# Combinatorial code analysis for understanding biological regulation

Peng Jiang

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Computer Science
Adviser: Mona Singh

September 2013

# Abstract

An important mechanism to achieve regulatory specificity in diverse biological processes is through the combinatorial interplay between different regulators, such as amongst transcription factors (TFs) during transcriptional regulation or between RNA binding proteins (RBPs) and microRNAs (miRNAs) during transcript degradation control. To advance our understanding of combinatorial regulation, we developed a computational pipeline called CCAT (Combinatorial Code Analysis Tool) for predicting genome-wide co-binding between biological regulators.

In the first part of this thesis, we applied CCAT to the *D. melanogaster* genome to uncover cooperativity amongst TFs during embryo development. Using publicly-available TF binding specificity data and DNaseI chromatin accessibility data, we first predicted genome-wide binding sites for 324 TFs across five stages of *D. melanogaster* embryo development. We then applied CCAT in each of these developmental stages, and identified from 20 to 60 pairs of TFs in each stage whose predicted binding sites are significantly co-localized. Several of the co-binding pairs we found correspond to TFs that are known to work together. Further, pairs of binding sites predicted to cooperate were found to be consistently enriched in their evolutionarily conservation and their tendency to be found in regions bound in relevant ChIP experiments. Finally, we found that TFs tend to be co-localized with other TFs in a dynamic manner across developmental stages.

In the second part of this thesis, we applied CCAT to explore whether RBPs and miRNAs cooperate to promote transcript decay. We concentrated on five highly conserved RBP motifs in human 3′UTRs. A specific group of miRNA recognition sites were enriched within 50 nts from the RBP recognition sites for PUM and UAUUUAU. The presence of both a PUM recognition site and a recognition site for preferentially co-occurring miRNAs was associated with faster decay of the associated transcripts. For PUM and its co-occurring miRNAs, binding of the RBP to its recognition sites was

predicted to release nearby miRNA recognition sites from RNA secondary structures. Overall, our CCAT analyses suggest that a specific set of RBPs and miRNAs work together to affect transcript decay, with the release of miRNA recognition sites via RBP binding as one possible model of cooperativity.

Our pipeline provides a general tool for identifying combinatorial cooperativity in biological regulation. All generated data as well as source code are available at: `http://cat.princeton.edu`.

# Acknowledgements

I would devote my sincerest gratitude and appreciation to my advisor Mona Singh and Hilary Coller. Without their efforts, my life would not be possible.

My advisor Mona Singh is one of the smartest people I ever met. She has been so patient and insightful on leading me across the journey of learning scientific research. At the very beginning, I had no research experience at all. She helped me to set up my projects, teaching me how to design control analysis and how to figure out solutions for the problems I encountered. Especially at several moments I treated some analysis artifacts as good results, she was so deep and sharp to point out the pitfalls and helped me to find out the final solution.

For Prof. Hilary Coller, by her deep insight and expertise of molecular biology, she greatly changed my view of thinking on computational biology. At each of our discussions, she trained me to think about the fundamental mechanism from the very bottom level of biochemistry, and how to figure out the underlying model in the simplest form. She also took great effort in training me, a computer science major, to do molecular biology experiments. This really helped me to better understand biology and better communicate with biologists.

I would also thank to Olivier Elemento for teaching me basic technique on statistical analysis and bioinformatics related programming in my first year. Also, thanks so much to Adam Evertts for tolerating my consistently bugging him to ask about details of experimental protocols. I would also devote my thanks to members of Singh Lab and Coller Lab for all the great discussions and their friendship.

My PhD study was supported by Princeton University and NSF ABI-0850063 and NIH/NIGMS R01 GM076275. Chapter 3 presented in this thesis has already been published [1].

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Regulation of eukaryotic gene expression

In virtually all biological processes, gene expression patterns are regulated in a time or space dependent manner [2, 3]. Transcriptional and post-transcriptional regulation are two basic layers of gene expression regulation (Figure 1.1).

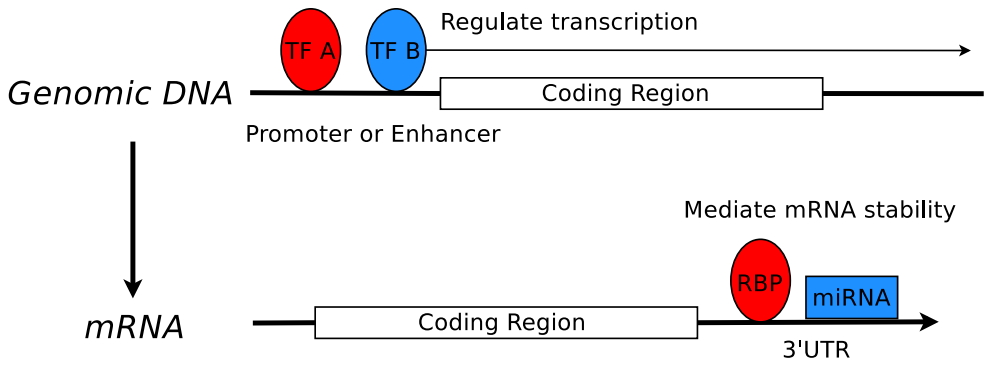



Figure 1.1: Regulation of gene expression. Levels of mRNA in eukaryotic cells are controlled by transcriptional and post-transcriptional regulation. In transcriptional regulation, transcription factors bind either promoters proximal to coding genes or enhancers distant from coding genes and modulate mRNA transcription. In post-transcriptional regulation, RNA binding proteins and microRNAs bind the 3 ′untranslated regions (3′UTRs) of mRNA transcripts and modulate their stability.

Transcription factors (TF) are a major class of regulatory proteins and they bind specific short DNA sequences (motifs) in gene promoters or enhancers [2, 4, 5, 6]. The binding of TFs, especially their combinatorial cooperativity, lead to the regulated patterns of gene transcription exhibited in diverse biological processes, including embryo development and cell differentiation [2].

After a gene is transcribed, post-transcriptional regulation can modulate the stability of the mRNAs and its translation efficiency [7, 8]. The two major regulators involved are microRNAs (miRNA) and RNA binding proteins (RBP). miRNAs are short RNAs approximately 21-23 nucleotides in length that generally bind their targets through complementary pairing in a short 7 bp seed sequence. This binding leads to translational repression or transcript decay [9, 10, 11, 12]. RBPs can affect transcript stability by binding to recognition sequences within 3 ′untranslated regions (3′UTRs); this either increases the degradation of target transcripts [13, 14, 13, 15, 16, 17, 18, 19, 20] or stabilizes the targeted message [21].

The final mRNA level in the cytoplasm is determined by the activity of both transcriptional regulation and post-transcriptional regulation (Figure 1.1).

## 1.2 Combinatorial regulation in biological processes

Recent progress in profiling the binding landscape of TFs has revealed that a single TF can bind thousands or tens of thousands regions in a genome [22, 23, 24], and that the binding of a single TF cannot achieve the complex and precise control of gene expression exhibited in organisms [25]. Similarly, at the level of transcript degradation control, the same RBP can regulate the stability of transcripts in totally different manners [26]. Thus, the existence of a single biological regulator is far from achieving the regulated outcome of the entire biological system.

Combinatorial cooperativity amongst different biological regulators is a central mechanism by which regulatory specificity is achieved (reviews, [2, 4, 27, 5, 28]). As an example, the Drosophila TF *dl* works with the TF *twist* to determine neurogenic activity in early embryo development [29]. Binding sites for *dl* and *twist* are observed close to each other in the enhancer regions of several genes and across several Drosophila species [30, 31]. As an example in transcript degradation control, the Pumilio (PUM) RNA binding protein has been proposed to modulate miR-221/222 activity on the 3'UTR of cyclin-dependent kinase inhibitor p27Kip1 by binding to sequences that can hybridize with miRNA recognition sites and thereby make them more accessible for the RISC complex [28]. Thus, the repression efficiency of miR-221/222 will depend on PUM expression.

In many studies of combinatorial cooperativity, it has been observed that certain pairs or groups of TFs tend to collaborate not only in a single region, but across many promoter or enhancer regions, following certain rules of binding motif positioning [30, 31, 32, 33, 29]; similarly, certain pairs of RBP and miRNAs recognition sites tend to co-localize with each other on certain 3'UTRs [34, 35, 36]. Thus, for a large class of combinatorial regulation cases, the problem can be abstracted as that of uncovering proximal recognition sites that are associated with the functional output of a biological process (Figure 1.2).

## 1.3 Computational challenges of finding combinatorial cooperativity

Given the prevalence and importance of cooperativity, several computational approaches have been developed to analyze genome-scale experiments in order to reveal interactions between TFs based upon ChIP experimental datasets [37, 38, 24, 39, 40, 41] or TF sequence motifs [42, 43, 44]. Despite these initial efforts, many com-

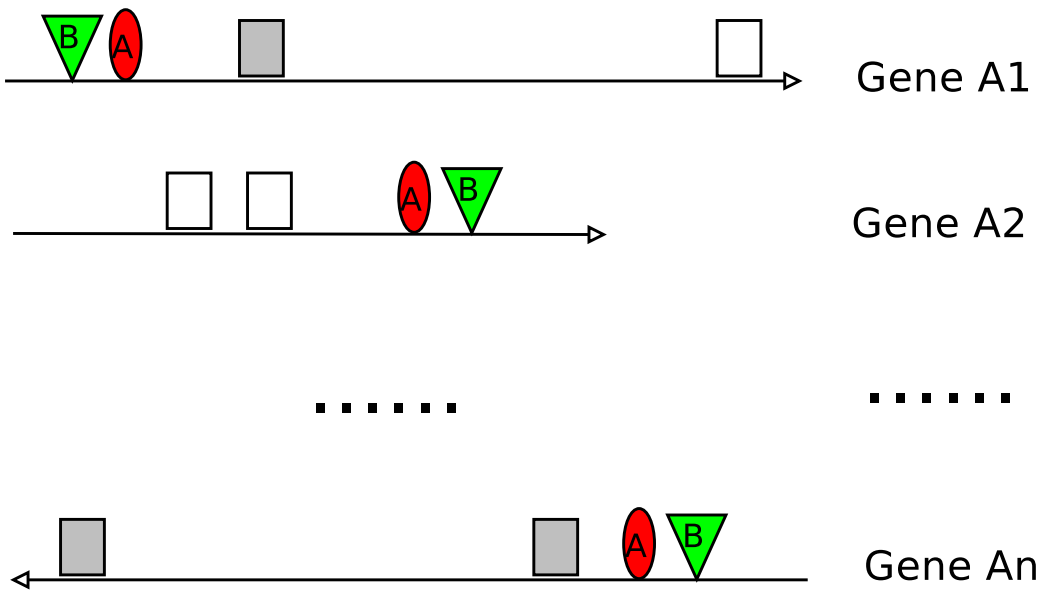



Figure 1.2: Common pattern of combinatorial regulation. For many examples of known combinatorial regulation, the recognition sites of the involved regulators are colocalized with each other over many genomic regions.

putational difficulties remain in uncovering TF interactions across diverse biological processes.

A common first step for all of the previous studies is that of finding the binding sites of TFs. Currently, it is estimated that there are about 753 TFs for fruitfly [45] and 1700 to 1900 TFs for human [46]. Considering the huge diversity of biological conditions, it is infeasible to profile hundreds of TF ChIP profiles under each condition of interest; thus ChIP based methods are not able to uncover TF cooperativity at the genome scale in a condition specific manner.

For the other class of studies, TF binding sites are computationally predicted in gene promoters from hundreds of available position weight matrices (PWM). However, given the short lengths of binding sites for most TFs and the degeneracy in sites that a TF can bind, matches to PWMs are frequently found by chance in genomic regions. Moreover, for higher eukaryotes, only a small fraction of TF binding sites are located in regions proximal to genes and a larger number of binding sites are

located further away and presumably regulate transcription by higher order genome organization [47, 48, 49, 50]. Thus, predicting binding sites with PWMs in promoters may suffer from reliability issues and may miss the majority of binding sites in higher eukaryotes.

For post-transcriptional regulation, previous computational approaches have initially attempted to find interactions between RBPs and miRNAs on a case by case basis. As an example, the miR-30 motif was found to be enriched on PUM targets [51] and miR-410 binding sites were found to be close to PUM recognition sites [52]. However, to the best of our knowledge, there has not been a genome scale screening of interactions between RBPs and miRNAs. Further their combinatorial impact on transcript stability is largely unknown. This is mostly due to the lack of known binding specificities for RBPs. For most RBPs, both their sequence motifs and experimental binding profiles are largely unknown.

## 1.4 Our contributions

In this thesis, we develop a computational pipeline named CCAT (Combinatorial Code Analysis Tool) to uncover combinatorially interacting motif pairs. CCAT is designed to overcome difficulties in previous studies, such as the requirement of having ChIP datasets for the condition of interest or limited searching of binding sites within promoter regions. While our methodology is general, we concentrate our efforts on two biological applications. First, we study the process of Drosophila embryo development and uncover genome-wide cooperativity between TFs across the five progressive embryo stages. Second, we study the cooperativity between RBPs and mammalian miRNAs in transcript degradation, and uncover interactions between PUM and a specific group of miRNAs. In both cases, we apply CCAT to profile genome scale cooperativity between the involved biological regulators. The predicted interactions

capture previously known cases of cooperativity and statistical analysis shows their consistency with other genomic datasets such as ChIP binding profiles and mRNA decay profiles.

Our front-to-end pipeline CCAT provides tools for clustering and manipulating regulatory motifs, predicting evolutionarily conserved regulatory motif sites, and uncovering preferentially co-occuring binding motifs. All source code is released as a Unix software package at `http://cat.princeton.edu`.

# Chapter 2

# Cooperativity between transcription factors for Drosophila embryo development

## 2.1 Introduction

Transcriptional regulation controls a diverse range of biological processes, from development to response to external stimuli. Recent progress in profiling the binding landscape of transcription factors (TF) has revealed that a single TF can bind thousands or tens thousands of regions in a genome [22, 23, 24], and it is clear that the binding of a single TF cannot achieve the complex and precise control of gene expression exhibited in organisms [25]. Combinatorial cooperativity amongst TFs is a central mechanism by which regulatory specificity is achieved (reviews, [2, 4, 27, 5]). Distinct modes of cooperativity between TFs have been identified, including physical interactions between TFs for proximal co-binding [53], collaborative competition of two TFs with a nucleosome for DNA binding [54], and changes in the local conformation of DNA by one TF's binding to assist the binding of other TFs [55, 56]. Further,

a TF may have different sequence specificities when interacting with different cofactor TFs [57, 58, 59, 60].

In many studies of TF cooperativity, it has been observed that certain pairs or groups of TFs tend to collaborate not only in a single region, but across many promoter or enhancer regions, following certain rules of binding motif positioning [30, 31, 32, 33, 29]. For example, the yeast TF *MCM1* interacts with several cofactor TFs to combinatorially regulate cell cycle and mating [53, 32]. Its binding motif is found near those of its cofactor TFs in many regulons and in several yeast species [32]. Another example comes from the Drosophila TF *dl*, which works with the TF *twist*. Binding sites for *dl* and *twist* are observed close in the enhancer regions of several genes and across several Drosophila species [30, 31], and binding motifs for other TFs, including *Su(H)*, also co-locate with them [29].

Given the prevalence of TF cooperativity, several computational approaches have been developed to analyze genome-scale experiments in order to reveal interactions between TFs. For example, ChIP experiments for TFs have been analyzed to find overlapping binding profiles [37, 38, 24] and to uncover enriched TF motifs corresponding to cofactors in distinct biological contexts [39, 40] or among related species [41]. Several studies have also computationally predicted TF binding sites in gene promoters from available position weight matrices (PWM), and used these to predict combinatorial TF interactions [42, 43, 44]. A common first step for all of these methods is to collect genomic binding sites for TFs. Computational methods using ChIP experiments clearly require the availability of ChIP datasets [37, 38, 39, 40, 41]; however, ChIP data is context-specific, and given the large number of TFs in higher eukaryotes(e.g., 753 in *D. melanogaster* [45] and 1700 to 1900 in *H. sapiens* [46]), it is currently prohibitive to perform these experiments for all TFs in each biological context of interest. On the other hand, the large numbers of TFs in model organisms for which binding specificities are known (e.g., 364 in *D. melanogaster* and 722 in *H.*

*sapiens*) provide a promising means for predicting binding sites for a significant fraction of TFs at the genome-scale. However, given the short lengths of binding sites for most TFs and the degeneracy in sites that a TF can bind, matches to PWMs are frequently found by chance in long genomic regions. To make higher quality predictions, binding sites are typically required to be conserved across organisms, and searched for within regions upstream of genes [61, 62] or within a small set of experimentally verified enhancer regions [63, 64]. However, for higher eukaryotes, only a small fraction of TF binding sites are located in regions proximal to genes and a larger number of binding sites are located further away and presumably regulate transcription by higher order genome organization [47, 48, 49, 50]. For example, less than 20% of the *D. melanogaster* TF ChIP binding regions included in the modENCODE project overlap gene promoter regions [37]. Thus, predicting binding sites only within promoter regions may miss the majority of regulatory binding sites in higher eukaryotes.

Recently, DNaseI digestion has been coupled with massively parallel sequencing to measure genome-wide chromatin accessibility and the occupancy patterns of DNA binding proteins [65, 66, 67, 68]. The binding of multiple regulators within a genomic region will increase its local chromatin accessibility to DNaseI nuclease digestion. Thus, finding DNaseI hypersensitive sites has proven to be a powerful means for mapping regulatory binding sites without requiring prior knowledge of specific DNA binding proteins. DNaseI digestion patterns have already been measured at the genome scale by high throughput sequencing for the five stages of Drosophila embryo development [66, 67] as well as for 125 diverse cell and tissue types for human [68]. Thus, the rapid progress of DNaseI experiments, when combined with predictions of TF binding sites, provides new opportunities for profiling genome-wide condition specific TF occupany [69, 70] as well as TF cooperativity under different conditions.

In this study, we develop a computational pipeline CCAT (Combinatorial Code Analysis Tool) to uncover combinatorially interacting motif pairs, which is designed to

overcome difficulties in previous studies, including the requirement of ChIP datasets under the studied condition or limited searching within promoter regions. We concentrate our efforts on the process of Drosophila embryo development, which involves extensive cooperativity amongst many TFs [2]. We leverage known binding site specificities for hundreds of *D. melanogaster* TFs [71, 72, 73, 23, 22, 74, 75, 76, 61, 77, 78, 79], full genome sequences for 12 *Drosophila* species, and genome-scale chromatin accessibility data as determined by DNaseI experiments [67, 66] across five conditions of embryo development. We first predict conserved binding sites for 324 TFs in these five conditions by focusing on accessible genomic regions in each condition. We show that our predictions exhibit good agreement with ChIP experiments, and are comparable in quality to high-throughput ChIP experiments, as judged via functional measures. We next search for pairs of TF regulatory motifs whose binding sites are significantly co-localized, by comparing real occurrences of binding motifs with randomized controls. We find that pairs of binding sites predicted to cooperate are consistently enriched in their evolutionarily conservation and in their tendency to be found in regions bound in relevant ChIP experiments. Further, our predicted combinatorial pairs tend to be used in specific stages of embryo development, which is consistent with the dynamic nature of combinatorial regulation. The source code for our front-to-end pipeline, from predicting evolutionarily conserved genomic binding sites for TFs to uncovering preferentially co-occuring binding motifs, is available online at `http://cat.princeton.edu`.

## 2.2 Results

### 2.2.1 Predicting TF binding sites in accessible genomic region

We collected 712 position weight matrices (PWM) representing 364 different *D. melanogaster* transcription factors (TFs) from several resources [71, 72, 73, 23, 22, 74, 75, 76, 61, 77, 78, 79]. We searched the 12 Drosophila genomes for matches to these PWMs using the fimo algorithm from the MEME package [80, 81]. For each binding site in *D. melanogaster*, we next calculated conservation scores based upon nearby matches in the 11 other Drosophila species [61, 82]. For each TF, only the top 20% most conserved binding sites were selected for further analysis. We next considered only conserved binding sites in *D. melanogaster* within the most accessible genomic regions, as determined by DNaseI experiments across five embryo development stages of Drosophila (stages S5, S9, S10, S11 and S14, corresponding to 3, 4, 5, 6, and 11 hours after fertilization) [67, 66]. For each stage, all genomic DNaseI accessibility scores were sorted from largest to smallest and the top 5% of scores were used for binding site selection.

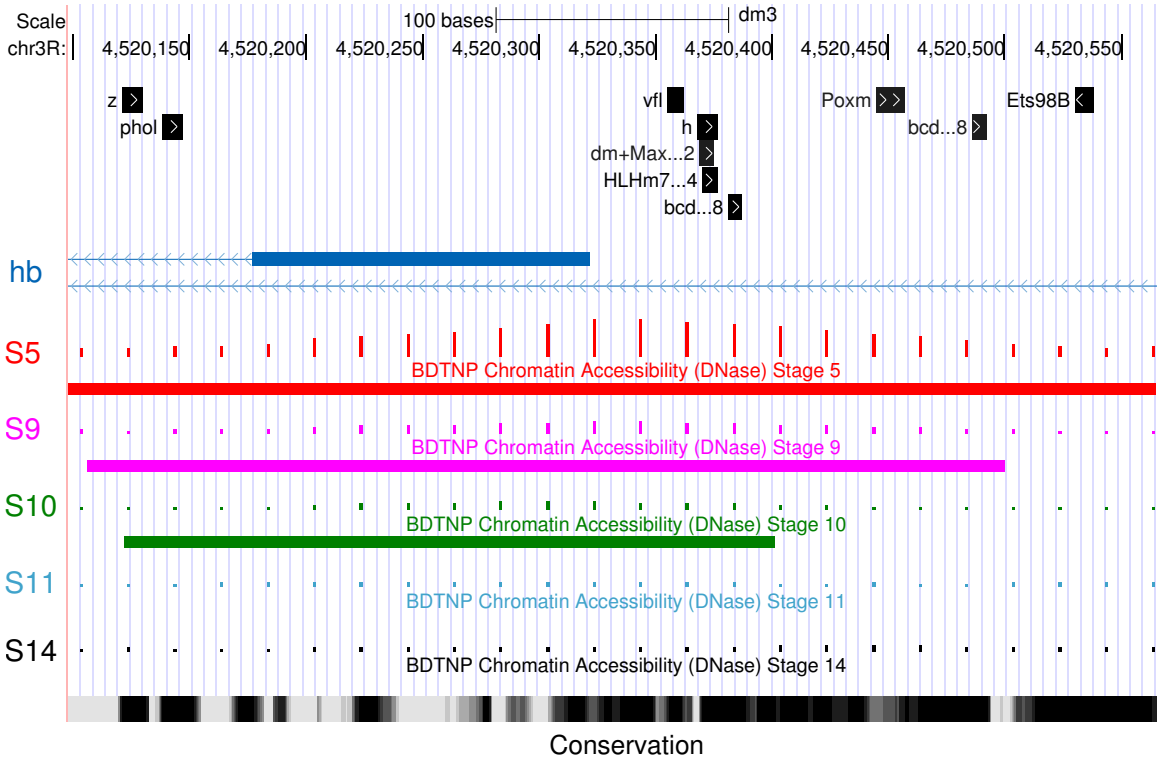
We noticed that many of the collected PWMs are very similar to each other (e.g., homeodomain proteins cluster into groups of TFs with similar binding specificities [72]), and thus have largely overlapping sets of predicted genomic binding sites. To address this, we grouped TFs with similar PWMs together using hierarchical clustering [83]. This resulted in 198 TF PWM clusters, and 44 of these contained two or more TFs. Binding site predictions for any TF in one of these clusters were assumed to be putative predictions for the other TFs in the same cluster. All 44 clusters with multiple TFs were assigned indices from 0 to 43, and were referred to by that index along with a representative TF contained within the cluster. As one example, we visualized binding site predictions near the transcription start site of *hb* (Figure 2.1A).

While several binding site predictions correspond to a single TF, a few correspond to predictions for a cluster of TFs. For example, binding sites for TFs in a cluster including *bcd* are found; this cluster, referred to by *bcd* and the index 8, contains 4 different TFs including *bcd*, *oc*, *Ptx1* and *Gsc* (Figure 2.1B).

