	miR-34a;miR-34c-5p;miR-449a;miR-449b	3	158
UUGGCAC	miR-96	3.83	175
ACCCGUA	miR-99a;miR-100;miR-99b	13	Near zero
ACAGUAC	miR-101	3.15	196
GCAGCAU	miR-103;miR-107	2.18	92
GGAGUGU	miR-122	1.7	28
AAGGCAC	miR-124;miR-506	5.63	212
CCCUGAG	miR-125b;miR-125a-5p	4.08	206
CGUACCG	miR-126	7.5	Near zero
CACAGUG	miR-128a;miR-128b	3.96	203
UUUUUGC	miR-129-5p	0.89	Near zero
AGUGCAA	miR-130a;miR-301a;miR-130b;miR-454;miR-301b	3.27	133
UUGGUCC	miR-133a;miR-133b	3.38	130
AUGGCUU	miR-135a;miR-135b	3.09	131
UAUUGCU	miR-137	3.12	258
GCUGGUG	miR-138	3.58	144
CUACAGU	miR-139-5p	1.13	Near zero
AGUGGUU	miR-140-5p	2.34	44
AACACUG	miR-141;miR-200a	2.43	135
GUAGUGU	miR-142-3p	4.45	94
GAGAUGA	miR-143	1.86	70
ACAGUAU	miR-144	1.44	75
UCCAGUU	miR-145	2.19	148
GAGAACU	miR-146a;miR-146b-5p	0.9	Near zero
CAGUGCA	miR-148a;miR-152;miR-148b	3.23	162
CUCCCAA	miR-150	0.89	Near zero
UGCAUAG	miR-153	3.73	135
UAAUGCU	miR-155;	1.62	56
ACAUUCA	miR-181a;miR-181b;miR-181c;miR-181d	2.14	232
UUGGCAA	miR-182	2.67	228
AUGGCAC	miR-183	3	66
GGACGGA	miR-184	2.06	Near zero
CGUGUCU	miR-187	0.65	Near zero

Supplemental Table 1 (continued)

GAUAUGU	miR-190;miR-190b	1.54	23
AACGGAA	miR-191	3.75	Near zero
UGACCUA	miR-192;miR-215	0.99	Near zero
ACUGGCC	miR-193a-3p;miR-193b	2.17	29
GUAACAG	miR-194	1.89	48
AGGUAGU	miR-196a;miR-196b	3.62	33
CCAGUGU	miR-199a-5p;miR-199b-5p	3.5	136
AAUACUG	miR-200b;miR-200c;miR-429	3.14	199
UGAAAUG	miR-203	1.33	52
UCCCUUU	miR-204;miR-211	1.44	81
CCUUCAU	miR-205	1.61	56
UAAGACG	miR-208;miR-208b	2.19	Near zero
UGUGCGU	miR-210	2.42	Near zero
AACAGUC	miR-212;miR-132	2.33	54
CAGCAGG	miR-214	2.1	61
AAUCUCU	miR-216b	1.09	Near zero
AAUCUCA	miR-216a	0.78	Near zero
ACUGCAU	miR-217	1.35	22
UGUGCUU	miR-218	3.25	230
GAUUGUC	miR-219-5p	3.79	82
GCUACAU	miR-221;miR-222	1.78	68
GUCAGUU	miR-223	1.69	41
AAGUGCU	miR-302a;miR-302b;miR-302c; miR-302d;miR-372;miR-373;miR-	2.24	70
CCAGCAU	miR-338-3p	1.14	Near zero
AAUGCCC	miR-365	2.11	30
UUGUUCG	miR-375	2.73	Near zero
GAUCAGA	miR-383	1.03	Near zero
AUGACAC	miR-425	1.07	Near zero
AACCGUU	miR-451	3.08	Near zero
AUGUGCC	miR-455-5p	1.25	Near zero
AACCUGG	miR-490-3p	0.88	Near zero
UAAGACU	miR-499-5p	1.19	16
AGCAGCG	miR-503	5.15	Near zero
CGACCCA	miR-551a;miR-551b	3.75	Near zero

*Values for signal above background of <16 sites were designated as "near zero." This cutoff was chosen because the most poorly performing 8mer had a background estimate 16 sites greater than the signal, putting a conservative upper limit on the ability to distinguish preferentially conserved sites from background.