In order to assess the quality of our binding site predictions, we used the ChIP datasets we collected from diverse sources [23, 22, 84, 77, 78, 79, 85, 86]. Among 53 TFs with at least one associated ChIP dataset, 39 of them are included in our TF binding site predictions. For each of these TFs, we computed the percent of its predicted binding sites that are located within ChIP bound regions. We found that by requiring conserved sites to be within a DNaseI accessible regions in at least one stage, a larger fraction of binding sites are located within ChIP regions than are when considering conserved sites over all genomic regions (Figure 2.2A and Figure 2.3).

We further compared our binding site predictions with other large-scale regulatory networks. The fly modEncode project released two physical regulatory networks: `motif_net` and `ChIP_net`. The `motif_net` network is computationally predicted by finding conserved binding sites within gene promoters [61, 84]. The `ChIP_net` network is determined via ChIP experiments [84]. We also constructed regulatory networks from BDTNP [23, 22] by assuming that a TF regulates a gene if there is a ChIP binding region within 2000 nts from the transcription start site. For our binding site predictions, we built six networks using predictions restricted to either DNA accessible regions in a specific stage or over the whole genome. Our regulatory network contains a significantly larger number of TFs (Table 2.1) than these previous networks.

A. Conserved binding sites in chromatin accessible region




B. Example of merged PWM cluster 8

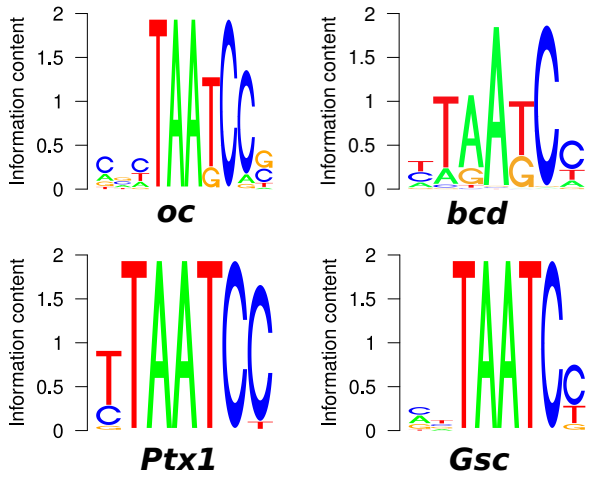



Figure 2.1: Predicted conserved TF binding sites in chromatin accessible regions. TF binding sites were predicted across the entire genome and overlapping sites were merged based upon the membership of TFs within uncovered clusters with similar binding specificities. For each motif instance, the local average DNaseI accessibility score are calculated for five embryo development stages [67, 66]. (caption continued on next page.)

Figure 2.1: (previous page) For each stage, only motif instances with DNaseI score larger than the top 5% of all scores were selected. (A) Examples of motif instances near the transcription start site of gene *hb*. Only predicted binding sites with conservation percentile score greater than or equal to 0.8 are selected. The TF names are shown over their motif instance position. For merged binding sites, only the name of one representative TF member is shown, followed by the index number of that TF cluster. DNaseI scores for five developmental stages are shown proportional to the height of vertical bars. Regions within the top 5% of scores are shown for each stage with horizontal bars (S5, S9 and S10 in this example). Conservation scores are shown on the bottom with denser colors representing higher conservation. (B) Example of TF cluster 8. TF members and their motif logos are shown.

| **DataSet** | **Network** | **#Regulators** | **#Targets** | **#Interactions** |
|---|---|---|---|---|
| **modEncode** | motif_net | 104 | 10921 | 92978 |
| | ChIP_net | 79 | 12411 | 158571 |
| **BDTNP** | BDTNP | 24 | 8686 | 43243 |
| **CCAT** | All_site | 324 | 13155 | 379192 |
| | S5 | | 6503 | 67153 |
| | S9 | | 6597 | 68693 |
| | S10 | | 6087 | 62019 |
| | S11 | | 6510 | 72097 |
| | S14 | | 6631 | 71305 |

(a) Network level statistics

| **Network** | **#Regulators** | **#Targets** | **#Interactions** | **#Binding Sites** |
|---|---|---|---|---|
| All_site | 198 | 13090 | 271609 | 1188101 |
| S5 | | 6406 | 49195 | 64275 |
| S9 | | 6491 | 50494 | 70085 |
| S10 | | 5969 | 45707 | 60679 |
| S11 | | 6409 | 52952 | 74118 |
| S14 | | 6530 | 52792 | 74735 |

(b) Binding site level statistics

Table 2.1: Regulatory network sizes. (A) For each dataset, the number of regulators, targets and interactions are listed in each column. (B) Statistics for the dataset generated by CCAT, with TFs with similar PWMs clustered and their overlapping binding sites merged.

We also assessed the quality of our network using known functional annotations from the Gene Ontology [87]. We reasoned that the target genes of a TF should be involved in similar biological processes as it. For each GO biological process term annotating a TF, we computed an enrichment ratio by dividing the fraction of

genes annotated with that term within the TF's target genes versus the fraction of non-target genes annotated with that term. The top 20% conserved binding sites have similar GO enrichment measures as the modEncode computational `motif_net` network, whereas the ChIP experimental networks `ChIP_net` and `BDTNP` have better functional quality measures than purely computationally predicted networks (Figure 2.2BC). However, when restricting binding site predictions to be in the top 5% of DNaseI accessible regions, the GO enrichment measures of our binding site predictions were significantly improved in all five stages (Figure 2.2BC). There were 17 TFs profiled in all four datasets. When we compared over these 17 common TFs, our binding site predictions have similar functional enrichment ratios as the modEncode experimental ChIP network (Figure 2.2B and Figure 2.4). When all TFs included in each network were used to compute the enrichment measures for each network, our approach obtained a higher GO enrichment ratio than the modEncode ChIP experimental network (Figure 2.2C).

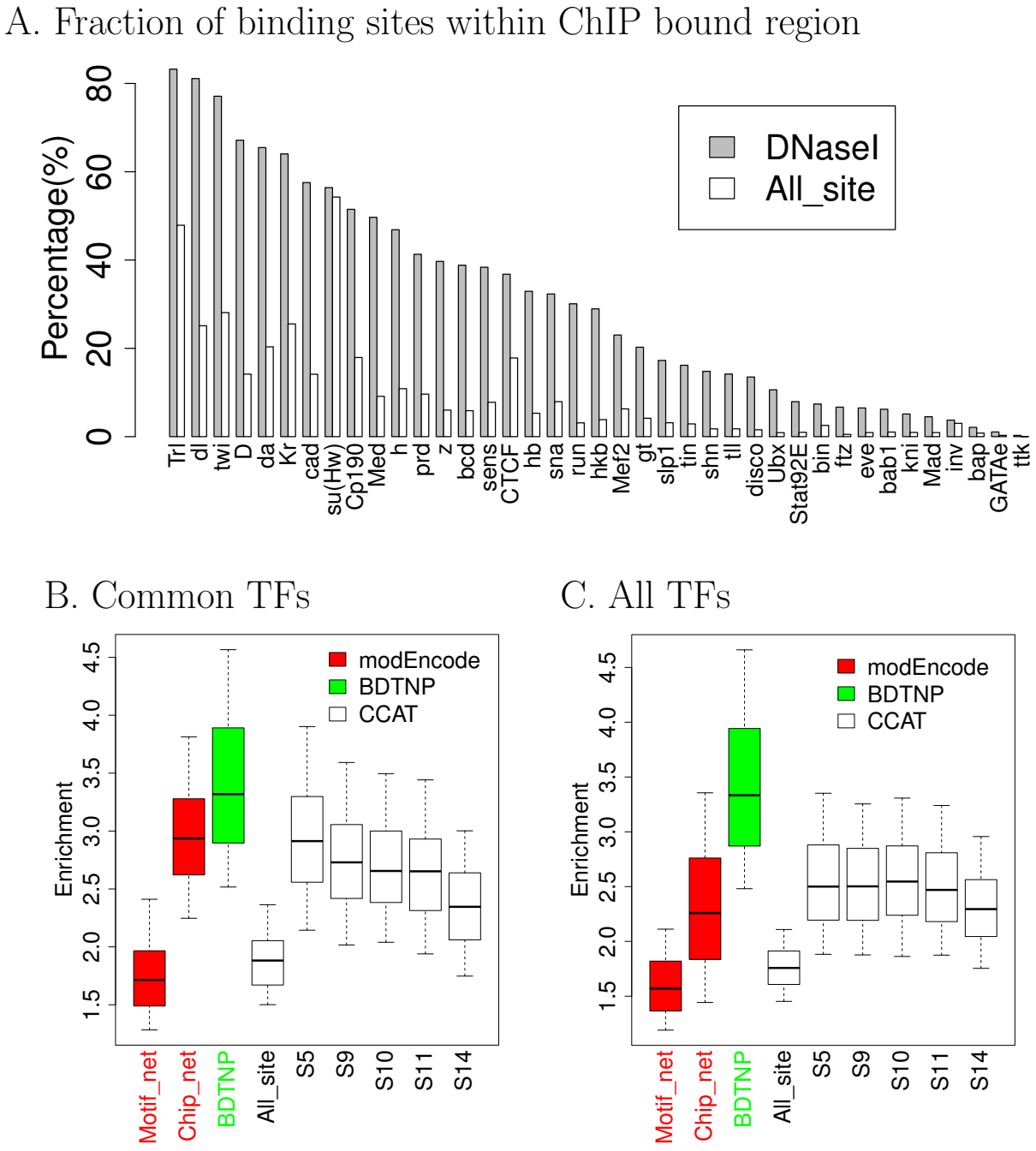



Figure 2.2: Predicted TF binding sites have quality comparable to ChIP experiment. (caption continued on next page.)

Figure 2.2: (previous page) (A) For each TF profiled with a corresponding ChIP dataset, the percent of predicted binding sites that are located within experimentally identified bound region was calculated. Only binding sites with conservation percentile scores greater than or equal to 0.8 were considered. White bars represent the percentages calculated with TF binding sites over the whole genome. Gray bars represent percentages calculated with binding sites with the top 5% of DNaseI scores in at least one stage. (B,C) Regulatory networks from TF to target genes were first built by connecting TFs to genes if binding sites are found within 2000nts of the transcription start site. For each GO term [87], the enrichment ratio among target genes was calculated as the fraction of target genes annotated with the term, divided by the fraction of non-target genes annotated with the term. For each TF, the GO enrichment ratios were calculated for all of its GO biological process annotations and visualized together by boxplots for each dataset. Motif_net and ChIP_net are two physical regulatory networks generated by modEncode [84]. BDTNP is the regulatory network constructed from BDTNP [23]. CCAT represents the networks generated by our computational predictions, constructed from conserved binding sites (conservation percentile score$\geq$0.8) over the entire genome or within DNaseI accessible regions of each stage. (B) Only the 17 TFs profiled by all datasets are considered. (C) All TFs profiled in each dataset are considered.

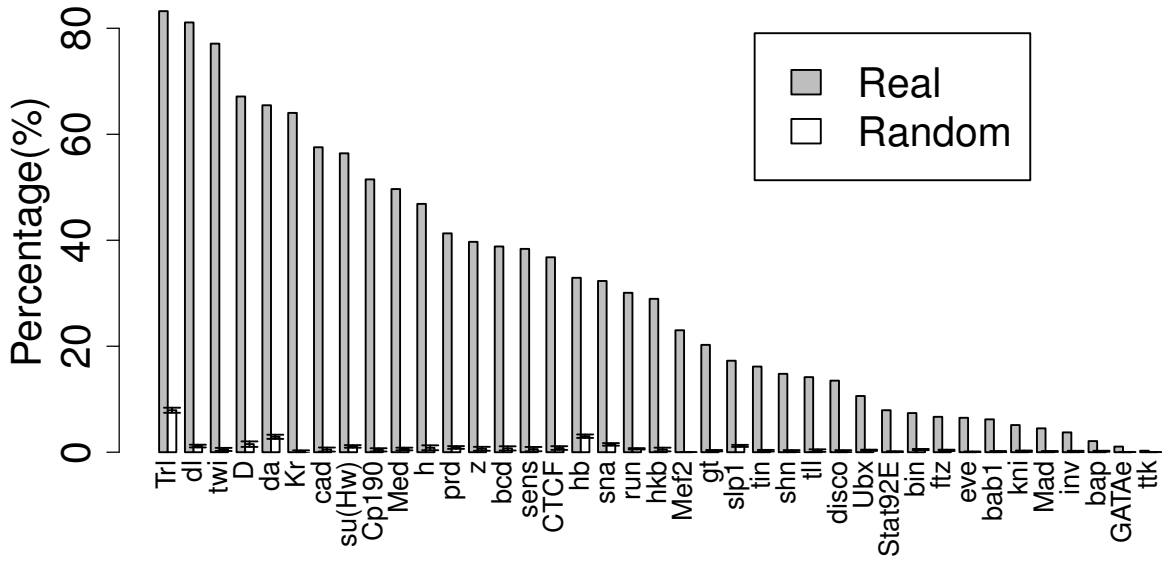



Figure 2.3: The CCAT predicted binding sites significantly overlap with ChIP experiments. We selected predicted binding sites within the top 5% of accessible DNaseI scores in at least one stage and having conservation percentile scores greater than or equal to 0.8. For each predicted binding site, the TF identity was randomly swapped with another TF if both of them were profiled in our collected ChIP datasets. The percent of binding sites in BDTNP ChIP bound region was calculated as Figure 2.2A. The standard deviations from 10 randomizations were represented by error bars.

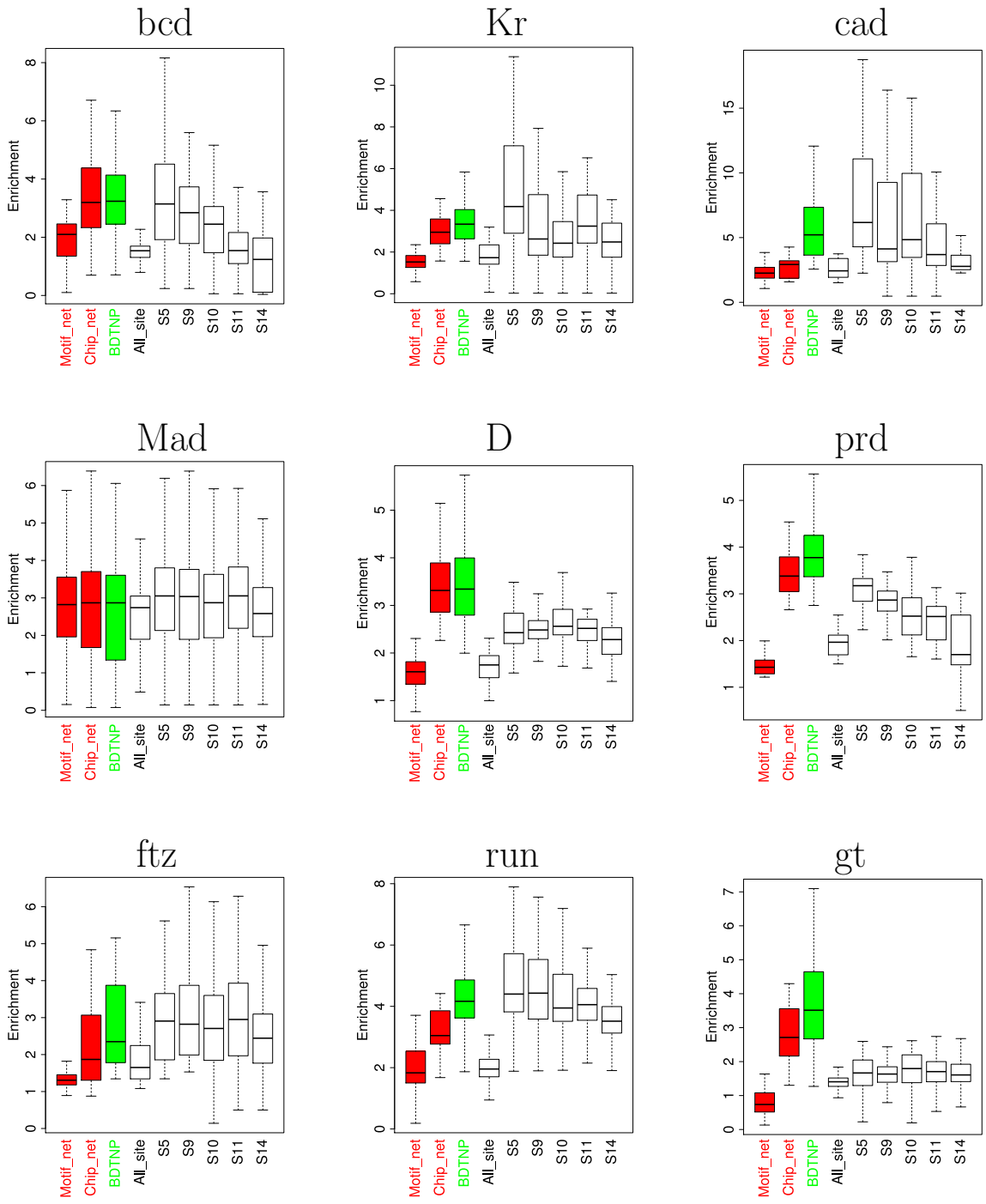



Figure 2.4: GO enrichment assessments for regulatory network targets. (caption continued on next page.)

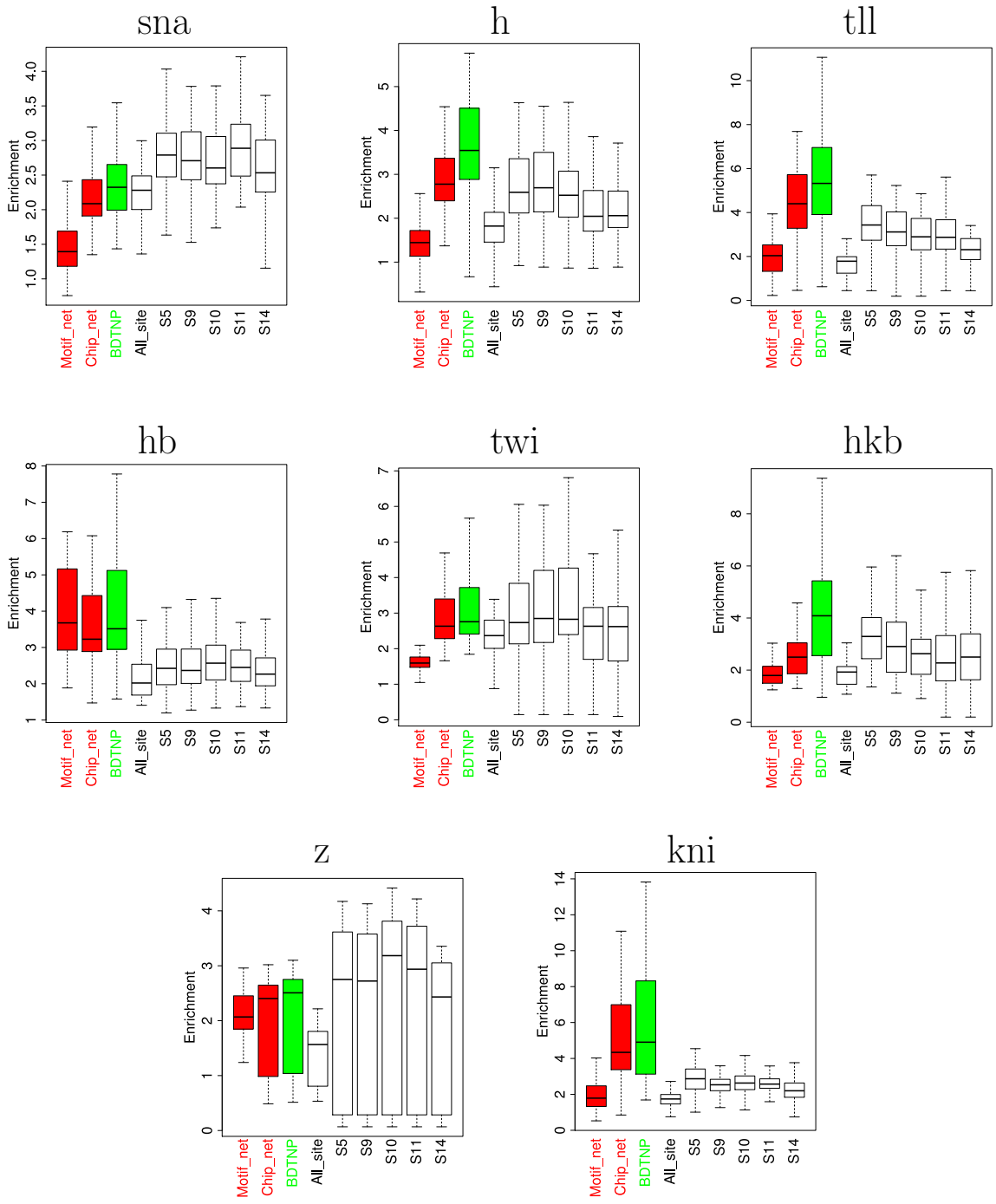



Figure 2.4: (previous page) For each TF and its annotated GO biological process terms, GO enrichment ratios among target genes were calculated and visualized as described in Figure 2.2B. Only 17 TFs that are included in all datasets are shown.

As an additional quality control, we utilized the Redfly regulatory network, a small, curated database of regulatory interactions in fly [88]. For each of the four

datasets, we computed the number of interactions that overlap those annotated in Redfly and compared this against the overlap found when the Redfly network is randomized by edge swapping [89]. The overlap enrichment is defined as number of overlapping interactions divided by the expected number of overlaps, as computed by averaging the number of overlapping interactions over 1000 edge-swapped [89] Redfly networks. We found our predicted regulatory network and the modEncode `motif_net` network consistently had higher enrichment levels than regulatory networks determined by ChIP (Figure 2.5); this finding is consistent with what was reported in the modEncode project [62].

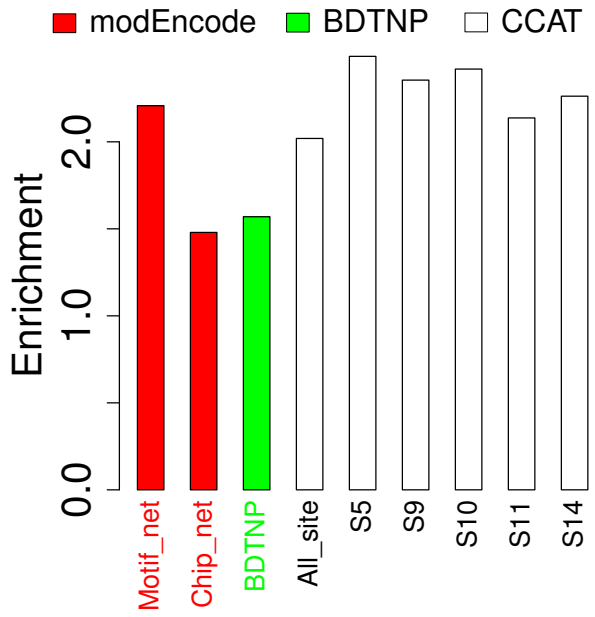



Figure 2.5: The CCAT regulatory network has high overlap with the Redfly dataset. 505 Redfly regulatory interactions were taken as a gold standard [88]. For each regulatory network, the overlap count was calculated with the real Redfly network and the randomized Redfly networks generated by edge swapping [89]. The enrichment ratio was defined as the real overlap count divided by the average overlap count from 1000 randomizations. Motif_net and ChIP_net are two networks integrated by modEncode [84]. BDTNP is derived from BDTNP ChIP experiments [23]. CCAT represents the networks predicted by our computational pipeline, which are constructed from conserved binding sites (conservation percentile score $\geq$0.8) over the whole genome or binding sites within DNaseI accessible regions for each developmental stage [67].

### 2.2.2 Dynamic usage of TF binding motifs in embryo development

The DNaseI accessibility data we used provides information about the dynamics of chromatin accessibility during embryo development [67, 66]. We utilized this dynamic information to determine whether binding site accessibility varies per TF across developmental progression. For each TF, we first computed its normalized degree in each stage as the number of its predicted accessible binding sites normalized by the total number of predicted accessible binding sites for all TFs [69]. We found that different regulatory motifs tend to vary in their degrees across different stages (Figure 2.6A). For example, *bcd* has larger fractions of the accessible binding sites at the early stages S5 and S9, but lower fractions at the later stages S10, S11 and S14 (Figure 2.6A).

For each TF, in order to check the significance of its degree variation across stages, we defined the variation ratio as the maximum normalized degree across five stages divided by the minimum normalized degree. For each motif instance, we randomly permuted the binary DNaseI accessibility determination across the five stages and counted the accessible binding sites in each stage. For each TF motif, the normalized degrees and variation measures across five stages were computed again for each randomization. We found the real data were consistently more abundant than randomized data for larger variation ratios (Figure 2.6B).

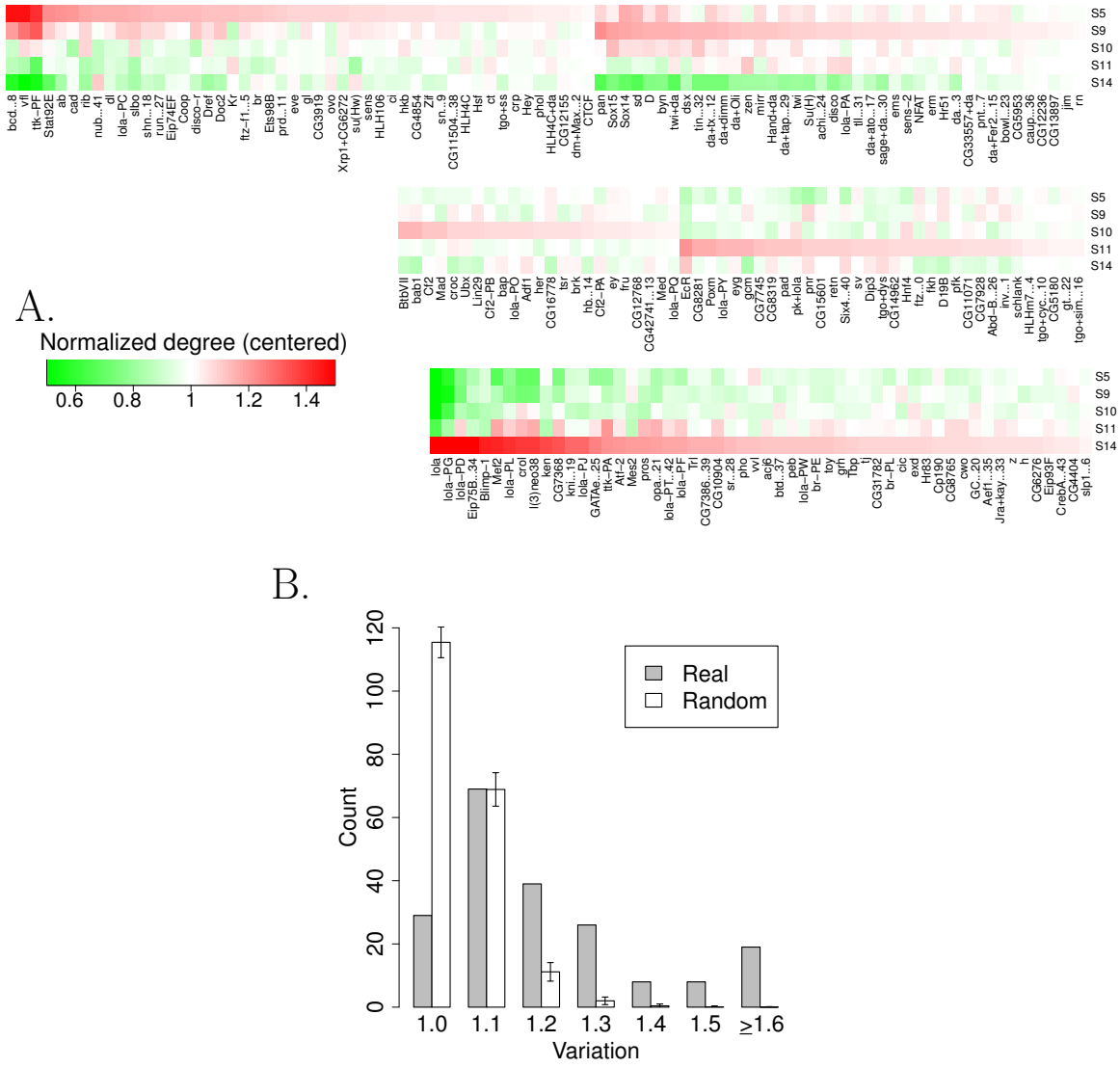



Figure 2.6: Stage specific usage of TF binding sites across embryo development. (A) For each regulatory TF motif, only predicted sites with conservation percentile score greater than or equal to 0.8 were considered. The normalized degrees in different stages were calculated as the number of binding sites normalized by the total number of binding sites over all TF motifs. For each regulatory motif, the centered normalized degree in each stage was computed by dividing the average degree across five stages and visualized by heatmap. (B) For each TF regulatory motif, a variation ratio was calculated as the maximum normalized degree among the five stages divided by the minimum normalized degree. As a random control, the DNaseI accessibility classifications for each predicted binding site were permuted across the five stages and the accessible binding sites were counted again to calculate the randomized normalized degrees and variations. Histograms for real and random variation ratios are shown, and the average and standard deviation values of random histograms were calculated from 1000 randomizations.

### 2.2.3 Finding combinatorial regulatory motif pairs

Based upon our stage-specific binding site predictions, we searched for pairs of regulatory motifs that show a co-localization enrichment based upon the frequency with which they occur within 100 nts of each other. For each pair of regulatory motifs, we first enumerated all predicted binding sites that fell within 1000 nts of each other (Figure 2.7A). For each enumerated pair, we assigned a weight between 0 and 1 by taking the minimum conservation percentile score between the two involved binding sites. Then, we classified all enumerated pairs into distance intervals corresponding to the number of nucleotides between them ($<$ 100 nts, 100–200 nts, etc.) and summed the weights that fell into each interval (Figure 2.8A). To estimate the expected weighted co-localization score for each pair of regulatory motifs, we permuted the identities of binding sites among TFs with similar base pair composition (Figure 2.7B and Figure 2.9), while keeping the genome position and conservation percentile score associated with each site fixed. For each pair of regulatory motifs, the distribution of distances between them were computed again (Figure 2.8A), and an empirical $P$-value for the motif pair co-localization was computed based upon the initial weighted count for motifs within 100 nts as compared to the weighted counts over 10,000 randomizations. FDRs for motif pairs were computed using the Benjamini-Hochberg procedure, and motif pairs with FDRs $\leq 0.05$ were determined to be co-localizing (Methods).

We ran the above procedure separately for each of the five studied stages and obtained 20 to 60 co-localizing regulatory motif pairs (Figure 2.8B and Figure 2.10). Several previously known examples of TF cooperativity were recapitulated in this set. For example, we found several TFs that co-localize with *vfl* (also known as *zelda*), a protein critical in embryo development [90]. In stage S5, we found that it co-localizes with *bcd*, which is known to be involved in early embryo development [91]. Similarly, previous studies showed that *dl* and *twi* cooperate in neurogenic enhancers that direct gene expression in the early embryo [29, 92], and we found that binding sites for *dl* and

*twi* co-localized with each other in stage S5 (Figure 2.8B). In additional to capturing known cooperativity among factors that play a role in development, our pipeline also predicted co-localization between *Hsf* and *Tbp* binding sites (Figure 2.8B); these TFs were previously found to physically interact with each other and cooperatively bind heat shock promoters [93].

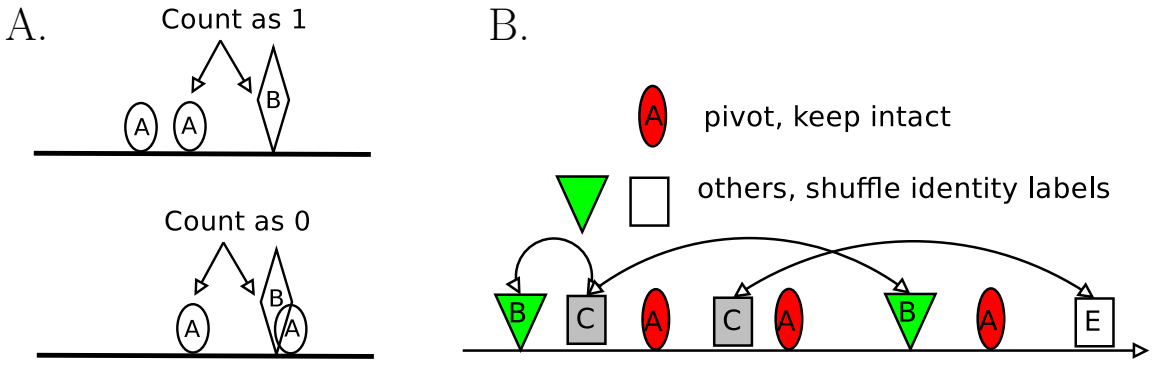
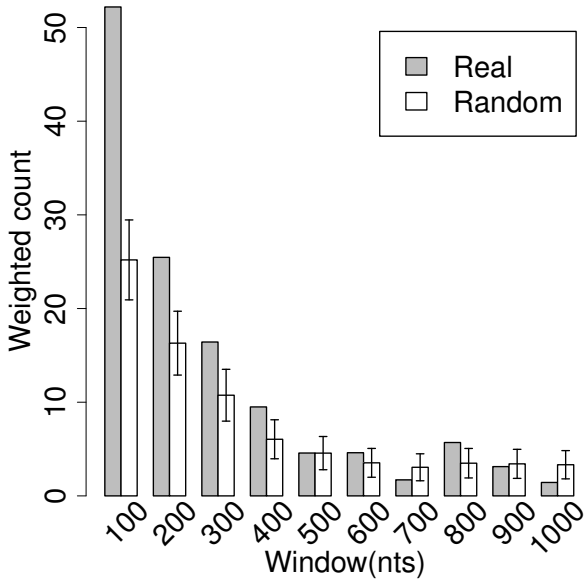



Figure 2.7: Finding combinatorial regulatory motif pairs. (A) Enumerating neighbor motif pairs. In case 1, two binding sites of motif A are close to the binding site of B. Since only the closest motif pair is considered, 1 pair is counted. In case 2, the binding site of motif A is close to the binding site of motif B. However, another binding site of A overlaps with site B, 0 pairs are counted. (B) Motif identity shuffling. Each regulatory motif was considered separately as pivot. The identities of all other regulatory motifs were randomly permuted within each composition cluster (listed in Figure 2.9), and significant proximal motifs for the pivot motif were profiled. For the final result of combinatorial motif pairs, a reciprocal mutual hit was required as both sides should identify each other as a proximal motif.

A. *vfl* ↔ *bcd* between distance




B. Combinatorial regulatory motif pairs in stage S5

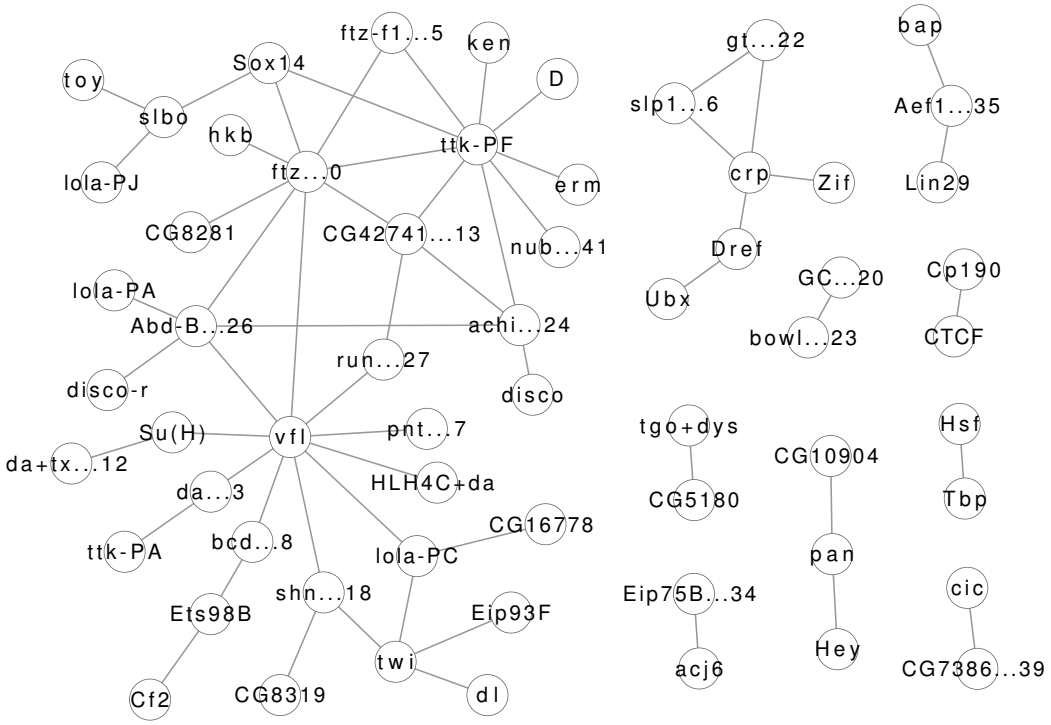



Figure 2.8: Combinatorial regulatory motif pairs with significantly co-localized sites. (caption continued on next page.)

Figure 2.8: (previous page) For each pair of TF regulatory motifs, all neighbor binding sites were enumerated within DNaseI accessible regions in each stage. The distances between all enumerated neighboring sites were profiled by a histogram with a step of 100nts, and the count in each bin was weighted by the conservation percentile scores of the enumerated binding site pairs. As a background, the identities of regulatory motifs with similar base pair composition were permuted across the same chromosome and the distance histograms were profiled again. For each window, an empirical $P$-value was calculated from 10000 randomizations, and the Benjamini-Hochberg procedure was used for multiple hypothesis correction. A FDR threshold of 0.05 was used. (A) For regulatory motifs represented by *vfl* and *bcd* in stage S5, the first window has FDR smaller than the threshold 0.05. (B) All predicted combinatorial TF regulatory motif pairs in stage S5 are shown.

The PWMs in our collection include binding specificities for *CTCF*, *Su(Hw)* and *Cp190*, which bind insulator elements. Our CCAT pipeline found that binding sites for *CTCF* and *Cp190* co-localize in all five stages (Figure 2.8B and Figure 2.10). Consistent with our findings, it was found that *CTCF* interacts with *Cp190*, and that its binding to targets requires *Cp190* in many cases [94]. Further, the ChIP binding profiles of *CTCF* and *Cp190* were previously observed to cluster together [85].

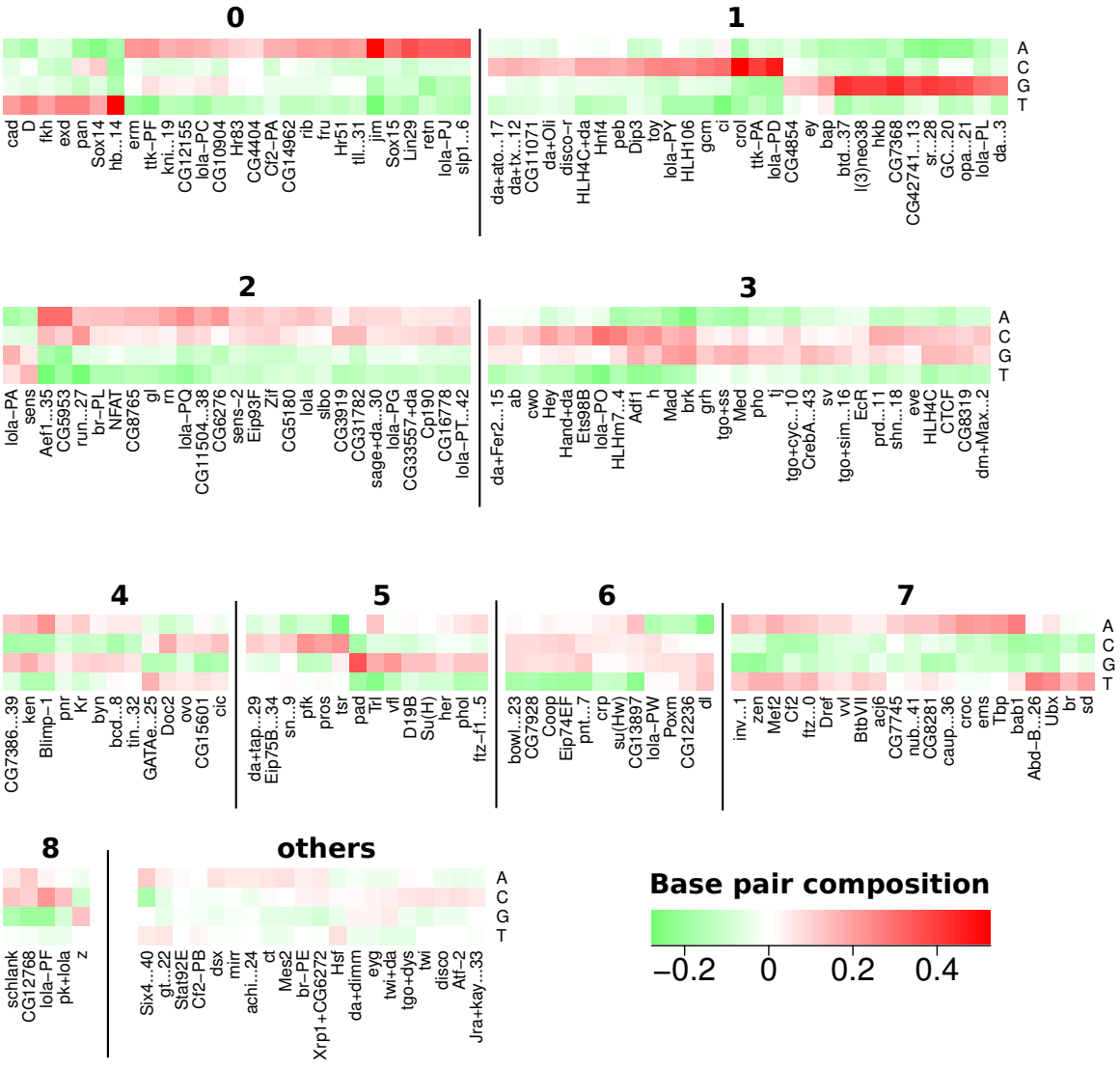



Figure 2.9: Clusters of regulatory motif base pair compositions. For all 198 TF regulatory motifs, we clustered them by the similarity of their base pair compositions. Each PWM was converted to a frequency vector of A,C,G,T content minus the background frequency of A,C,G,T over the whole fly genome. The standard deviation of each PWM composition was computed as a measure of base pair composition bias. The bottom 20% of them were excluded from further clustering as they don't have strong preference of compositions (shown as cluster "others"). Then all of the rest composition vectors were clustered by average link hierarchical clustering based upon the Pearson correlation coefficient. The hierarchical tree was cut at a Pearson correlation of 0.8.

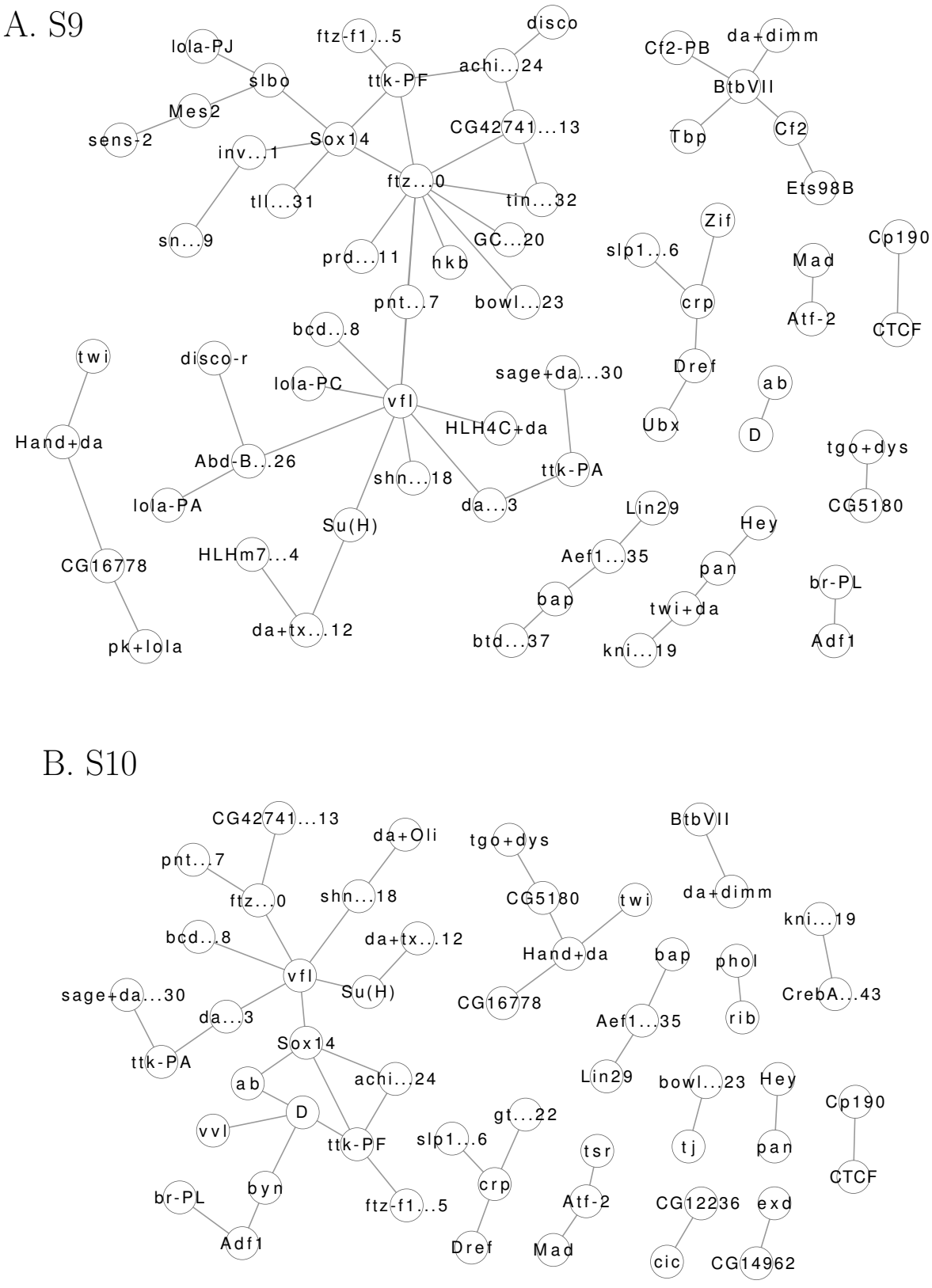
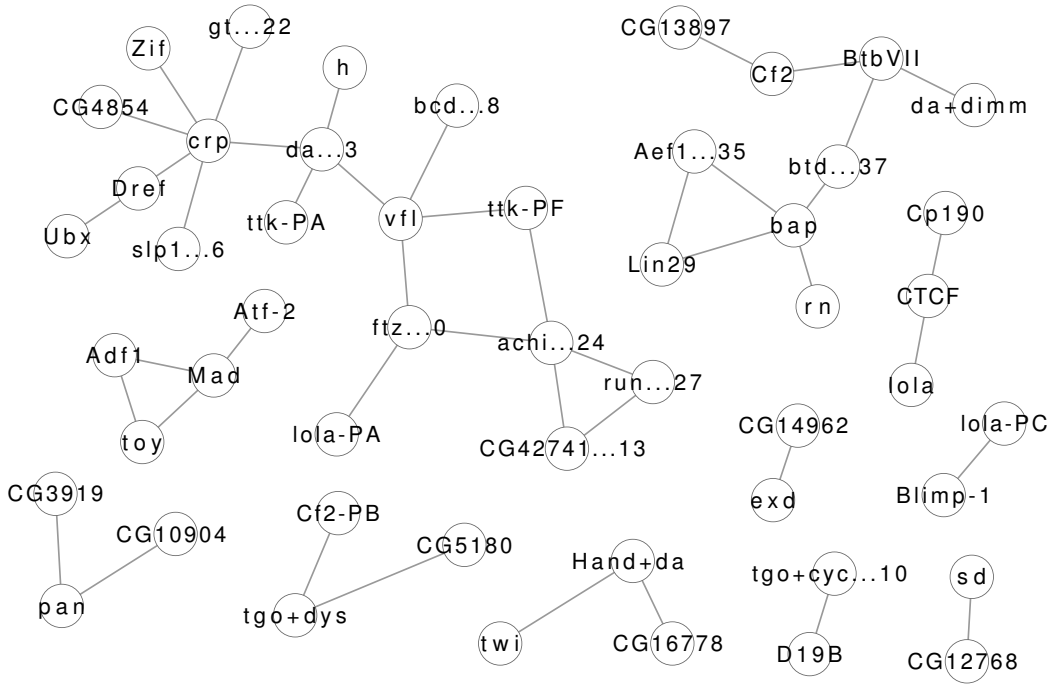



Figure 2.10: Combinatorial regulatory motif pairs. (caption continued on next page.)