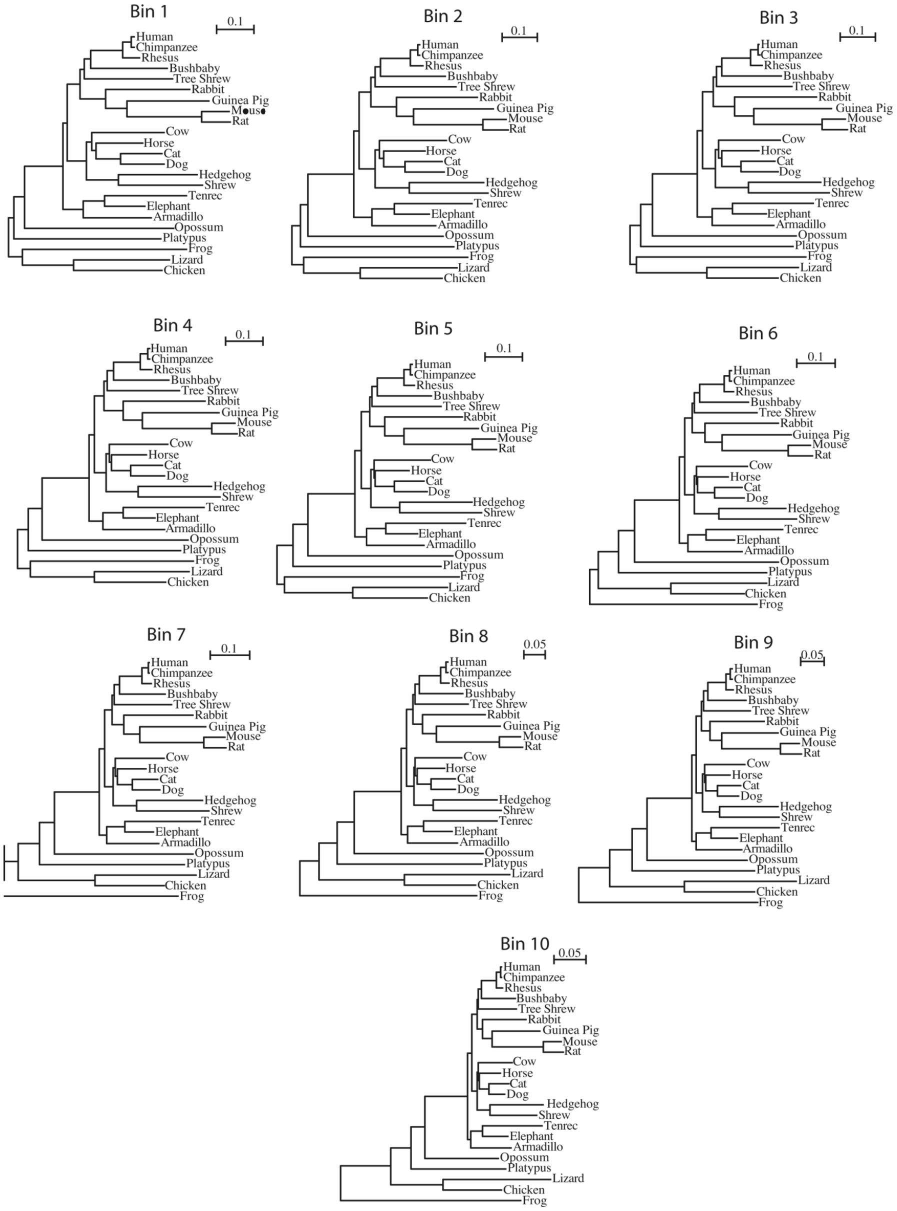
Supplemental Table 2. Mammalian-specific miRNA families.