C. S11




D. S14

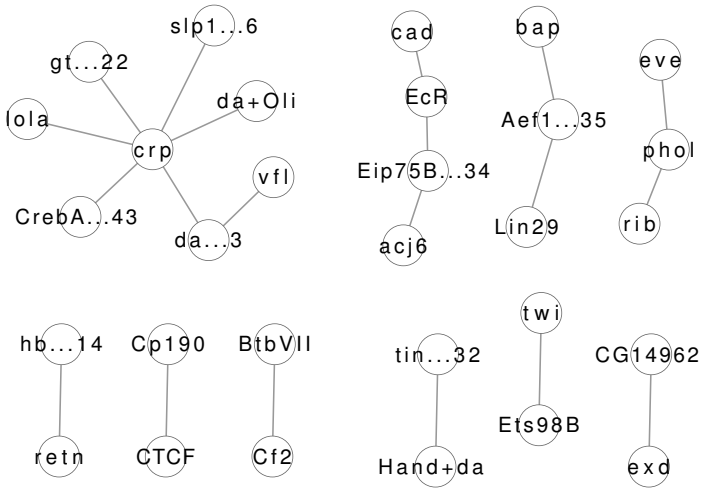



Figure 2.10: (previous page) For combinatorial regulatory motif pairs predicted, we plotted them by interaction graphs. (A) Stage S9. (B) Stage S10. (C) Stage S11. (D) Stage S14.

Encouraged by the coherence of our findings with previous studies, we set out to systematically assess the quality of our predicted combinatorial motif pairs. We reasoned that if a predicted binding site for a TF is close to a predicted binding site for one of the TFs with which it cooperates, then these predictions are more likely to be correct than other predicted sites for these TFs. To check this, we used our collected ChIP datasets. In order to use a TF in this assessment, for each TF that was profiled in at least one ChIP dataset, we considered its top 20% most conserved genome-wide binding sites, and required that at least 5% of them be located in a ChIP bound region. If several ChIP datasets existed for the same TF, we selected the ChIP dataset with the maximum percentage of predicted binding sites within ChIP bound regions.

For each TF considered, we classified its conserved binding sites into two categories: (1) those with a predicted conserved binding site within 100nts of it for a TF that was found to be co-localizing and (2) those with other predicted binding sites within 100nts, but none of the binding sites are for TFs that were found to be combinatorial pairs. For each category, the percent of binding sites within a ChIP bound region was computed (Figure 2.11A).

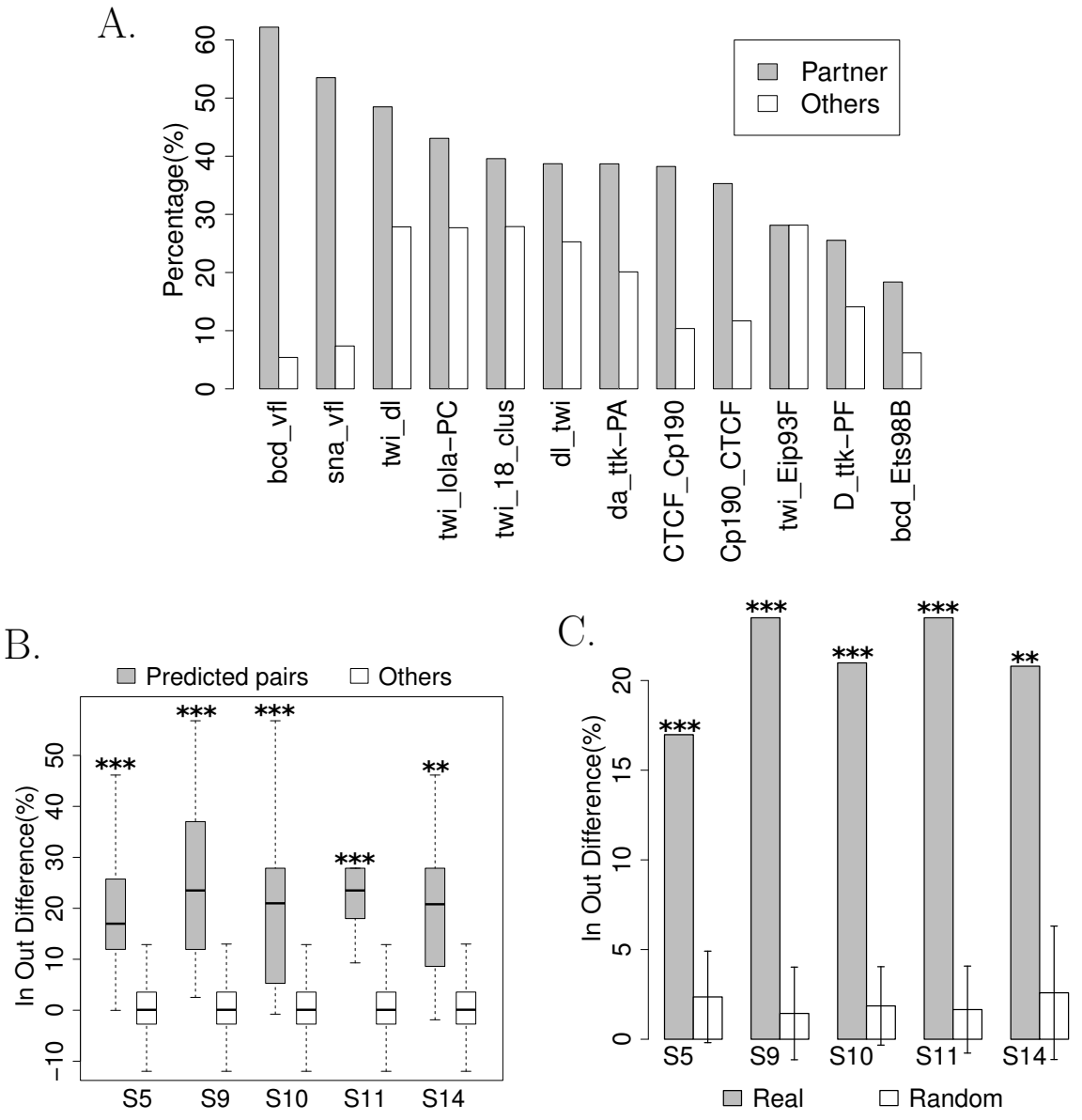



Figure 2.11: TF binding sites of combinatorial pairs are enriched in ChIP bound regions. All collected ChIP datasets are used to assess the quality of predicted motif pairs. The TF binding sites were classified into two categories: (1) those with another TF binding site within 100nts and the neighbor TF pair was predicted to be a combinatorial motif pair; (2) those with other TF binding sites within 100nts, but none of them comprise the predicted combinatorial pairs. For each category, the percent of TF binding sites within the ChIP bound regions was computed. (caption continued on next page.)

Figure 2.11: (previous page) (A) Combinatorial pairs profiled in stage S5 are considered. The ChIP percentages are plotted for the two categories. (B) The difference of ChIP percentages between the two categories is shown for all TF pairs profiled as combinatorial regulatory motif pairs and all other TF pairs that were not predicted. For each group of TF motif pairs, the measures are visualized by boxplots. The bottom and top of the box are the 25th and 75th percentiles (the inter-quartile range). Whiskers on the top and bottom represent the maximum and minimum data points within the range represented by 1.5 times the inter-quartile range. The Wilcoxon rank sum test was used to compare the two groups, and $P$-values were Bonferroni corrected for each of the five stages. One asterisk represents a $P$-value $\leq 0.05$, two asterisks represent a $P$-value $\leq 0.01$ and three asterisks represent a $P$-value $\leq 0.001$. (C) The combinatorial regulatory motif pairs in each stage were randomized by network edge swapping [89]. For each stage, the median difference is plotted for real pairs and randomized motif pairs. Average, standard deviation and empirical $P$-values were calculated from 10000 randomizations. $P$-values were Bonferroni corrected for the five stages and visualized by asterisks as in (B).

We assume that if a predicted combinatorial motif pair is used by TFs at a specific stage, the first category should have more binding sites located in ChIP bound region than the second category in that specific stage. We took the difference of percentages between two categories as a quality measure for the combinatorial motif pairs. For each stage, when comparing against other pairs of regulatory motifs that were not predicted, the difference in these two categories is significantly larger for predicted combinatorial pairs in all stages (Figure 2.11B). For each stage, we also built randomized motif pairs by treating the predicted combinatorial motif pairs as a network where edges correspond to uncovered combinatorial pairs between TFs and then randomizing the network via edge-swapping; note that this maintains the number of combinatorial pairs each motif is involved with [89]. The difference measures for real motif pairs are consistently better than randomized motif pairs in all stages (Figure 2.11C).

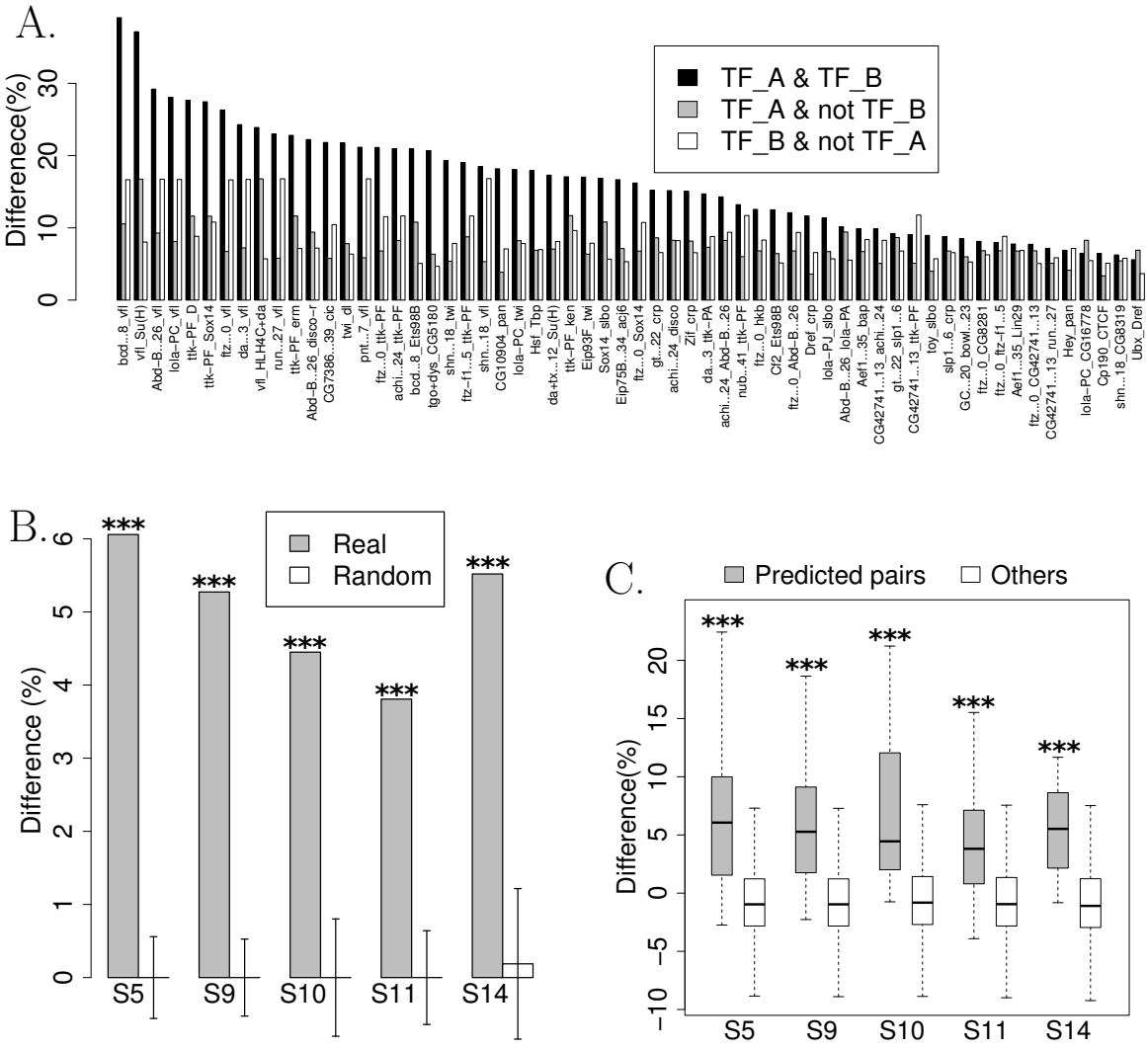



Figure 2.12: TF binding sites of combinatorial pairs are more conserved. (A) For each profiled combinatorial pair between TF A and B, the conservation percentile scores were compared among three categories: 1 (TF_A & TF_B) those with motif sites of TF_A and TF_B within 100nts; 2 (TF_A & not TF_B) those with motif site of TF_A and another motif site which is not TF_B within 100nts; and 3 (TF_B & not TF_A) those with motif site of TF_B and another motif site which is not TF_A within 100nts. The percent of site pairs with conservation percentile scores greater than or equal to 0.8 for both involved binding sites was calculated for all three categories. (caption continued on next page.)

Figure 2.12: (previous page) (B) For each motif pair, the measure "Difference" is defined as the percent of conserved site pairs (conservation percentile score≥0.8) of category "TF_A & TF_B" - maximum(the percentage of category "TF_A & not TF_B", the percentage of category "TF_B & not TF_A"). For each stage, boxplots were used to visualize difference measures for all predicted combinatorial pairs and other pairs not predicted to be preferentially co-localized. The bottom and top of the box are the 25th and 75th percentiles (the inter-quartile range). Whiskers on the top and bottom represent the maximum and minimum data points within the range represented by 1.5 times the inter-quartile range. The Wilcoxon rank sum test was used to compare between two groups. $P$-values were Bonferroni corrected for five stages. One asterisk represents a $P$-value $\leq 0.05$, two asterisks represent a $P$-value $\leq 0.01$ and three asterisks represent a $P$-value $\leq 0.001$. (C) The regulatory motif pairs predicted in each stage were randomized by network edge swapping [89]. The median difference was plotted for real pairs and randomized pairs. The average and standard deviation values and empirical $P$-values were calculated from 10000 randomizations. $P$-values were Bonferroni corrected for each of the five stages and visualized by asterisks as (B).

We also used evolutionary conservation to assess the quality of our uncovered combinatorial pairs of TFs. A previous study in mammalian embryonic stem cells revealed that a TF binding site would be more evolutionarily conserved if it was found near a binding site for a TF with which it cooperates [95]. We also searched for whether there was evolutionary constraint for our predicted combinatorial motif pairs. For each profiled combinatorial pair between TF A and B, the conservation percentile scores were compared among three categories: 1 (TF_A & TF_B) those with motif sites of TF_A and TF_B within 100nts; 2 (TF_A & not TF_B) those with motif site of TF_A and another motif site which is not TF_B within 100nts; and 3 (TF_B & not TF_A) those with motif site of TF_B and another motif site which is not TF_A within 100nts. Then for each of these three categories, we computed the percent of site pairs where both binding sites had conservation percentile scores greater than or equal to 0.8 (Figure 2.12A). Similar to the comparison based upon ChIP experiments (Figure 2.11), we computed the difference of the fraction of highly conserved pairs between the first category and the other two categories and found

that our predicted pairs consistently have more significant measures than pairs that are not predicted or than randomized pairs (Figure 2.12BC).

### 2.2.4 Dynamic usage of combinatorial pairs in embryo development

For all combinatorial motif pairs predicted in any of the five stages, we checked the extent to which they had stage-specific usage. We first computed the stage specific enrichment ratio for each stage by dividing the weighted counts of site pairs within 100nts between real and randomized data, as plotted in Figure 2.8A. We grouped all predicted motif pairs by their maximum stage specific enrichment ratios and visualized them in heatmap format (Figure 2.13A). For each motif pair, the variation ratio is defined as (maximum stage specific enrichment ratio)/(minimum stage specific enrichment ratio). The histogram of all predicted combinatorial pairs are compared with combinatorial pairs resulting from edge-swapping randomizations [89]. We found that for larger variation ratios, our predicted pairs are consistently more frequent than randomized pairs (Figure 2.13B).

We further concentrated on *twi* and *dl*, a combinatorial pair that we found to be strongly preferred in the stages S5 and S9 (Figure 2.13A), and for which we have ChIP experiments in several stages of development. In particular, *twi* has been profiled via ChIP in three consecutive embryo developmental stages (S5-7, S8-9 and S10-11) [77]. We found that *twi* binding sites with 100nts of a *dl* site had higher fraction in ChIP bound region in the first two stages (S5-7 and S8-9) than the third stage (S10-11) whereas when *twi* binding sites had other TF sites nearby, the fraction within ChIP bound regions was similar across the three stages (Figure 2.13C). Thus, this case study is coherent with our stage specific usage profiling (Figure 2.13A).

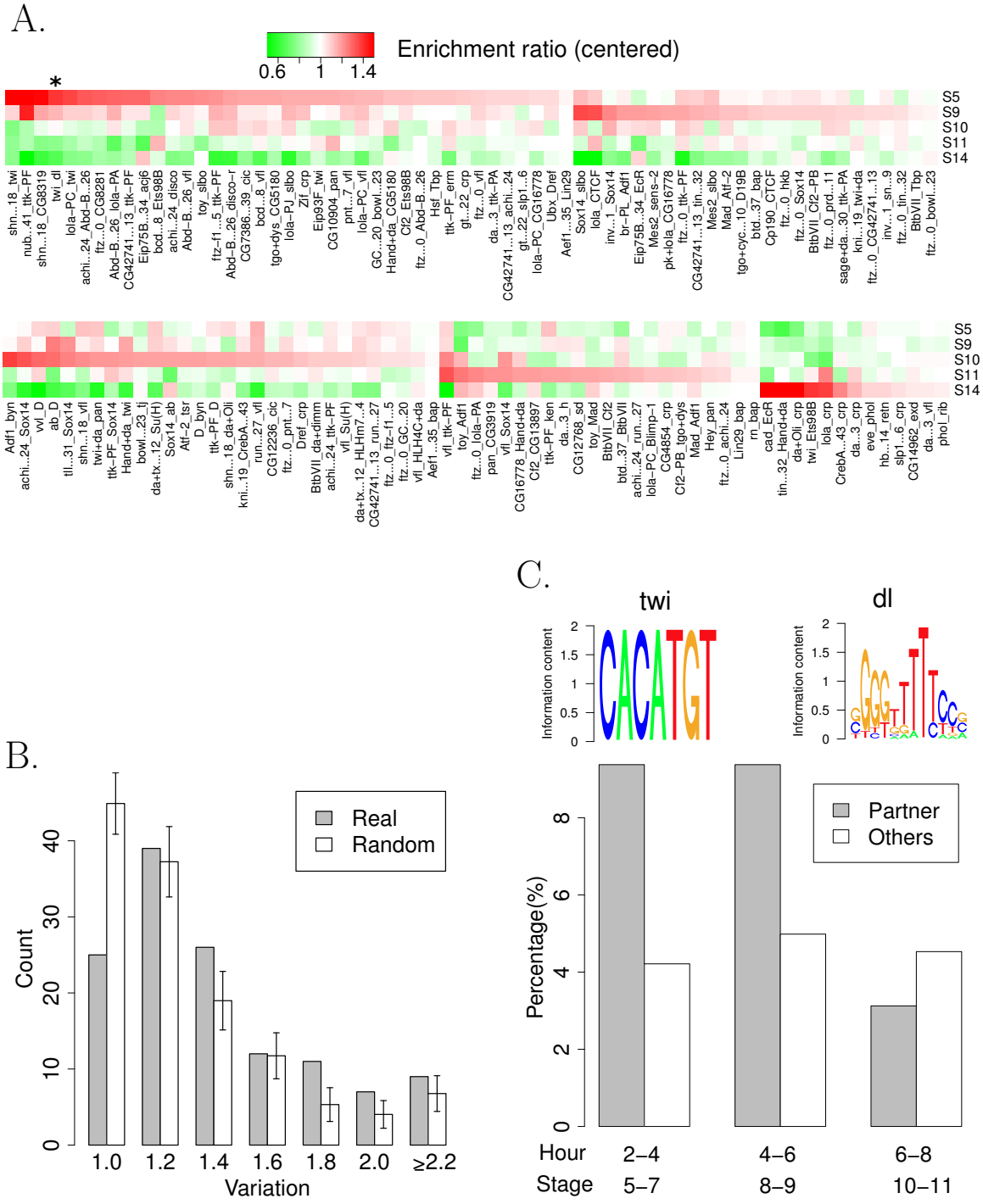



Figure 2.13: (A) For each pair of TF regulatory motifs, five stage enrichment ratios were determined as (#site pairs within 100nts in a stage specific DNaseI accessible region)/(#average site pairs within 100nts from random shuffles). For all predicted motif pairs across five stages, the enrichment ratio in each stage was centered by dividing the average ratio across five stages and visualized by heatmap. (caption continued on next page.)

Figure 2.13: (previous page) (B) A variation ratio is calculated as the maximum enrichment ratio across five stages divided by the minimum stage enrichment ratio. The histograms of variation ratios were plotted for all predicted TF regulatory motif pairs in any stages, and randomized TF regulatory motif pairs by edge swapping. The average and standard deviation values of randomized histogram were calculated from 10000 shuffles. (C) For the combinatorial motif pair *twi* and *dl*, the percent of predicted *twi* binding sites in ChIP bound region was plotted over three consecutive stages of embryo development as Figure 2.11A [77].

## 2.3 Discussion

We developed a pipeline for predicting combinatorial TF interactions based upon known TF binding motifs and DNaseI data. In addition to capturing some known cases of TF cooperativity, our systematic quality assessments revealed that our predicted TF pairs are coherent with experimental ChIP data and supported by evolutionary analysis. Thus for specific biological processes of focus, without requiring hundreds of ChIP experiments, our pipeline enabled the genome scale profiling of the landscape of transcriptional cooperativity from a single DNaseI-seq experiment. In addition, we also developed accompanying tools to cluster similar PWMs of different TFs into clusters which removed redundancy in our predicted TF pairs.

### 2.3.1 Positional constraints between regulatory motifs

Currently, we only considered proximity in binding sites when predicting whether two TFs combinatorially cooperate. However, several studies have shown that within enhancer regions, TF binding sites may follow specific positioning and orientation rules [42, 56, 96]. For example, in the human Interferon-$\beta$ enhancer, eight TFs bind together with a very specific motif order within 55bps of DNA [56]. Further, in several well-characterized developmental enhancers in Drosophila, binding sites show a periodic distribution that reflects the geometry of helical turns of DNA [96]. A

computational study in yeast also showed that interacting TF binding sites follow very strict spacing and orientation preferences [42].

To date, we have not found evidence that our predicted combinatorial motif pairs exhibit any spacing or orientation preferences between binding sites. Instead, we observe a relatively flexible spacing in our data, consistent with the billboard model of enhancers [4]. For example, one study of Drosophila cardiac development revealed that five TFs cooperatively bind a large set of enhancers that have diverse motif compositions along with flexible positioning between binding sites [97].

One possibility for the difference of positioning constraint might come from the biological processes studied. Our study is focused on embryo development and it is possible that developmental combinatorial binding allows flexible spacing. On the contrary, the Interferon-$\beta$ enhancer needs to rapidly response to viral infection [56] and may prefer a highly ordered structure among TF binding sites. Another possibility for differences in positional constraints might come from the evolutionary differences between organisms. In a single cell organism such as yeast, protein physical interaction between TFs might facilitate the strength of cooperativity and the spacing and orientation constraints reflect the constraints of physical interactions [42]. In higher eukaryotes such as Drosophila, spacing flexibility might allow TF cooperativity in more enhancers and allow more cooperativity with different TFs. However, without a systematic study of TF motif positioning under many different biological contexts, it is not possible to conclude whether flexible or strict positioning is more common in combinatorial regulatory motif pairs.

### 2.3.2 Systematic profiling of combinatorial regulatory code among diverse biological processes

We have shown that combining binding sites matches to TF PWMs with DNaseI accessibility experiments can result in high-quality genome-wide TF binding site pre-

dictions that are comparable in quality to those obtained by ChIP experiments (Figure 2.2BC). Several previous studies also reached a similar conclusion [70, 98, 50]. Our high-quality binding site predictions allowed us to find combinatorial interactions between regulatory motifs at the genome scale. The different conditions of DNaseI experiments also enabled us to uncover the dynamic usage of combinatorial motif pairs in different stages. Compared to ChIP experiments, which require one experiment for each single TF under each condition, DNaseI experiments enable genome-scale profiling of combinatorial code in only one single experiment. The recent release of the ENCODE project includes genome-wide DNaseI experiments for 125 different human cell lines and tissue types [68]. Thus, the growing availability of DNaseI experiments would enable reliable combinatorial regulatory motif pair profiling in many biological conditions.

## 2.4 Methods

### 2.4.1 Binding site search, conservation and accessibility scoring for position weight matrix

Multiple genome alignments of *D.melanogaster* and 11 other sequenced Drosophila species were downloaded from the UCSC genome browser [99]. Each position weight matrix (PWM) was searched on both strands of the genome sequences via the algorithm fimo from the MEME package [80, 81], using the default $P$-value threshold 1E-4. We excluded all binding sites in protein coding exons, as annotated by Flybase [100].

For each match to a PWM on the *D.melanogaster* genome, we looked for matches in the other 11 genomes on either strand within an offset of 10 nts. These additional matches were considered conserved instances, and were used to calculate a branch length score (BLS) as follows [61]. We obtained the minimum phylogenetic subtree that included all conserved instances. The BLS was computed as the total branch

length of this subtree as a fraction of the entire tree [61]. We observed that it was possible to get a high BLS score if there was an isolated match in a species distant to *D. melanogaster*. Since such a match may be spurious, we ignored the match in the genome most distant from *D. melanogaster* if there was a gap of more than four species from the second most distant match and the evolutionary distance from *D. melanogaster* was two times bigger than the second most distant match. Then for each TF PWM, all of its binding sites were ranked by BLS scores from largest to the smallest. These BLS scores were then converted to conservation percentile scores, which represent the relative ranks among all predicted binding sites. For example, a conservation percentile score of "0.6" means the current binding site has BLS score greater than 60% of all predicted binding sites for that PWM.

For each predicted PWM binding site in *D.melanogaster*, we derived accessibility scores based upon DNaseI experiments over five embryo development stages (S5, S9, S10, S11 and S14 corresponding to 3, 4, 5, 6, and 11 hours after fertilization) [67, 66]. For each predicted PWM binding site, its DNaseI accessibility score was estimated by averaging all DNaseI experimental scores ±50nts around it. For each stage, the top 5% of DNaseI scores across the whole genome was set as a threshold, and PWM sites with an average DNaseI score larger than this threshold were defined as accessible binding sites.

### 2.4.2 Collection and selection of TF regulatory motifs

We collected 712 PWMs representing the binding specificities of 364 different DNA binding proteins from FlyFactorSurvey [71, 72, 73], BDTNP [23, 22], Flyreg [74], JASPAR [75], Transfac 6.0 [76], a collection of Kellis and colleagues [61] and several ChIP experiment papers [77, 78, 79]. Of the 364 collected DNA binding proteins, 170 have two or more PWMs associated with them. For example, the well-studied *bcd* has 12 different PWMs in our data set. It has been shown previously in *S.cerevisiae*

that different PWMs for the same TF may differ in quality even if they share motif similarity [101]. Thus, in order to control for PWM quality and correct for study bias, only one PWM was selected for each TF as follows.

We collected ChIP datasets for 53 TFs from BDTNP [23, 22], modEncode [84] and several publications [77, 78, 79, 85, 86]. For any TF, if there is ChIP experimental data for it, the percent of PWM binding sites within the ChIP experiment regions was calculated. If several different ChIP datasets existed for the same TF, the median value was taken for comparison. For different PWMs of the same TF, the PWM with the highest ChIP percentage was selected. For TFs without available ChIP experiment, the five embryo stage DNaseI dataset was utilized [67, 66]. The regions with DNaseI accessibility score in the top 5% were treated as the ChIP region for PWM selection.

Finally 364 PWMs were selected for 364 DNA binding proteins. After searching with the fimo algorithm with the default $P$-value threshold 1E-4 [80, 81], 324 of them have matched binding sites on the *D. melanogaster* genome.

### 2.4.3 Clustering of highly similar TF regulatory motifs

Many TFs, especially from certain structural families, exhibit similar binding specificities [72]. Thus predicted binding sites for different TFs may overlap extensively with each other. We clustered our selected PWMs by hierarchical clustering and then merged their overlapping binding sites.

For each pair of PWMs A and B, the Pearson correlation coefficient (PCC) was calculated as follows. The PCC between two PWM columns X and Y was computed as $\sum(X-\overline{X})(Y-\overline{Y})/\sqrt{\sum(X-\overline{X})^2\sum(Y-\overline{Y})^2}$ with $\overline{X}$ (or $\overline{Y}$) representing the average of X (or Y). The information content (IC) was computed for each PWM column as $2+\sum_{i\in(A,C,G,T)} P_i * log(P_i)$. For an aligned column between A and B, the information content was computed as $\sqrt{IC_{column\ A} * IC_{column\ B}}$.

For each possible ungapped alignment between PWM A and B, a weighted PCC was computed as $\sum_{i \in Aligned\ columns}(IC_i * PCC(A_i, B_i))/\sum_{c \in All\ columns} IC_c$ . Since we search PWMs on both the positive and negative strands of genome, the PCC is also calculated between PWM A and the reverse complement motif of PWM B. A final PCC for the similarity between A and B was calculated as the maximum possible PCC with the optimal column offset and orientation.

Average linkage clustering was applied to group all PWMs into a hierarchical tree. To uncover clusters of PWMs, the tree was cut at PCC $\geq 0.8$. 198 PWM clusters were acquired by this threshold and 44 of them contained two or more PWMs. For each cluster of TFs, a binding site is predicted for that cluster if it is predicted for at least one of its members.

### 2.4.4 GO enrichment analysis for the predicted TF target genes

In order to measure the biological function similarity between a TF and its predicted target genes, we utilized Gene Ontology (GO) annotations [87]. We only considered GO biological process terms with more than 5 but less than 1000 genes annotated in *D. melanogaster*. For each TF and a GO term annotated with it, we computed the fraction of its target genes that are also annotated with that term. For all genes that are not a target of that TF but are annotated with at least one GO term, we also calculated the fraction annotated. The enrichment ratio for that specific GO term and its annotated TF is computed as (fraction of target genes with that annotation)/(fraction of non-target genes with that annotation). For each TF, we considered all annotated biological process terms and use the median among all enrichment ratios over all TFs as an overall measure for each dataset (Figure 2.2BC).

### 2.4.5 Finding combinatorial regulatory motif pairs

Within the top 5% of accessible DNaseI regions of each stage, the predicted binding sites with conservation percentile score greater than or equal to 0.6 were first selected. Then, for each regulatory motif in turn as "pivot", we set out to find other regulatory motifs that significantly co-localize with it.

First, we enumerated all pairs of predicted binding sites between the pivot regulatory motif and other regulatory motifs, weighted by the minimum conservation percentile score between the two involved binding sites. For each genomic region where several binding sites for the same TF are clustered, only the neighboring sites closest to the pivot motif were considered (Figure 2.7A). Then for each pivot TF motif and other motifs, we classified all enumerated binding site pairs by the distance between the two involved sites and derived a histogram in steps of 100nts (Figure 2.8A). In each histogram bin, the weights of all enumerated binding site pairs were added up as a weighted count (Figure 2.8A).

To estimate how often TF binding sites would co-localize by chance, we randomized the identities of the other regulatory motifs. The binding site identity of the pivot regulatory motif is not changed; but the identities of all other motif binding sites are shuffled across the same chromosome (Figure 2.7B). As an extra constraint in the shuffling process, we classified the 198 TF PWM clusters by the similarity in their nucleotide compositions and created 10 composition clusters (method described in the next section). Only regulatory motifs in the same composition cluster can exchange their identities in the shuffling process (Figure 2.9). In this way, the local base pair composition will be similar to the initial data after randomization. Then, the motif pair sites neighboring the pivot regulatory motif were enumerated again. Histograms were plotted according to the distances of all motif pairs for real data and randomized data (Figure 2.8A).

Empirical $P$-values were computed in each histogram bin as the fraction of randomized weighted counts that were greater than or equal to the real weighted count among 10000 randomizations. For each histogram bin, we only considered motif pairs with weighted count greater than or equal to 1% of the total sum of all conservation percentile scores of each involved regulatory motif (pivot and other TF). The Benjamini-Hochberg procedure was applied on the empirical $P$-values, and a FDR threshold of 0.05 was used to select significant combinatorial motif pairs [102]. The final set of combinatorial motif pairs were selected by reciprocal hit in the first histogram bin (distance within 100nts), where FDR$\leq$0.05 was required when each motif was used as the pivot (Figure 2.7B). We also required the final set of predicted combinatorial motif pairs to have at least 10 pairs with conservation percentile score greater than or equal to 0.6 within 100nts of each other.

### 2.4.6 Categorizing regulatory motifs by base pair composition

For each of the 44 PWM clusters that contained more than one PWM, one centroid PWM was generated by averaging over all included PWMs over their shared columns. Then for each of the 198 PWM clusters, we computed the A,C,G,T content of its centroid PWM by averaging over columns. The background frequency of the whole fly genome was subtracted from these compositions (A:28.87%, C:21.15%, G:21.11%, T:28.86%).

The standard deviation of each PWM composition was calculated as a measure of base pair preference. The bottom 20% of all compositions ranked by their standard deviations were excluded for further clustering, since they show weak preference of base pair compositions (the cluster "others" in Figure 2.9). All of the rest frequency compositions were then clustered by average link hierarchical clustering using the Pearson correlation coefficient (PCC). Since we search the regulatory motifs on both

strands on the fly genome, the PCC between base pair composition vectors were calculated in both the same direction and the reverse complement direction, and the maximum of the two was used.

The hierarchical tree was cut at PCC 0.8. If there were small clusters with less than 5 members, this would lead to a very restricted space in our motif identity shuffling process. We merged them to the last cluster (labeled with "others"). Finally, 10 composition clusters were generated (Figure 2.9).

# Chapter 3

# Cooperativity between RNA binding proteins and microRNAs in transcript decay

## 3.1 Introduction

Transcript degradation is an important mechanism for regulating the levels of proteins in a time or space-dependent manner [3]. One mechanism through which transcript degradation can be controlled is via miRNAs, short RNAs approximately 21-23 nucleotides in length that regulate diverse biological processes [10, 12]. miRNAs are initially transcribed as pri-miRNAs, processed to form pre-miRNAs, which are hairpins of approximately 70-80 nucleotides, exported from the nucleus, and further processed to generate the final mature dsRNA [10]. Mature miRNAs are then loaded into the RISC complex, where they associate with target transcripts, resulting in transcript degradation and translation inhibition [103].

miRNAs generally bind their targets through complementary pairing in a short 7 bp seed sequence [9, 11]. There are likely other factors that also determine whether a

miRNA will effectively target a particular recognition site, and some of these factors may be 3′UTR sequences that reside outside of the complementary sequence that the miRNAs bind. As an example, AU-rich sequences surrounding the miRNA binding sites have been reported to enhance the efficacy of miRNA-mediated mRNA decay [11, 104, 34]. The location of the recognition site at the 5' or 3' end of the 3′UTR, and especially far away from the center of long 3′UTRs, has also been associated with improved miRNA efficiency [11]. Thus, given a target transcript with a specific miRNA recognition site sequence, its decay efficiency is likely to be determined by a number of variables not all of which are currently well-understood.

Transcript degradation can also be regulated by RNA binding proteins (RBPs). These proteins can affect transcript stability by binding to recognition sequences within 3′UTRs. Some RBPs, for instance, AU rich element (ARE) binding proteins or Pumilio (PUM), increase the degradation of target transcripts [13, 14, 13, 15, 16, 17, 18, 19, 20]. Others, like the HuR family of ARE-binding proteins [21], cause stabilization of the targeted message. Several genomewide studies have suggested that RBPs and miRNAs may functionally interact [105]. Mukherjee and colleagues found that microRNA depletion had a less dramatic effect on sites at which the HuR binding protein could also bind, indicating that HuR was likely competing with microRNAs for binding sites and stabilizing the targeted transcript [106]. In another study, an analysis of gene expression changes after miRNA transfection revealed that U-rich motifs similar to HuD binding sequences were associated with transcript down-regulation [107]. Finally, immunoprecipitation with antibodies to the PUM protein followed by microarray analysis of surrounding RNA sequences revealed that miRNA binding sites are overrepresented in 3′UTR sequences within close proximity to PUM binding sites [51].

Specific instances in which RBPs enhance or inhibit the effectiveness of miRNAs have been experimentally verified. Competition between miRNAs and RBPs for the

same sequence has been reported [108, 109, 110]. For example, down-regulation of the cationic amino acid transporter 1 (CAT-1) mRNA by miR-122 is inhibited by stress, and the de-repression requires binding of HuR to the 3′UTR [108]. As another example, the RBP CRD-BP binds to the coding region of TrCP1 mRNA and stabilizes it by competing with miR-183 and thus preventing miRNA-dependent processing [110]. miRNAs and RBPs have also been reported to cooperate. HuR and the miRNA let-7 repress c-MYC expression though a mechanism that requires both HuR and let-7 [35]. The C. elegans PUM homolog puf-9 is required for 3′UTR-mediated repression of the let-7 target hbl-1 [36]. In Drosophila, an association between the RBP dFXR and RISC is required for efficient RNA interference [111]. As a final example, an AU-rich motif located upstream of the miR-223 binding site in the 3′UTR of RhoB has been reported to enhance miRNA function [34].

One specific mechanism through which the PUM RNA binding protein has been proposed to modulate miRNA function is by binding to sequences that can hybridize with miRNA recognition sites and thereby make them more accessible for the RISC complex. Binding of PUM to the 3′UTR of the cyclin-dependent kinase inhibitor p27Kip1 has been reported to cause a local change in structure that promotes p27Kip1 repression by miR-221/miR-222 [28]. Another study demonstrated that binding of PUM facilitated miR-503 regulation of the E2F3 3′UTR [112]. The authors hypothesized PUM binding was able to relax the 3′UTR secondary structure elements that would otherwise block miR-503 binding sites. A final study on the pyrimidine-tract-binding (PTB) protein proposed that PTB binding can modulate the secondary structure of the GNPDA1 3′UTR to facilitate let-7b binding [26].

We hypothesized that miRNAs and RBPs might cooperate to facilitate transcript decay more extensively than had been realized. Using computational models, we systematically explored RBP-miRNA interactions within human and mouse 3′UTRs and discovered that RBP recognition sites co-occur with subsets of miRNA recognition

sites. Our analyses revealed that PUM is likely to cooperate with specific miRNAs to promote decay. Moreover, we found that a subset of miRNAs that co-occur with PUM recognition sites have recognition seed sequences that are the reverse complements of the PUM recognition motif, and thus, may form hairpin secondary structures that would be disrupted by PUM binding. Based on our computational analysis, we discovered seven miRNAs in human and five in mouse that followed this pattern. Approximately 4% of the target sites for these miRNAs colocalize with PUM sites in a pattern that would have the potential for miRNA binding site rescue.

## 3.2 Results

### 3.2.1 RBP and miRNA recognition motif selection

We performed a literature search and identified 15 instances in which an RBP and its putative recognition motif were reported [16, 113, 114, 115, 116, 117, 118] (Figure 3.1). We reasoned that true RBP recognition motifs that are functional in 3′UTRs would be present more frequently than expected by chance, especially at high levels of evolutionary conservation. Using a method adapted from Kellis and colleagues [119], we found 5 out of the 15 RBPs had significantly increased conservation frequencies compared to their shuffled control motifs (Figure 3.2AB, 3.3). All five of these motifs have been demonstrated to be present in 3′UTRs by previous studies. The motifs are recognition motifs for the transcript decay factors PUM (UGUANAUA) [116], the Fox-1 family of proteins associated with splicing (UGCAUGU) [120, 121, 122], U1A (AUUGCAC)(a component of the snRNP complex) [123, 124], Nova (YCAUUUCAY) [117], and the AU-rich element (ARE) UAUUUAU, which is bound by many different ARE binding proteins [13, 125].

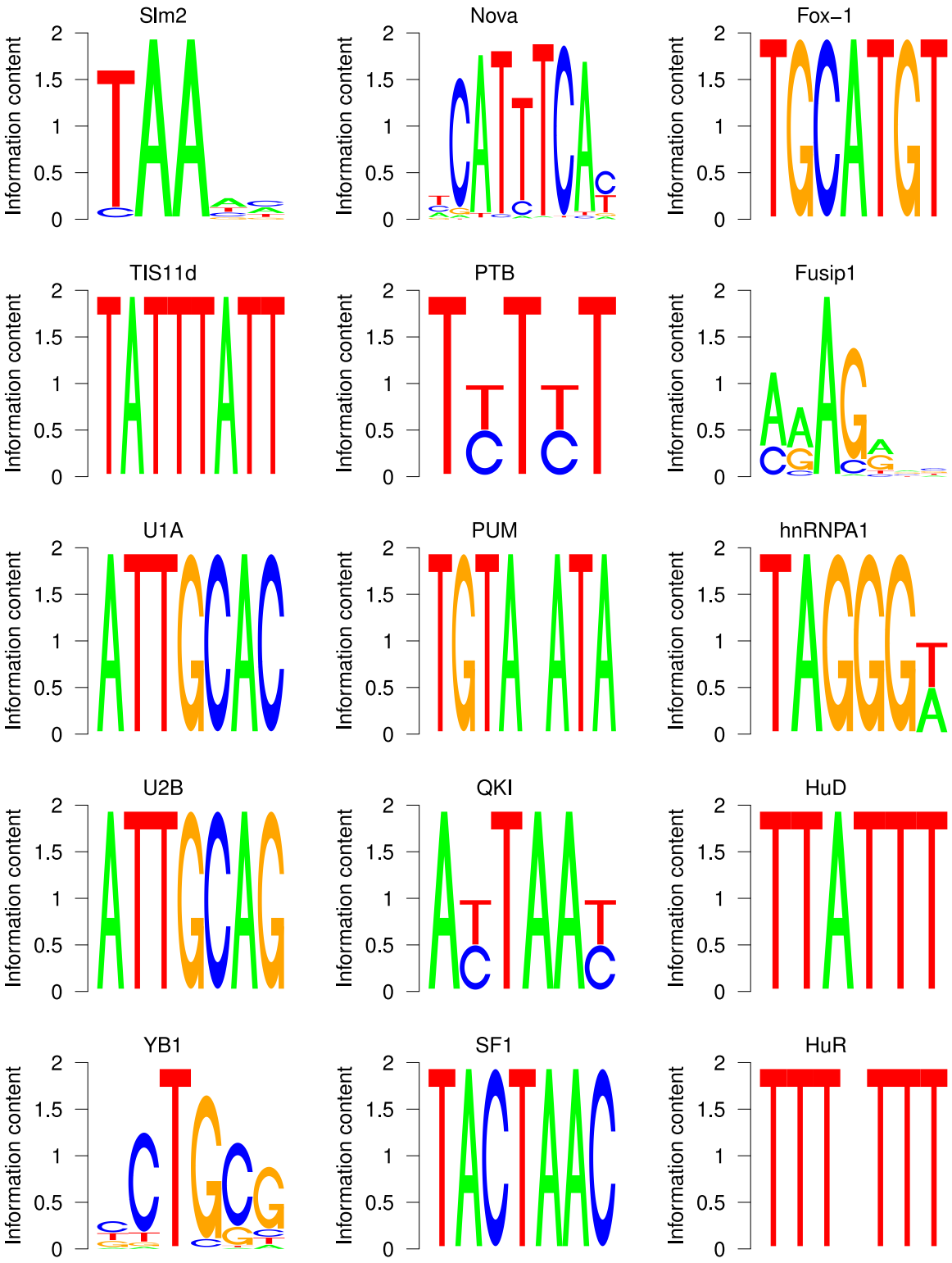



Figure 3.1: RBP recognition motifs. The position weight matrix logos are shown for the 15 RBP motifs that were evaluated.

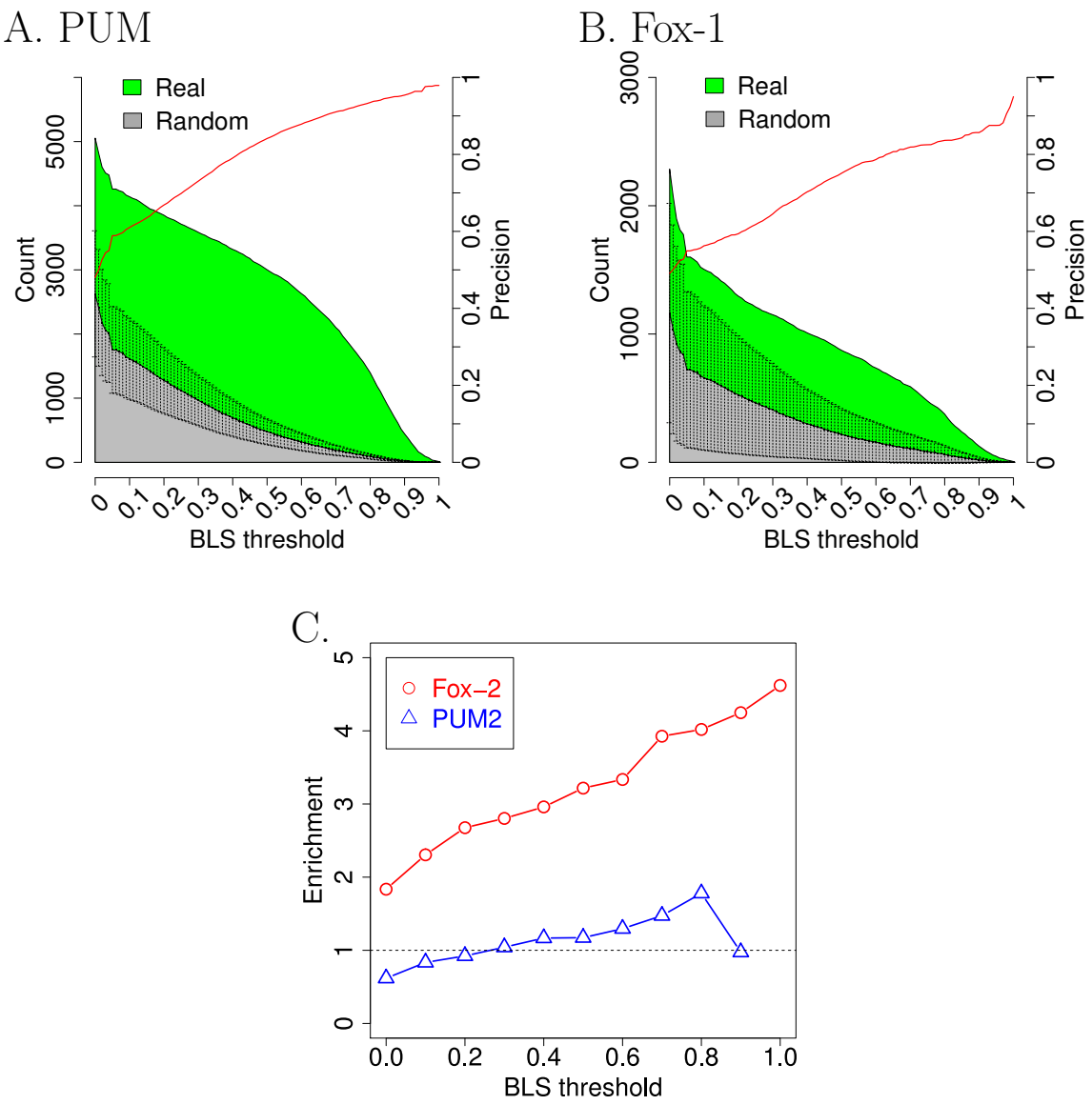



Figure 3.2: Identifying RBPs with binding sites evolutionarily conserved in 3′UTRs. (A, B) For each instance of a putative RBP recognition motif site in 3′UTRs, the Branch Length Score (BLS) was determined based on multiple genome alignments. The number of motifs at different levels of conservation (BLS) is plotted. The area below the curve for the true RBP is shaded green. The frequency with which randomly shuffled motifs were present in the genome is indicated in gray according to the y-axis on the left. Error bars indicate the standard deviation for the different shuffled versions of the motif. The precision ratio (1-[The average number of matches of shuffled motifs]/[The number of matches of the canonical motif]) is indicated by the red line according to the y-axis on the right. (caption continued on next page.)