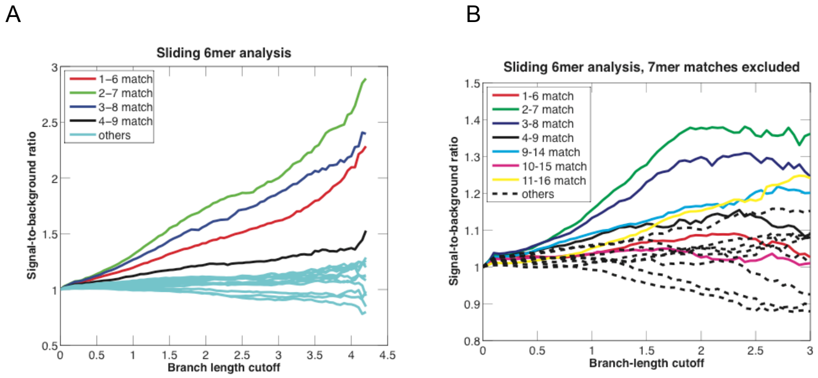
Seed + nt 8	Human miRNAs in family	8mer signal-to-background ratio in placental mammals (cutoff 0.85)
CAGGUGA	miR-125a-3p	0.61
CGGAUCC	miR-127-3p	2.94
GUGACUG	miR-134	0.85
CUCCAUU	miR-136	0.84
AGGUUAU	miR-154	0.93
GGAGAGA	miR-185	1.10
UCACCAC	miR-197	1.24
AGGAGCU	miR-28-5p;miR-708	1.40
AGGGUUG	miR-296-3p	0.68
AUGUGGG	miR-299-3p	0.85
GCAUCCC	miR-324-5p	0.69
UGGCCCU	miR-328	1.14
CUCUGGG	miR-330-5p;miR-326	1.80
CAAGAGC	miR-335	1.67
CCCUGUC	miR-339-5p	1.01
UAUAAAG	miR-340	1.46
CUCACAC	miR-342-3p	0.69
GUCUGCC	miR-346	1.10
UAUCAGA	miR-361-5p	1.34
ACACACC	miR-362-3p;miR-329	1.33
CCUGCUG	miR-370	1.29
CUCAAAC	miR-371-5p	0.71
UAUAAUA	miR-374a;miR-374b	0.88
UCAUAGA	miR-376a;miR-376b	1.14
ACAUAGA	miR-376c	0.69
UCACACA	miR-377	1.16
CUGGACU	miR-378;miR-422a	0.99
GGUAGAC	miR-379	1.46
AUACAAG	miR-381;miR-300	1.38
AAGUUGU	miR-382	1.16
UUCCUAG	miR-384	0.84
AUAUAAC	miR-410	1.05
UCAACAG	miR-421	1.19
GUCUUGC	miR-431	1.27
UCAUGAU	miR-433	1.30
UUUGCGA	miR-450a	0.59
GAGGCUG	miR-485-5p	0.92
AUCGUAC	miR-487b	0.77
UGAAAGG	miR-488	1.02
GUGGGGA	miR-491-5p	0.75
GAAACAU	miR-494	0.89
AACAAAC	miR-495	1.28
GAGUAUU	miR-496	1.01
GACCCUG	miR-504	1.25
GUCAACA	miR-505	1.29
GAGAAAU	miR-539	1.16
GUGACAG	miR-542-3p	1.14
AACAUUC	miR-543	1.38
UUCUGCA	miR-544	1.28
UGUUGAA	miR-653	0.85
UUGUGAC	miR-758	1.27
UGCCCUG	miR-874	1.30
GGAUUUC	miR-876-5p	1.01

Supplemental Table 3: miRNA families with intermediate conservation. These families were not found in enough non-mammalian genomes to be placed in the broadly conserved set (Supplemental Table 1), yet they were found in too many non-placental animals to be part of the mammalian-only set (Supplemental Table 2).