Figure 3.2: (previous page) (C) CLIP-Seq binding regions for Fox-2 were mapped on human 3′UTRs [126]. At each conservation BLS threshold, an enrichment ratio was determined by comparing the density of binding sites within the CLIP region versus outside the CLIP region for the Fox-1 and Fox-2 binding motif UGCAUGU [127]. The BLS threshold is shown on the X-axis and enrichment is shown on Y axis. As a control, enrichment of the Fox-1 and Fox-2 motif in PUM2 Par-CLIP sequences was also plotted [118].

For the PUM recognition motif, for instance, there is a large increase in the number of observed recognition sites compared with the number expected based on shuffled controls (Figure 3.2A). In contrast, U2B is reported to bind the sequence AUUGCAG [128], however, its putative binding sites were present a comparable number of times in 3′UTRs compared with shuffled versions of the motif at all levels of evolutionary conservation (Figure 3.3D). U2B and the nine other such RBPs were therefore not included in our subsequent analyses.

One example of a RBP motif that passed our threshold was the Fox-1 family binding site (UGCAUGU), which represents a family of RBPs that are well-conserved in metazoans. In mammals, there are three members of the Fox-1 family, Fox-1, Fox-2 and Fox-3 [127]. The Fox-1 RBP family recognizes sites with a consensus sequence of UGCAUGU [120] and matches to this sequence were consistently present in 3′UTRs at a higher frequency than shuffled controls (Figure 3.2B). This was somewhat unexpected because Fox-1 family RBPs are generally considered to be splicing factors [127]. To further confirm that the recognition sites on 3′UTRs that we designated as a Fox-1 family binding sites are bound by Fox-1 family RBPs, we analyzed 34,111 non-overlapping regions on human 3′UTRs identified in a previous study of Fox-2-associated sequences using next generation sequencing [126]. As a member of the Fox-1 RBP family, Fox-2 also binds UGCAUGU, so we compared the density of Fox-1 family motifs within the immunoprecipitated 3′UTR sequences with the density in 3′UTRs outside the immunoprecipitated sequences. The enrichment for the Fox-1 family motif increased monotonically with an increasing conservation threshold,

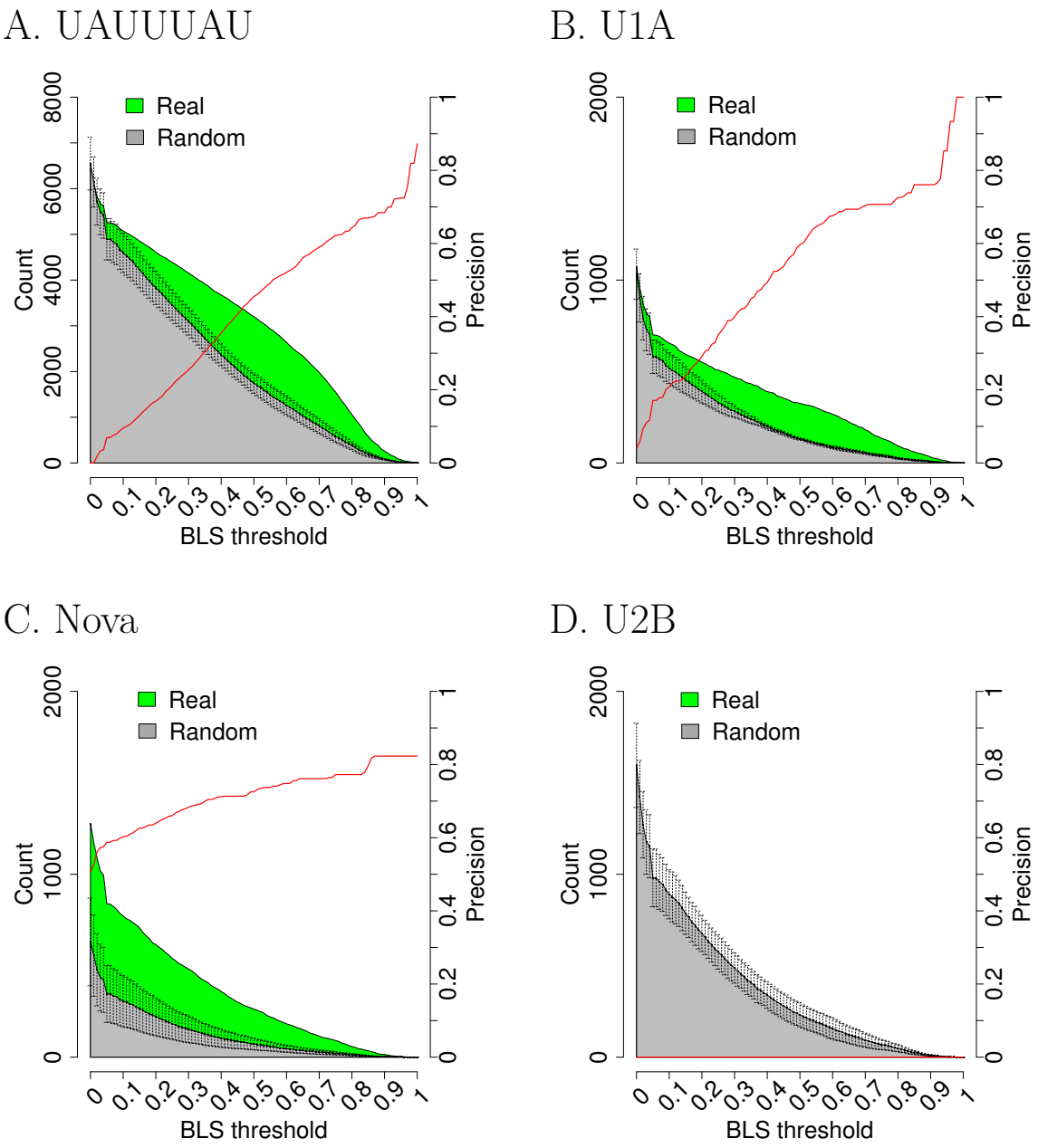



Figure 3.3: RBP recognition motif selection. The number of instances of each RBP was plotted for different Branch Length Scores (BLS) as in Figure 1. These values are plotted using the y-axis on the left. True motifs are indicated in green and shuffled motifs are indicated in gray. Precision is shown in red and plotted according to the y-axis on the right. (A) UAUUUAU. (B) U1A. (C) Nova. (D) U2B.

from twice as frequent for all binding sites to 4 times more frequent when requiring perfect conservation through all placental mammals (Figure 3.2C). As a control, we didn′t observe a significant enrichment for the Fox-1 motif within sequences immunoprecipitated with antibodies to PUM2 [118] (Figure 3.2C). We conclude that the computational approach that we are using to define RBPs that bind 3′UTRs is consistent with experimental data, and that members of the Fox-1 family likely do bind 3′UTRs.

### 3.2.2 RBP motifs tend to localize to the end of long 3′UTRs

Previous analyses showed human miRNA recognition motifs tend to localize at the 5′ beginning or 3′ end of long 3′UTRs [129, 130]. For the five RBP recognition motifs included in this study, we investigated the localization of the associated RBP binding sites along 3′UTRs. We first classified the human 3′UTRs into three length categories: 3′UTRs with length $< 500$ nts (6622 transcripts), 3′UTRs with length $\geq 500$ nts and $< 2000$ nts (7385 transcripts) and 3′UTRs with length $\geq 2000$ nts (3759 transcripts). Within each length category, we divided 3′UTRs into 10 equal parts and counted the percentage of motif occurrences in each of the 10 bins. We observed that for RBP motifs PUM and UAUUUAU, the number of recognition sites is highest at the very end of the 3′UTRs longer than 500 nts (Figure 3.4A). For 3′UTRs longer than 2000 nts, we created ten 100-nt-windows from the 5′ beginning and 3′ end of the full UTR and counted the percentage of RBP motifs found in each window. The number of RBP motifs PUM and UAUUUAU was highest in windows located 100 nts and 200 nts from the end of the 3′UTRs (Figure 3.4B).

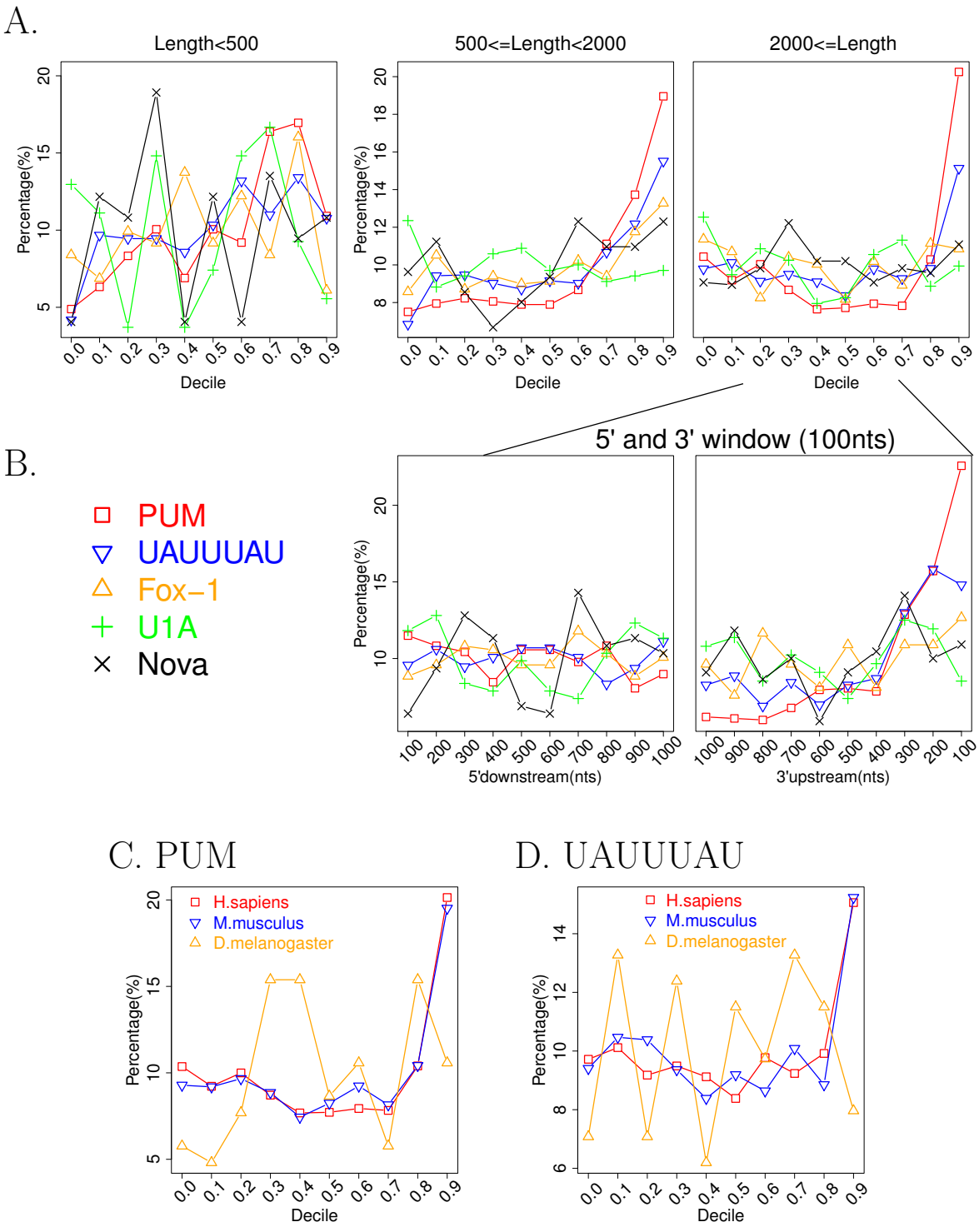



Figure 3.4: RBP motifs tend to localize to the end of long 3′UTRs. (A) All human 3′UTRs were classified into three length categories: smaller than 500 nts, longer than 2000 nts, or between 500 and 2000 nts. Each 3′UTR was equally divided into ten bins, numbered from 0.0 to 0.9. Within each length category, the percentage of RBP recognition sites in each bin was plotted. (caption continued on next page.)

Figure 3.4: (previous page) (B) For 3′UTRs longer than 2000 nts, ten 100-nt-windows from the 5′ start towards downstream and the 3′ end towards upstream were analyzed. The percentage of RBP recognition sites in each window was plotted. (C, D) Human (H.sapiens), mouse (M.musculus) and fly (D.melanogaster) (but not the worm C.elegans) contain 3′UTRs longer than 2000 nts. The localization patterns of PUM and UAUUUAU recognitions sites were plotted in ten bins for 3′UTRs longer than 2000 nts.

For RBP motifs such as PUM and UAUUUAU with high AU content, their preferential distribution at the very end of 3′UTRs could, in principle, reflect the higher AU-content at the end of long 3′UTRs. We analyzed the fraction of AU base pairs in different deciles of 3′UTRs and found that 3′UTRs tend to have high AU-content at the 3′ end region in human, mouse, fly and worm (Figure 3.5). In order to control for AU-content, we generated shuffled control motifs that have the same base pair composition as the initial motif for each RBP. We compared the percentage of RBP recognition sites in each 3′UTR bin with the average from all shuffled RBP motifs in the same bin (Figure 3.6). In the human genome, PUM recognition sites (binomial test $P$-value =2.24E-23) and UAUUUAU (binomial test $P$-value =6.57E-3) were significantly more frequent at the very 3′ end of 3′UTRs, even after correcting for the high AU content in this region of 3′UTRs (Figure 3.6A, B).

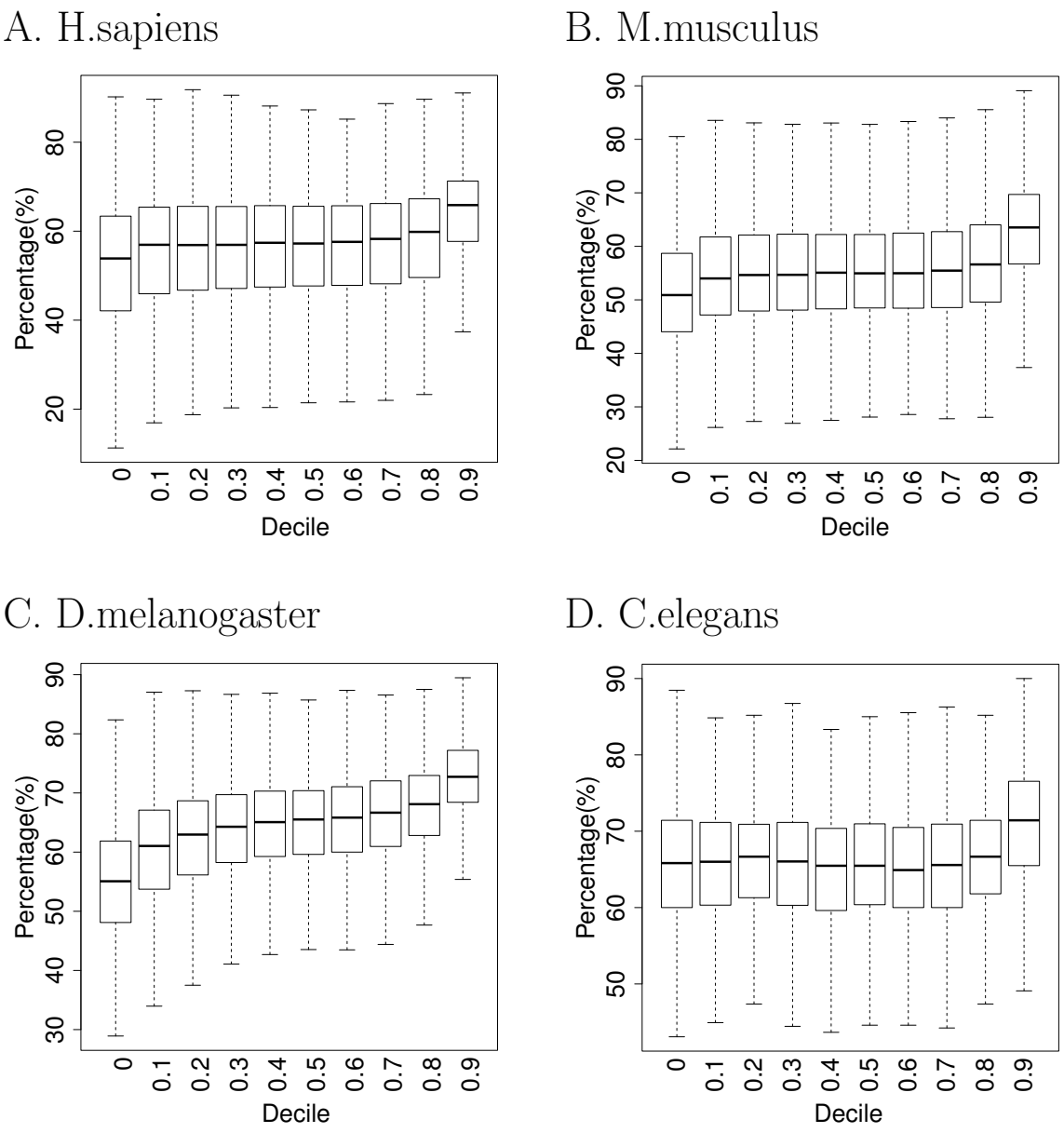



Figure 3.5: AU-content is high at the end of 3′UTRs. Each 3′UTR longer than 500 nts was equally divided into ten deciles. For each decile, AU-content was calculated and box-plots across all genes are shown. (A) Human. (B) Mouse. (C) Fruit fly. (D) Worm.

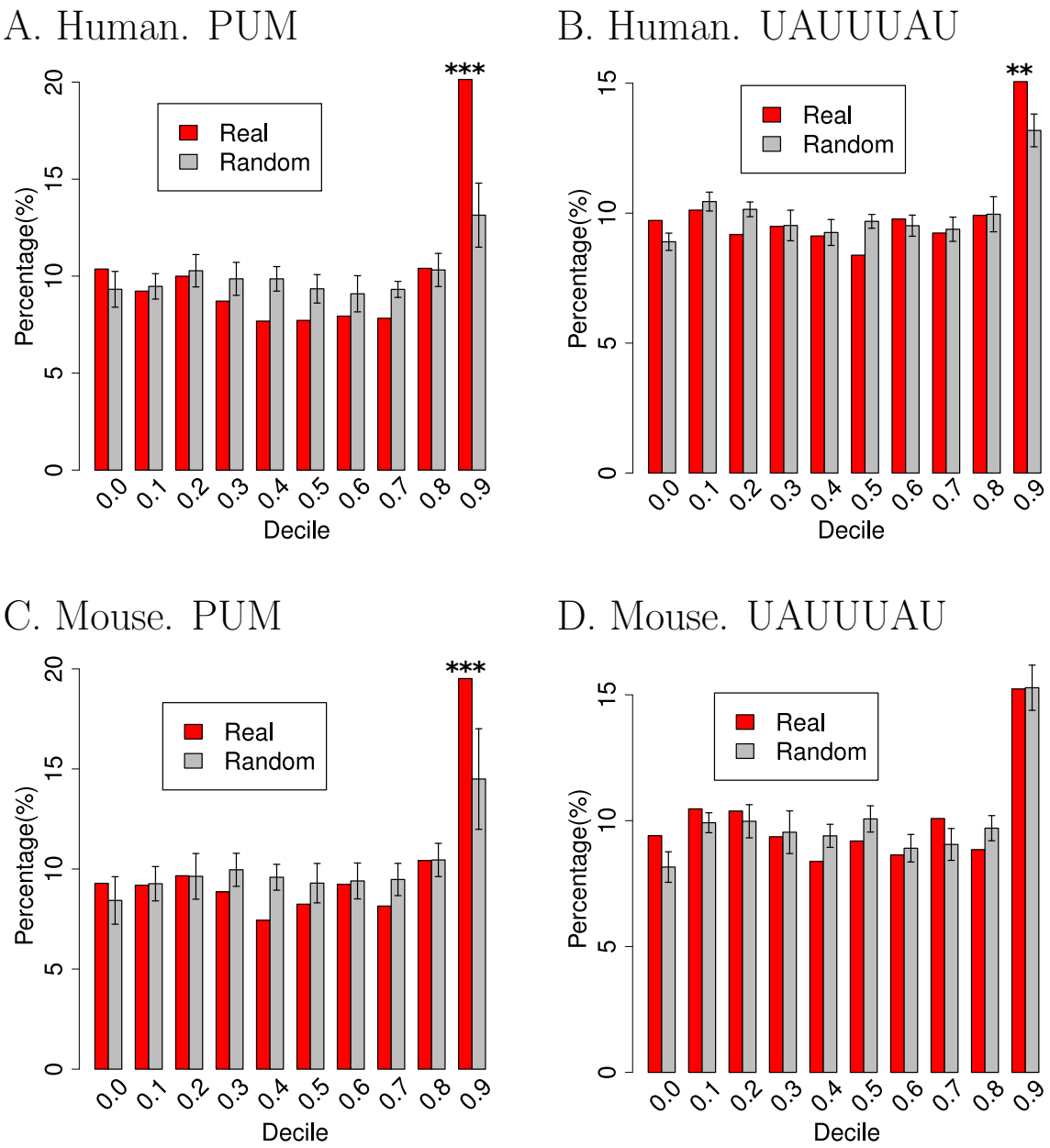



Figure 3.6: RBP localization compared with shuffled control motifs. Each 3′UTR longer than 2000 nts was equally divided into ten bins, numbered from 0.0 to 0.9. The percentage of RBP recognition sites in each bin was compared with its shuffled control motifs, which have the same AU-content. For each 3 ′UTR bin, asterisks represent comparisons of percentages between real and random motifs using the binomial test with a Bonferroni correction for 10 tests. One asterisk indicates a $P$-value $\leq 0.05$, two asterisks indicate a $P$-value $\leq 0.01$, and three asterisks indicate a $P$-value $\leq 0.001$. (A) PUM localization in human. (B) UAUUUAU localization in human. (C) PUM localization in mouse. (D) UAUUUAU localization in mouse.

Having discovered that certain RBP recognition motifs are enriched at the 3' ends of long 3′UTRs in human, we then asked whether this localization pattern is present in other species as well. PUM is part of a well-conserved family of PUF proteins [18, 131]. There are PUM proteins that bind the consensus sequence UGUANAUA in human [116], mouse [132], fly [133] and worm [134]. UAUUUAU is also a binding motif for RBPs in human, mouse [135], fly [136] and worm [137]. We analyzed the localization of the PUM and UAUUUAU consensus sequence within 3′UTRs in these four species and discovered that the preference for the 3' most region of long 3′UTRs exists in human and mouse, but not fly and worm (Figure 3.4C, D for 3′UTRs longer than 2000 nts and Figure 3.7 for 3′UTRs shorter than 2000 nts but longer than 500nts). For mouse, we also determined the extent to which AU content can explain the enrichment for PUM and UAUUUAU at the 3′ end of longer 3′UTRs.