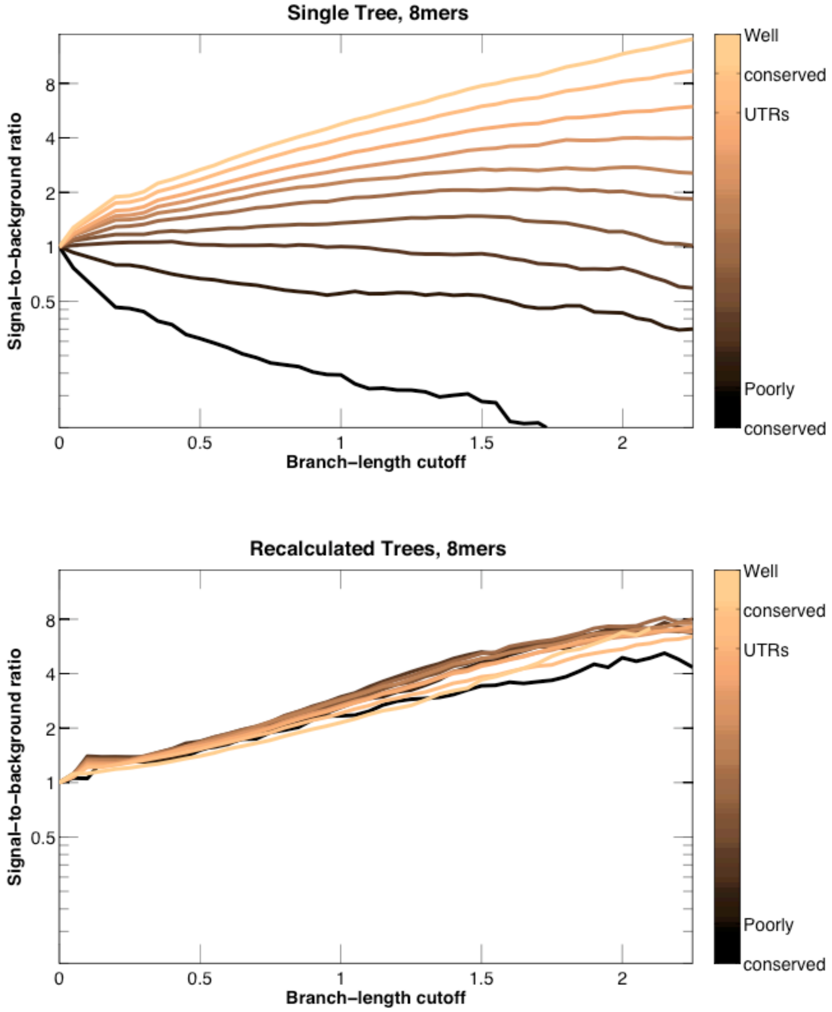
Seed + nt 8	Human miRNAs in family	8mer signal-to-background ratio in placental mammals (cutoff 0.85)	Notes
CUGGCUC	miR-149	2.14	non-placental mammal conservation
AAAGAAU	miR-186	1.06	found in opossum, platypus
GAGGUAU	miR-202	4.30	seed changes in mammals and in fish
AAGUCAC	miR-224	1.46	non-placental mammal conservation
AAAGCUG	miR-320	2.23	non-placental mammal conservation
AGUAGAC	miR-411	1.79	non-placental mammal conservation
AUGACAC	miR-425	1.00	found in opossum, platypus, lizard
UGCAUAU	miR-448	1.91	non-placental mammal conservation
CCUGUAC	miR-486-5p	1.35	non-placental mammal conservation
AGCAGCG	miR-503	5.45	non-placental mammal conservation
CGACCCA	miR-551a;miR-551b	1.67	found in opossum, platypus, chicken
AAUUUUA	miR-590-3p	0.76	found in platypus
UUGUGUC	miR-599	1.01	found in chicken, platypus, opossum
CCGAGCC	miR-615-3p	1.72	chicken, platypus, lizard alignment
CAGGAAC	miR-873	1.06	found in opossum, platypus
AUACCUC	miR-875-5p	0.97	found in opossum, platypus, lizard

Supplemental Figure 1. Trees for the ten UTR bins, with bin 1 being least conserved and bin 10 being most conserved.





Supplemental Figure 2. A systematic analysis of matches to each 6-nt segment across the 87 broadly conserved miRNA families. (*A*) Analysis performed without excluding those sites that also possessed canonical 7mer matches. Comparison to panel B indicates that much of the preferential conservation is due to overlap with parts of larger, canonical sites. (*B*) Analysis performed excluding those sites that also possessed canonical 7mer matches, revealing no other segment with appreciable enrichment in conservation.



Supplemental Figure 3. Reduced variability of signal-to-background ratios when using UTR bin-specific trees. Top panel: Analysis performed using a single tree, which was estimated using all UTRs. UTRs were divided into ten equally populated bins, based on their conservation rates, and the signal-to-background ratio for 8mer sites matching the 87 broadly conserved miRNAs is plotted separately for each bin. For each UTR bin, the fraction of conserved sites (the signal) was divided by the fraction of conserved controls (the background), using the same background estimate for all ten bins, which was the overall background, estimated using all UTRs. Bottom panel: As above but signal and background were calculated for each UTR conservation bin with a unique tree (Supplemental Fig. 1), estimated using only the UTRs from that bin. In other words, sites were compared only with background in the same UTR bin. Note that the bin that deviates most at high cutoffs is the one with the most poorly conserved UTRs. Because at high cutoffs this bin had too few conserved sites for reliable signal-to-background determination (less than one site per miRNA family conserved at branch-length cutoffs exceeding 1.5), its deviation at high cutoffs is not informative, and even if it were informative, it would involve too few sites to be of concern.