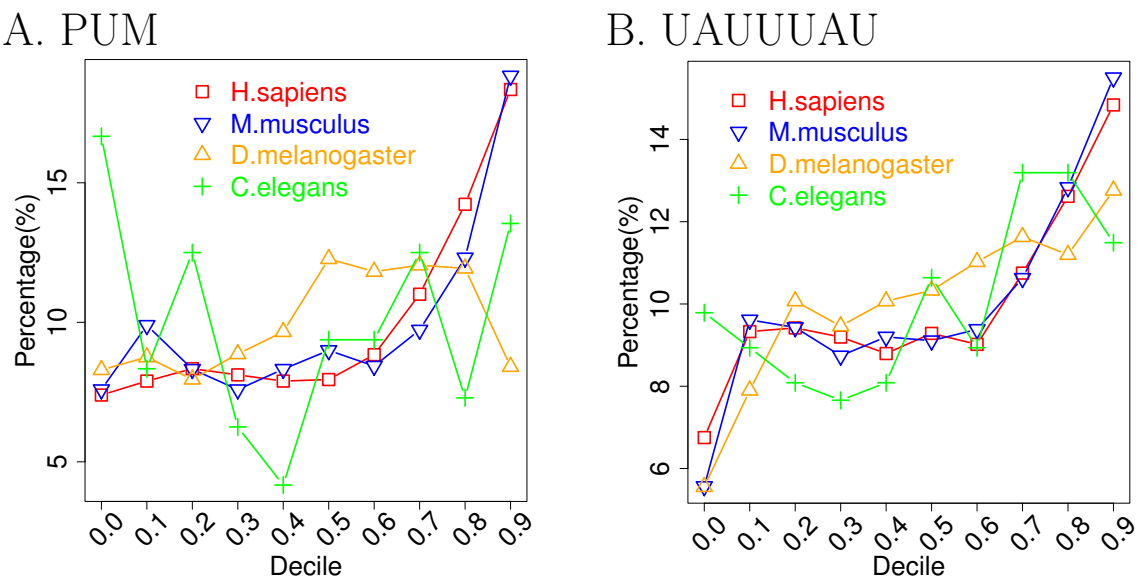



Figure 3.7: Localization of RBP motifs in 3′UTRs across four organisms. For each organism, 3′UTRs with length longer than 500 nts, but shorter than 2000 nts, were considered. Each 3′UTR was equally divided into 10 bins, numbered from 0.0 to 0.9. The percentage of RBP recognition sites in each bin was plotted. (A) PUM localization pattern. (B) UAUUUAU localization pattern.

In a pattern similar to that observed in human, PUM strongly localized to the most 3′ decile of mouse 3′UTRs compared to shuffled control motifs (Figure 3.6C, binomial test $P$-value = 1.95E-9). However, UAUUUAU was present in a similar percentage of 3′UTRs compared to shuffled control motifs with the same AU content (Figure 3.6D), even though it is highest in the most 3′ decile. Thus, for mouse 3′UTRs, both PUM and UAUUUAU are enriched at the very end of 3′UTRs, but the UAUUUAU enrichment is likely explained by the high AU-content at the end of mouse 3′UTRs.

### 3.2.3 Recognition sites for RBPs and specific miRNAs colocalize

We then asked whether the recognition sites of RBPs and miRNAs tend to be present close to each other on the same transcripts, as previous studies have reported that RBPs and miRNAs that functionally interact are often located close to each other [34, 35, 36]. For each pair of RBP and miRNA, we counted the number of neighboring RBP and miRNA recognition sites within 50 nts. As a control, we shuffled the identities of predicted miRNA recognition sites, while keeping their positions intact. An empirical $P$-value was calculated by comparing the observed number of neighboring RBP and miRNA recognition sites within 50 nts with the number of neighboring sites when the miRNA identities were randomized. For each RBP, miRNAs were classified as interacting miRNAs if they had a false discovery rate (FDR) less than 0.05 as determined by the Benjamini Hochberg procedure [138].

Among the five RBPs investigated, only PUM and UAUUUAU have miRNAs that are more abundant than expected within 50 nts of the RBP recognition site using this procedure (Table 3.1). For ten 50-nt windows upstream and downstream from RBP recognition sites, we plotted the ratio of the observed number of miRNA recognition sites to the expected number of sites, as estimated by randomly shuffling

miRNA site identities (Figure 3.8A). As expected, for interacting miRNAs, the ratio of observed to expected events is high around the RBP sites, and is lower in more distant windows.

We performed a similar analysis to determine interacting miRNAs for the same RBPs in mouse and discovered that there are miRNAs that colocalize with the PUM recognition site or UAUUUAU in mouse 3′ UTRs (Figure 3.8B). Some miRNAs have recognition sites that co-localize with PUM and UAUUUAU in both species (Figure 3.8C). For example, five of the seven miRNAs identified as PUM-interacting miRNAs in the human genome are also PUM-interacting miRNAs in mouse.

| | **50** | **100** | **150** | **200** | **250** | **300** | **350** | **400** | **450** | **500** |
|---|---|---|---|---|---|---|---|---|---|---|
| **PUM** | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **UAUUUAU** | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Fox-1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **U1A** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Nova** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 3.1: RBP and miRNA recognition motifs colocalize. For each RBP, the number of predicted interacting miRNAs is shown for each of ten 50-nt-windows from the RBP motif.

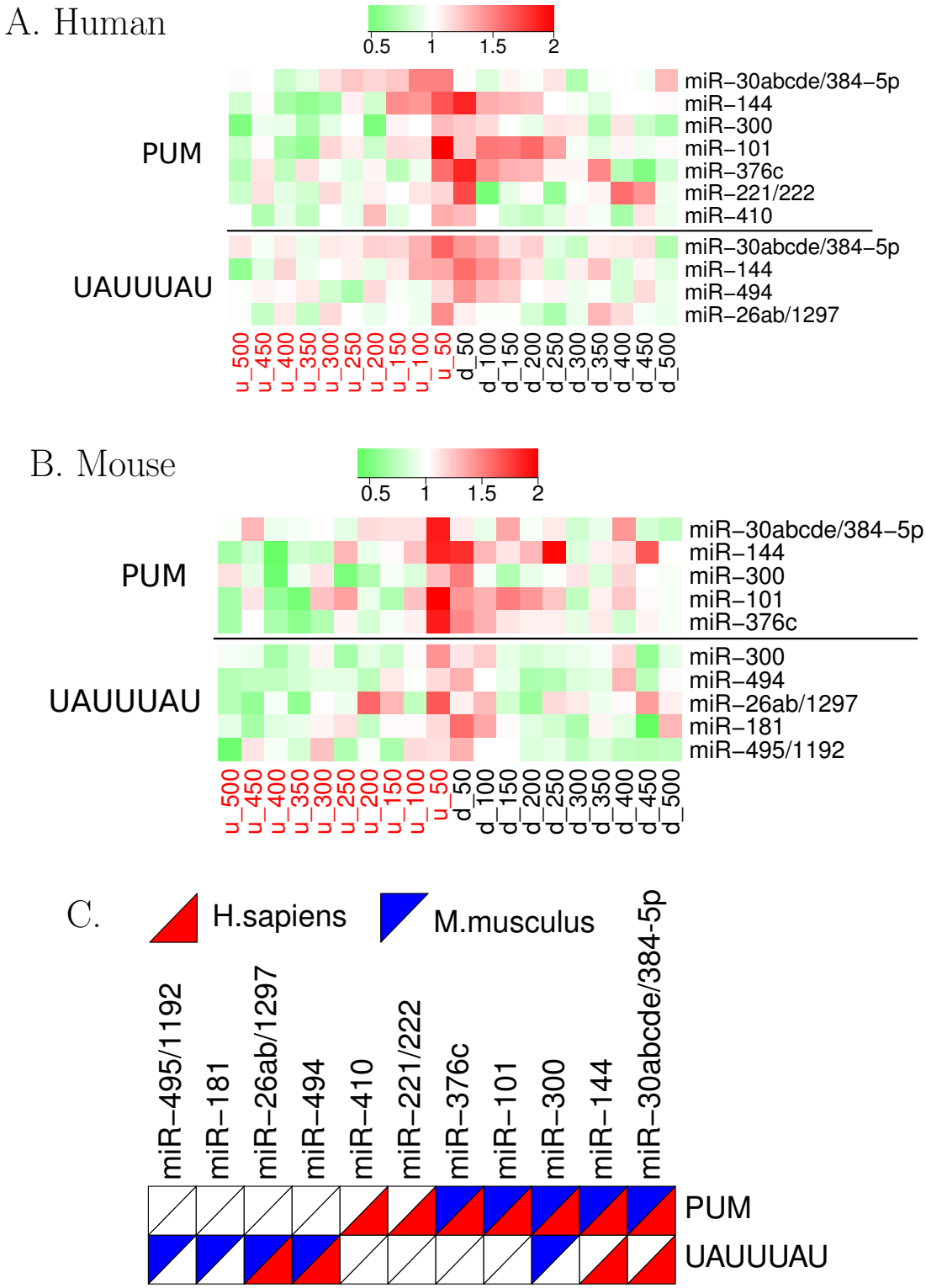



Figure 3.8: The recognition sites of RBP and specific miRNAs colocalize. (caption continued on next page.)

Figure 3.8: (previous page) For each RBP, the set of miRNAs with recognition sites that co-localize with the analyzed RBP with an FDR $\leq$0.05 are shown. The number of neighboring miRNA sites in ten 50 nt windows up- or down-stream of the RBP recognition site were compared to the number when the miRNA identities were shuffled. For each window, the ratio was determined as (number of miRNA sites)/(expected number based on shuffling). The miRNA site ratios were visualized in heatmap format with red indicating a high ratio and green indicating a low ratio. (A) Human miRNAs. (B) Mouse miRNAs. (C) A pairwise matrix between RBPs and interacting miRNAs is shown (FDR$\leq$0.05). In each cell, a red upward-sloping triangle is used to indicate colocalization in human and a blue downward-sloping triangle is used for mouse.

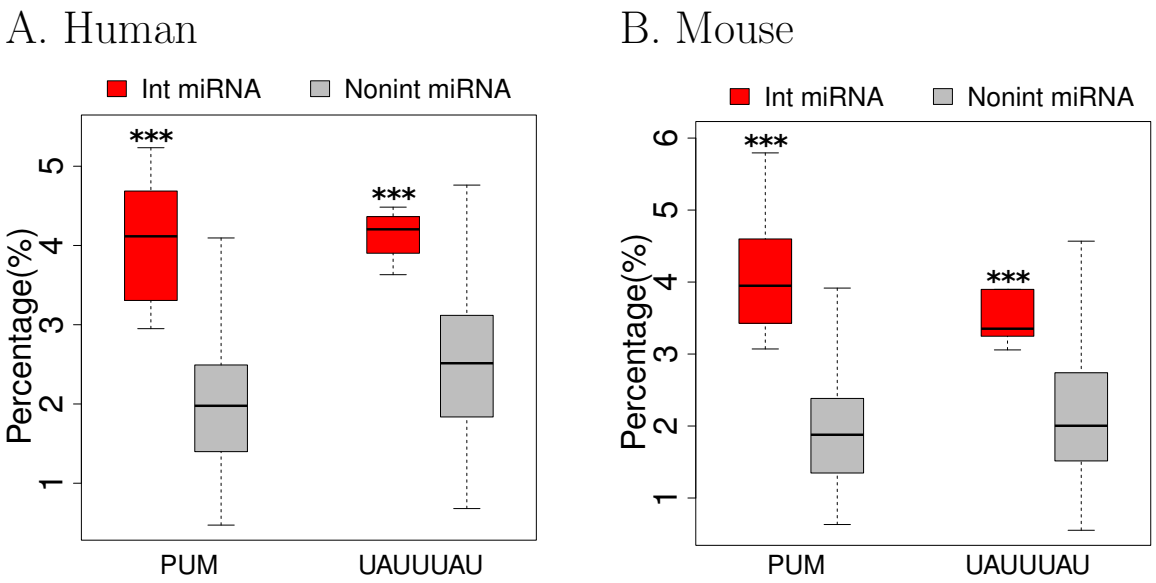



Figure 3.9: Percentage of miRNA recognition sites within 50 nts from RBP sites. For each RBP, the percentage of miRNA recognition sites within 50 nts of an RBP recognition site was determined for the set of all interacting miRNAs (Figure 3.8) and for the set of all non-interacting miRNAs. All values are shown with box-plots. For each RBP, asterisks represent comparisons of percentages between Int miRNA and Nonint miRNA determined by Wilcoxon rank sum tests. One asterisk indicates a $P$-value $\leq 0.05$, two asterisks indicate a $P$-value $\leq 0.01$, and three asterisks indicate a $P$-value $\leq 0.001$. Two separate data plots are shown for (A) Human and (B) Mouse.

For the interacting miRNAs, we calculated the percentage of all miRNA recognition sites that are located within 50 nts from the sites of their preferentially co-localized RBPs (Figure 3.9). For both PUM and UAUUUAU, the fraction of their interacting miRNA binding sites that are found proximal to RBP sites is around 4%. As expected, a smaller fraction of binding sites are proximal to the RBP recognition sites for non-interacting miRNAs in both human and mouse (Figure 3.9).

We further tested whether PUM and its predicted interacting miRNAs are enriched in experimental data in which the binding sites for both PUM2 and AGO were experimentally profiled in HEK293T cells by Par-CLIP (Photoactivatable Ribonucleoside Enhanced Crosslinking and Immunoprecipitation) [118]. We restricted the predicted PUM and miRNA recognition sites to only those identified by Par-CLIP analysis of PUM2 and AGO and counted the number of miRNA sites within 50 nts of PUM recognition sites. To define the background expectation, we permuted the identities of miRNA sites across chromosomes and counted the number of neighboring sites after restricting our analysis to sites within Par-CLIP regions. For each miRNA, an enrichment ratio was calculated as (the true number of neighboring sites)/(the average number of neighboring sites from 10000 shuffles). The interacting miRNAs showed significantly higher enrichment ratios than non-interacting miRNAs (Figure 3.10A).

We also recognized that not all miRNAs are expressed in all cells. To address this issue, the interacting miRNAs were further classified based on the sequence read abundance in HEK293T cells [118]. Expressed miRNAs, which were defined as the miRNAs with the top 25% read frequency, showed significantly higher enrichment ratios than non-expressed miRNAs (Figure 3.10B). Thus, the set of interacting miRNAs we predicted based on computational analysis of the genome sequence are also enriched based on Par-CLIP experimental data, and this enrichment shows the expected dependency on cell type-specific miRNA expression.

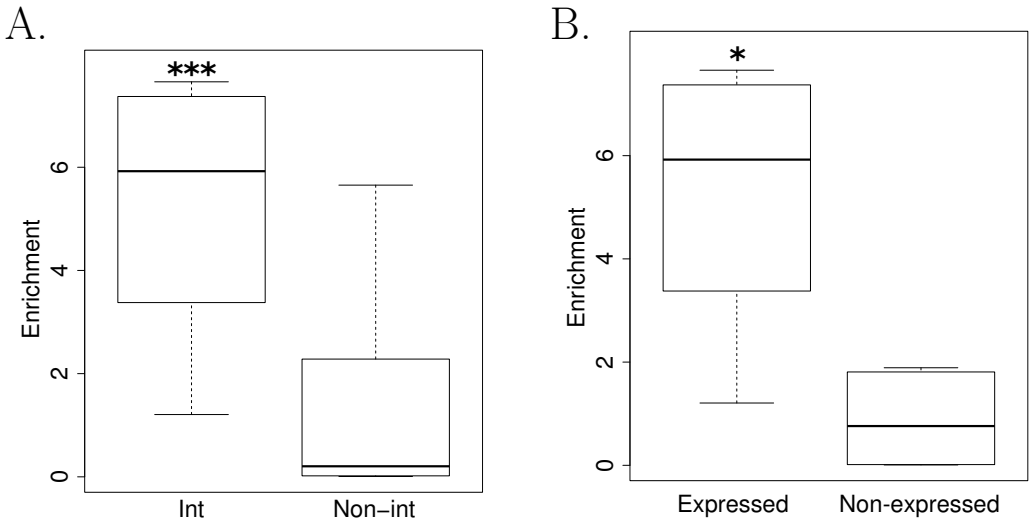



Figure 3.10: PUM co-localizes with its interacting miRNAs in the Par-CLIP region. The recognition sites of PUM and miRNAs were restricted to those experimentally identified by Par-CLIP analysis of PUM2 and AGO binding in the HEK293T cell line [118]. The number of neighboring PUM and miRNA recognition sites within 50 nts was counted. To determine the number of neighboring recognition sites expected by chance, the labels of all miRNA recognition sites were shuffled across chromosomes and the number of neighboring PUM and miRNA sites within the Par-CLIP region was counted again. For each miRNA, the enrichment ratio was calculated as (#neighboring sites)/(#expected sites). (A) Enrichment ratios for PUM-interacting miRNAs (Figure 3.8A) and non-interacting miRNAs are shown with box-plots. Asterisks represent comparisons of enrichment ratios between the two groups determined by Wilcoxon rank sum tests. One asterisk indicates a $P$-value $\leq 0.05$, two asterisks indicate a $P$-value $\leq 0.01$, and three asterisks indicate a $P$-value $\leq 0.001$. (B) The PUM-interacting miRNAs were classified as expressed if they were among the 25% of the most frequently sequenced small RNAs in HEK293T cells [118] (miR-30abcde/384-5p, miR-101 and miR-221/222). The rest of the PUM-interacting miRNAs were classified as non-expressed (miR-144, miR-300, miR-376c, miR-410). Enrichment ratios between the two groups were visualized and compared in the same way as described in (A).

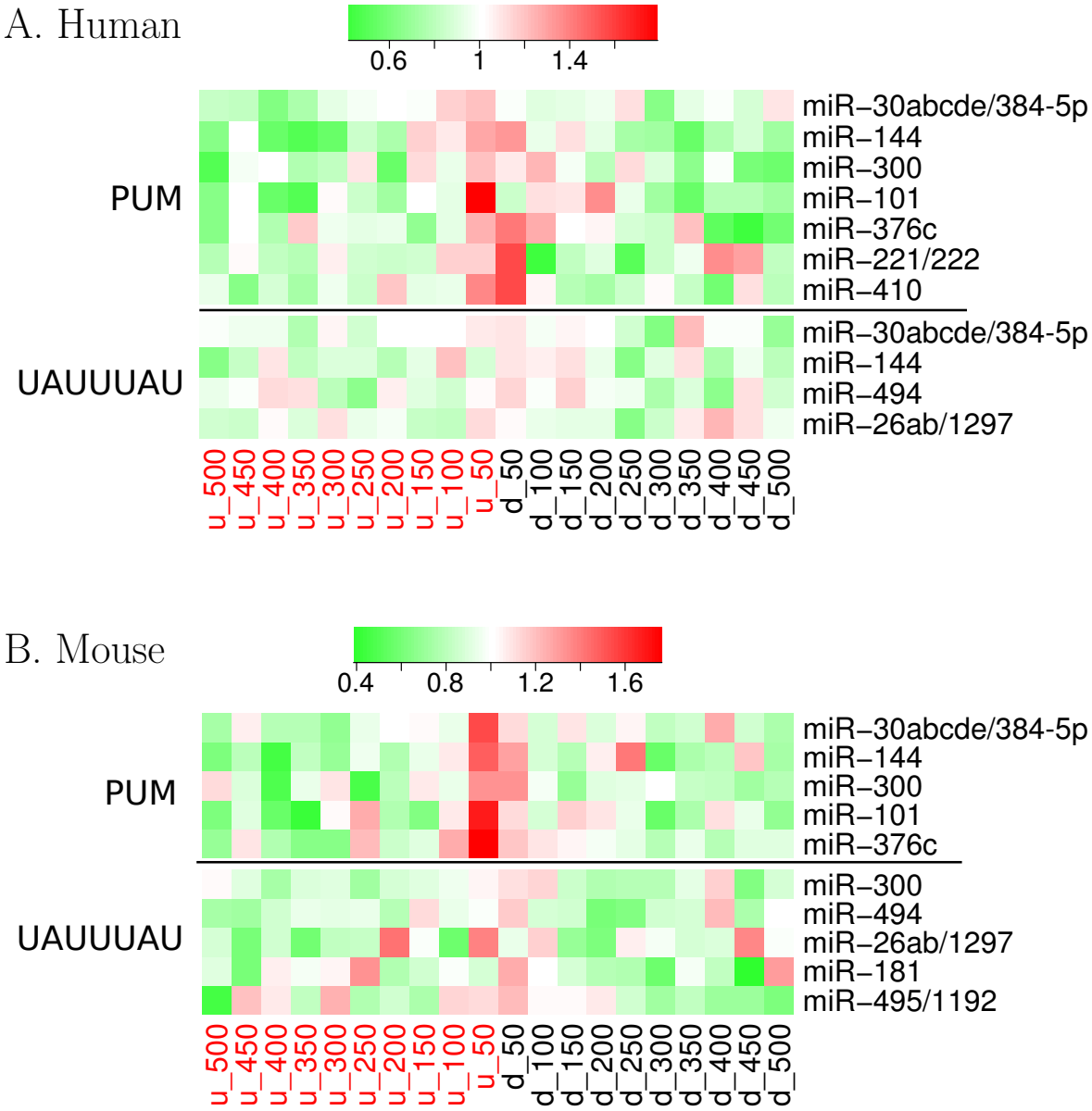



Figure 3.11: miRNA-RBP colocalization is not simply a consequence of AU-content. For RBPs and their interacting miRNAs (Figure 3.8), we considered each of ten 50-nt windows upstream and ten 50-nt windows downstream of a RBP binding motif. We determined the enrichment ratio of the number of miRNA recognition sites located in that window compared to the number of miRNA sites localized to shuffled RBP motifs with the same nucleotide content, normalized by their overall numbers across all 50-nt windows (Methods). The enrichment ratio in each window is shown in heatmap format. (A) Human enrichment heatmap. (B) Mouse enrichment heatmap.

We also assessed the possible effects of AU content on co-localization of miRNA and RBP recognition sites. AU content has been reported to affect miRNA site effectiveness [11]. Indeed, the recognition motifs for PUM and UAUUUAU and their co-localizing miRNA recognition seeds tend to be AU-rich (Table 3.2). To ensure that the co-occurrence observed between miRNAs and RBPs was not caused exclusively by the high AU composition of these motifs and their colocalization in AU-rich regions of 3′UTR, we evaluated shuffled RBP motifs with the same AU composition (Methods). For the two RBPs with co-occurring miRNAs, miRNA recognition motifs exhibited enrichment around true RBP recognition motifs compared to shuffled RBP control motifs generated to preserve AU content in the windows 50 nts up- or down-stream of the RBP (Figure 3.11). The signal remained strong for PUM in both mouse and human, but was weaker for UAUUUAU after this correction. Thus, the RBP-miRNA co-localization that we observed, especially for PUM, cannot be explained simply by the AU composition of the recognition sites.

| **RBP** | **Recognition motif** | **Interacting miRNA seeds** |
|---|---|---|
| PUM | 81.3% | 69.4% |
| UAUUUAU | 100% | 71.4% |

Table 3.2: AU composition of RBP recognition motifs and interacting miRNA recognition seeds. The fraction of nucleotides that are AU was calculated for each RBP recognition motif and the recognition seeds of miRNAs that interact with each RBP. The average value is shown. The average AU fraction for all miRNA seeds is 52.7%.

Using the same procedure as for human and mouse, no interacting miRNAs were predicted for fly and worm. We also determined the enrichment of miRNA recognition site density around each RBP site compared to the overall miRNA site density across all 3′UTRs with a RBP or miRNA recognition site. With this analysis, enrichment near the RBP recognition site for two RBP motifs was higher in human and mouse than the other two organisms (Table 3.3). However, since the quality of the 3′UTR annotations or the miRNA family member annotations may differ across organisms, more research will be needed to determine whether RBP-miRNA interactions are prevalent or limited to specific species.

| | **PUM** | **UAUUUAU** |
|---|---|---|
| H.sapiens | 1.12 | 1.14 |
| M.musculus | 1.12 | 1.06 |
| D.melanogaster | 0.87 | 0.95 |
| C.elegans | 1.07 | 0.87 |

Table 3.3: miRNA sites are enriched around RBP sites in human and mouse 3′UTRs. For each organism, the ratio of miRNA recognition site density 50 nts upstream or downstream of the RBP recognition sites to miRNA site density across all 3′UTRs is reported for PUM and UAUUUAU.

### 3.2.4 PUM and miRNAs cooperate to affect mRNA decay

We next examined the functional effect of RBPs and miRNAs on transcript decay using three genome-wide mRNA half-life datasets. These datasets are genomewide

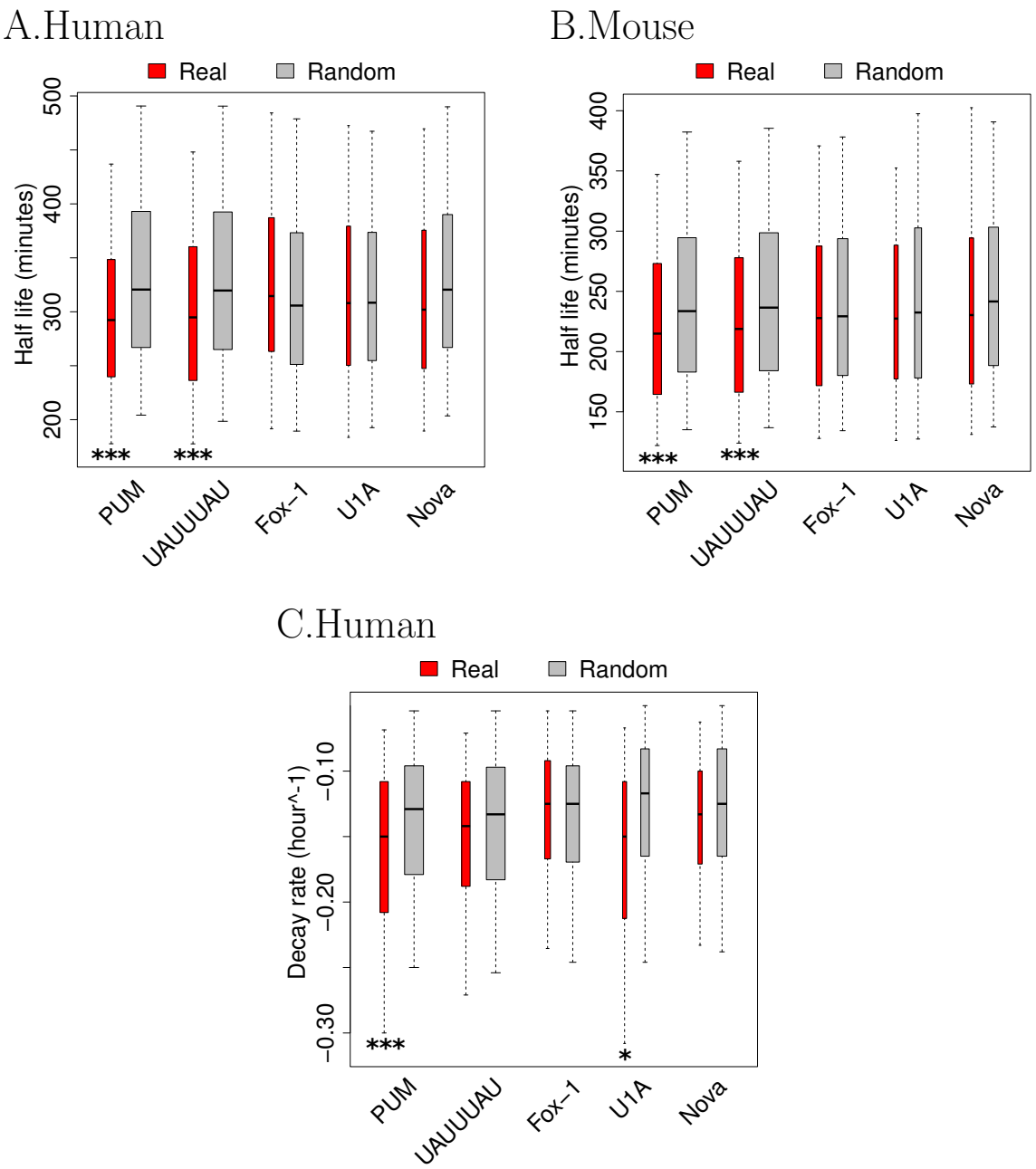



Figure 3.12: The presence of PUM and UAUUUAU results in faster transcript decay. Decay rates are shown in box-plots for transcripts with recognition sites for the designated RBP or shuffled RBP motifs. The Wilcoxon rank sum test with a Bonferroni correction was applied to measure the difference between real RBP sites (Real) and shuffled RBP controls (Random). For each RBP, an asterisk designates a significant difference between transcripts with RBP recognition sites and transcripts with shuffled RBP motif sites. One asterisk indicates a $P$-value $\leq 0.05$, two asterisks indicate a $P$-value $\leq 0.01$, and three asterisks indicate a $P$-value $\leq 0.001$. (A, B) Half-lives are based on the published dataset [139]. (C) Decay rates are based on the published dataset [140].

measurements of mRNA half-lives in human B cells and mouse fibroblasts [139] and mRNA decay rates in human HepG2 cells [140]. We determined the median half-life or decay rate for the set of transcripts that contain each of the recognition sites of interest. Transcripts containing a 3′UTR PUM or UAUUUAU site decayed faster than transcripts containing shuffled RBP motifs in all datasets (Figure 3.12). The presence of Fox-1, U1A or Nova recognition sites was not consistently associated with faster decay.

Having determined that for certain pairs of RBPs and miRNAs, their binding sites are frequently present in 3′UTRs in close proximity (Figure 3.8), we set out to further dissect the cooperativity between RBPs and miRNAs in mediating transcript decay. For each RBP, we divided all transcripts into four categories: Int-proximal: transcripts containing RBP recognition sites, and within 50 nts, a miRNA recognition site for one of the interacting miRNAs for that RBP; Int-distant: transcripts containing RBP recognition sites and miRNA recognition sites for one of the interacting miRNAs for that RBP, but none of the miRNA recognition sites and RBP recognition sites are within 50 nts of each other; Nonint-proximal: transcripts containing RBP recognition sites with at least one miRNA recognition site within 50 nts, but the miRNA is not an interacting miRNA for that RBP; Nonint-distant: transcripts containing RBP recognition sites with at least one non-interacting miRNA recognition site, but none of the RBP recognition sites and miRNA recognition sites are within 50 nts of each other.

For each RBP, mRNA half-life values or decay rates determined experimentally by Friedel [139] and Yang [140] were considered for all transcripts in each of the four classes of transcripts defined above. For each mRNA decay dataset, data from small RNA sequencing experiments in the same cell line were used to define the set of miRNAs expressed, and only those miRNAs in the top 25% most sequenced miRNAs were considered for further analysis [141, 142, 143]. For PUM, the presence of nearby

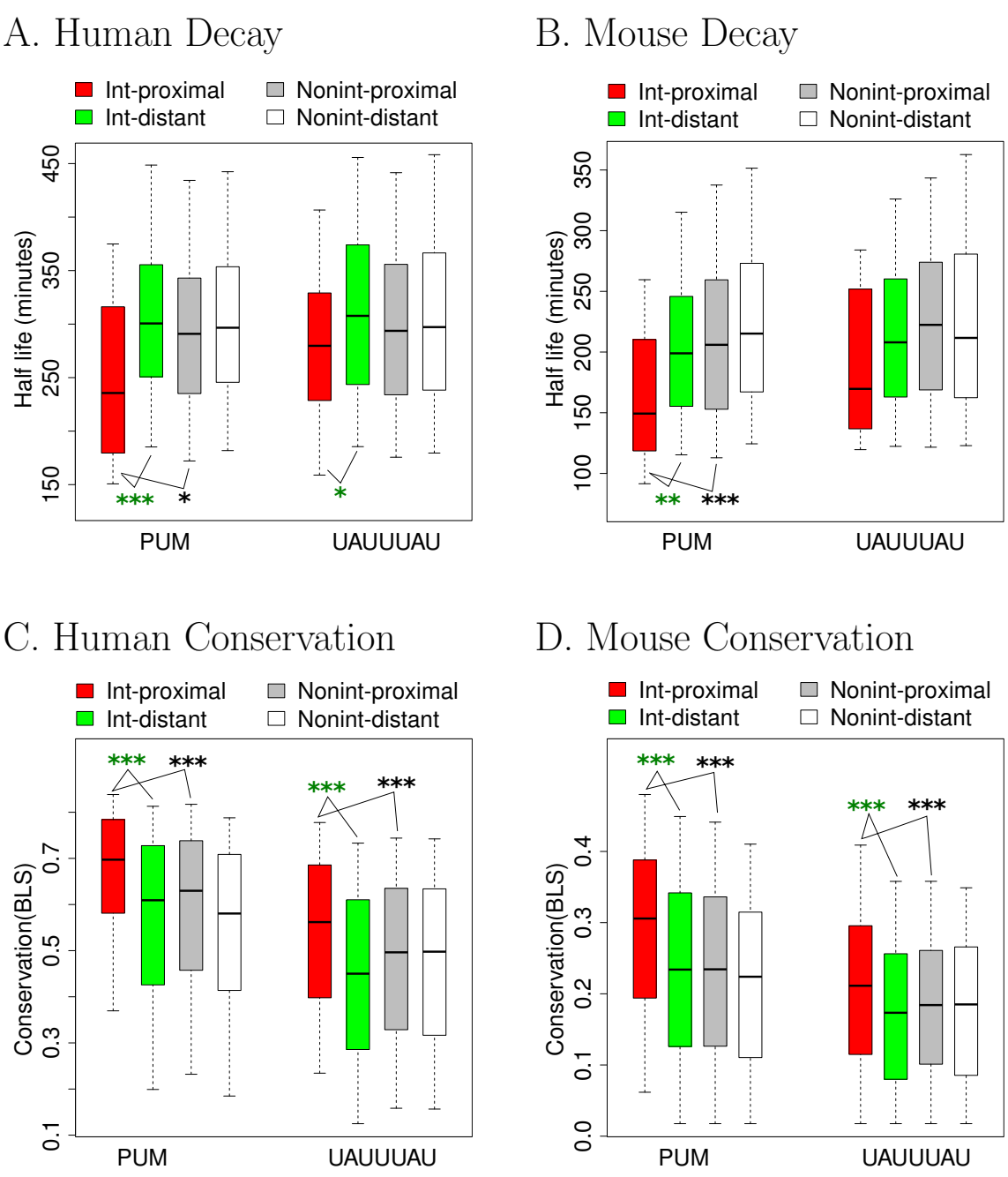



Figure 3.13: PUM recognition sites promote decay more effectively and are better conserved when present with interacting miRNAs. (caption continued on next page.)

Figure 3.13: (previous page) Transcripts with a specific RBP recognition site were divided into four groups. Group Int-proximal contained transcripts with at least one RBP site and its interacting miRNA recognition site within 50 nts. Group Int-distant contained transcripts with both a RBP recognition site and a recognition site for its interacting miRNA, but no pair of RBP-miRNA site is within 50 nts. Group Nonint-proximal and Nonint-distant were similar to group Int-proximal or Int-distant except non-interacting miRNAs (not predicted in Figure 3.8) were analyzed. For each group of transcripts, the half lives (or conservation scores) were ranked and the (25%, 75%) range of the data were extracted and plotted with box-plots for visualization. The bottom and top of the box are the 25th and 75th percentiles (the inter-quartile range). Whiskers on the top and bottom represent the maximum and minimum data points within the range represented by 1.5 times the inter-quartile range. For each RBP, asterisks represent comparisons of half-life (or conservation score) between Int-proximal and Nonint-proximal or between Int-proximal and Int-distant by Wilcoxon rank sum test on the full range of data. One asterisk indicates a $P$-value $\leq 0.05$, two asterisks indicate a $P$-value $\leq 0.01$, and three asterisks indicate a $P$-value $\leq 0.001$. (A, B) Half-lives for mRNAs are plotted for human and mouse [139]. (C, D) Conservation BLS scores for RBP recognition sites are plotted.

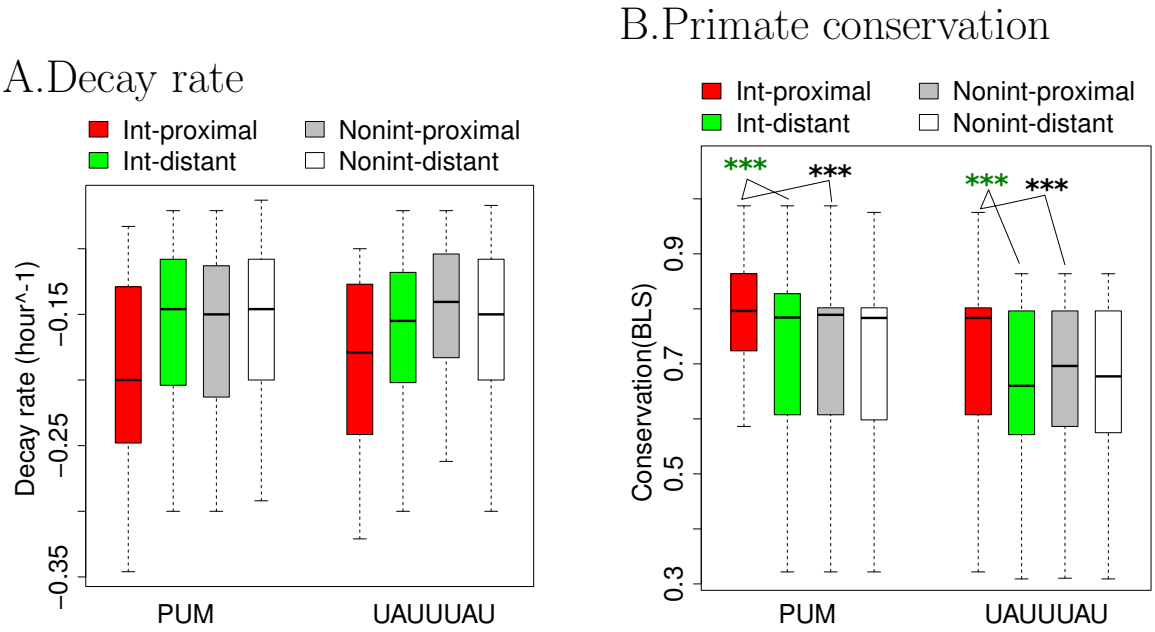



Figure 3.14: PUM recognition sites promote decay more effectively and are better conserved when present with interacting miRNAs. For each RBP, miRNAs were classified into four groups as described for Figure 3.13. (A) Decay rates were plotted based on dataset [140] as described in Figure 3.13A, B. (B) Conservation BLS scores were calculated based on ten primate species alignment, and plotted as described in Figure 3.13C, D.

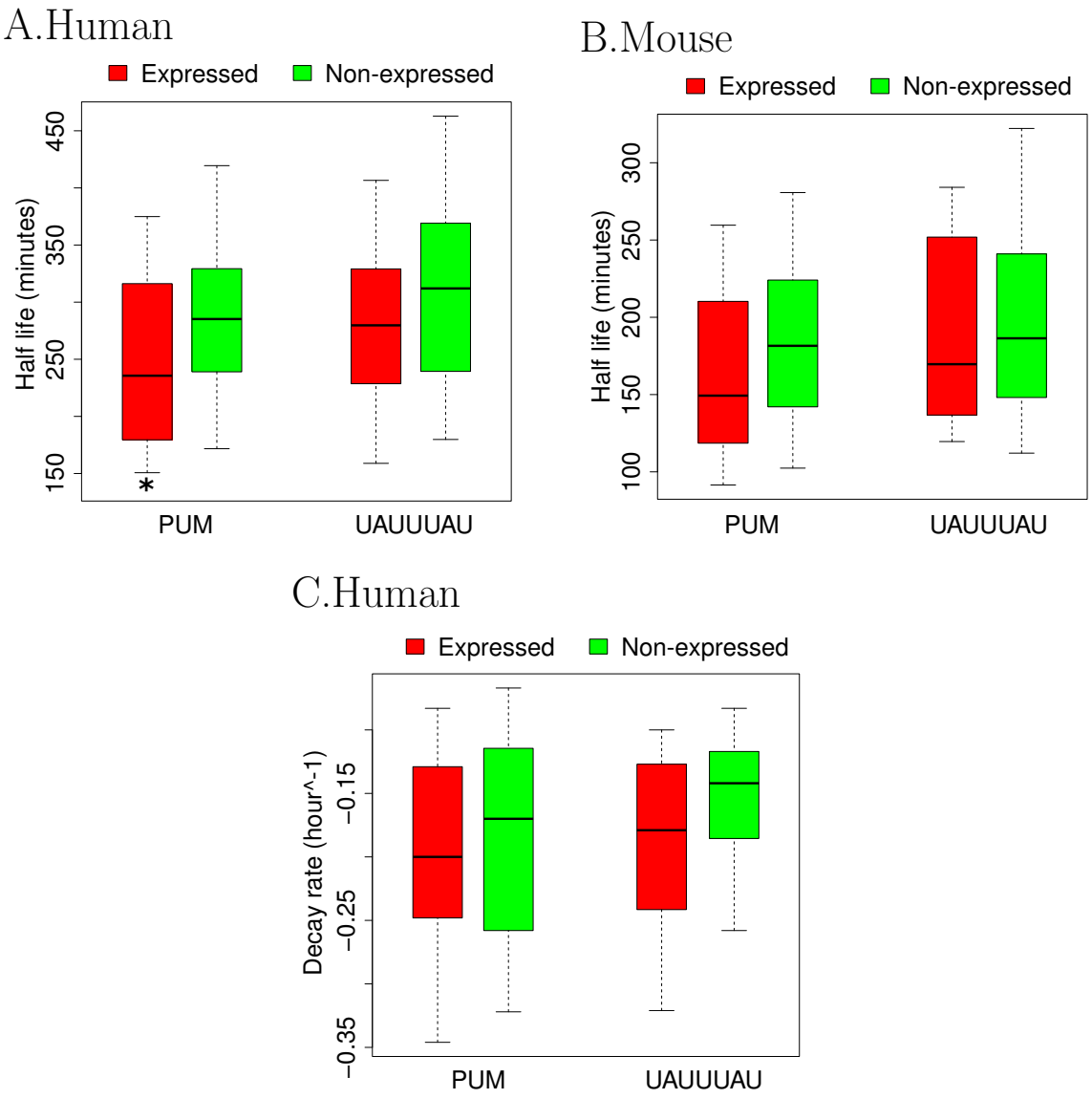



Figure 3.15: Expressed miRNAs promote decay more effectively than non-expressed miRNAs. For each of the cell lines used in half-life or decay rate datasets, companion small RNA sequencing datasets were identified from the literature. In each dataset, the reads of miRNAs were ranked and the most frequently expressed 25% of small RNA reads was established as a threshold for classifying interacting miRNAs for each RBP as Expressed or Non-expressed. Transcripts with proximal miRNA and RBP sites were compared with respect to their half-lives or decay rates. Asterisks represent comparisons of half-lives between two groups determined by Wilcoxon rank sum tests. One asterisk indicates a $P$-value $\leq 0.05$, two asterisks indicate a $P$-value $\leq 0.01$, and three asterisks indicate a $P$-value $\leq 0.001$. (A) For mRNA half-lives measured in Human B cells (BL41) [139], miRNA sequencing reads were derived from a dataset generated by Landgraf and colleagues [141]. (caption continued on next page.)

Figure 3.15: (previous page) (B) For mRNA half-lives measured in mouse fibroblasts (NIH-3T3) [139], miRNA sequencing reads were derived from a dataset generated by Zhu and colleagues [142]. (C) For mRNA decay rates measured in human HepG2 [140], miRNA sequencing reads were derived from a dataset generated by the ENCODE project [143].

interacting miRNA recognition sites, but not distant miRNA sites or nearby non-interacting miRNA sites, consistently increased the decay rate in both human B cells and mouse fibroblasts [139] (Figure 3.13A, B and Table 3.4). Similar results were also observed in an independently derived human mRNA decay rate dataset [140] (Figure 3.14A). Moreover for transcripts with recognition sites for PUM and its interacting miRNAs within 50 nts, expressed miRNAs promote decay consistently faster than non-expressed miRNAs (Figure 3.15).

We also tested whether the more rapid decay observed in transcripts with recognition sites for both PUM and miRNAs was a consequence of the high AU-content of the PUM recognition sites and the recognition sites of its interacting miRNAs. We utilized shuffled control motifs of PUM and miRNAs that have the same AU-content as the real motif. We established three groups of transcripts according to the presence of PUM and miRNA recognition sites on 3′UTRs: (Real), real RBP recognition sites and real recognition sites for interacting miRNAs within 50 nts; (miR_control), real RBP recognition sites and sites for shuffled interacting miRNA motifs within 50 nts; and (RBP_control), shuffled RBP motif sites and real interacting miRNA sites within 50 nts. We observed that transcripts with real PUM and interacting miRNA recognition sites have consistently shorter half-lives compared to transcripts in the two other control groups (Figure 3.16). Thus, the more rapid decay rate observed in 3′UTRs with interacting PUM and miRNA recognition sites is not simply a consequence of the high AU content of recognition motifs of PUM and its interacting miRNAs.

For UAUUUAU, transcripts with both UAUUUAU sites and recognition sites of its interacting miRNAs in close proximity tend to have shorter mRNA half lives

than transcripts in other groups (Figure 3.13A, B, Figure 3.14, Figure 3.15 and Figure 3.16). However, the effect is less strong than the effect observed for PUM.

| RBP | miRNA | Half life (minutes) | | Number of mRNAs | | P-value |
|---|---|---|---|---|---|---|
| | | Proximal | Distant | Proximal | Distant | |
| PUM | miR-30abcde/384-5p | 214.7 | 302.1 | 55 | 605 | 1.21E-2 |
| | miR-101 | 254.6 | 300.2 | 88 | 585 | 9.06E-3 |
| UAUUUAU | miR-30abcde/384-5p | 262.4 | 304.8 | 69 | 691 | 2.78E-3 |
| | miR-26ab/1297 | 309.0 | 310.3 | 64 | 696 | 4.29E-1 |

(a) Human Friedel

| RBP | miRNA | Half life (minutes) | | Number of mRNAs | | P-value |
|---|---|---|---|---|---|---|
| | | Proximal | Distant | Proximal | Distant | |
| PUM | miR-30abcde/384-5p | 138.8 | 187.1 | 57 | 539 | 5.76E-2 |
| | miR-101 | 153.3 | 195.9 | 72 | 557 | 4.03E-2 |
| UAUUUAU | miR-26ab/1297 | 169.65 | 208.0 | 64 | 565 | 2.18E-1 |

(b) Mouse Friedel

| RBP | miRNA | Decay rate (hour^-1) | | Number of mRNAs | | P-value |
|---|---|---|---|---|---|---|
| | | Proximal | Distant | Proximal | Distant | |
| PUM | miR-410 | -0.221 | -0.15 | 48 | 273 | 1.45E-2 |
| | miR-376c | -0.22625 | -0.15 | 32 | 165 | 9.95E-2 |
| | miR-30abcde/384-5p | -0.146 | -0.179 | 21 | 251 | 2.50E-1 |
| | miR-101 | -0.1405 | -0.15 | 34 | 241 | 4.08E-1 |
| UAUUUAU | miR-30abcde/384-5p | -0.233 | -0.163 | 21 | 271 | 3.20E-1 |
| | miR-26ab/1297 | -0.179 | -0.15 | 25 | 286 | 1.82E-1 |

(c) Human Yang

Table 3.4: Expressed RBP-interacting miRNAs and their effects on mRNA decay. For each of the mRNA decay datasets, we considered each RBP and its expressed interacting miRNAs (corresponding to Figure 3.13A, B). Proximal and distant pairs of recognition sites were defined as in Figure 3.13. The median mRNA half-lives or decay rates were shown for each combination. P-values were calculated from Wilcoxon rank sum test as a measure of the difference between proximal and distant group. (A) Results from mRNA half-life datasets of human B cells [139]. (B) Results from mRNA half-life datasets of mouse fibroblasts [139]. (C) Results from mRNA decay rate dataset of HepG2 cell line [140].

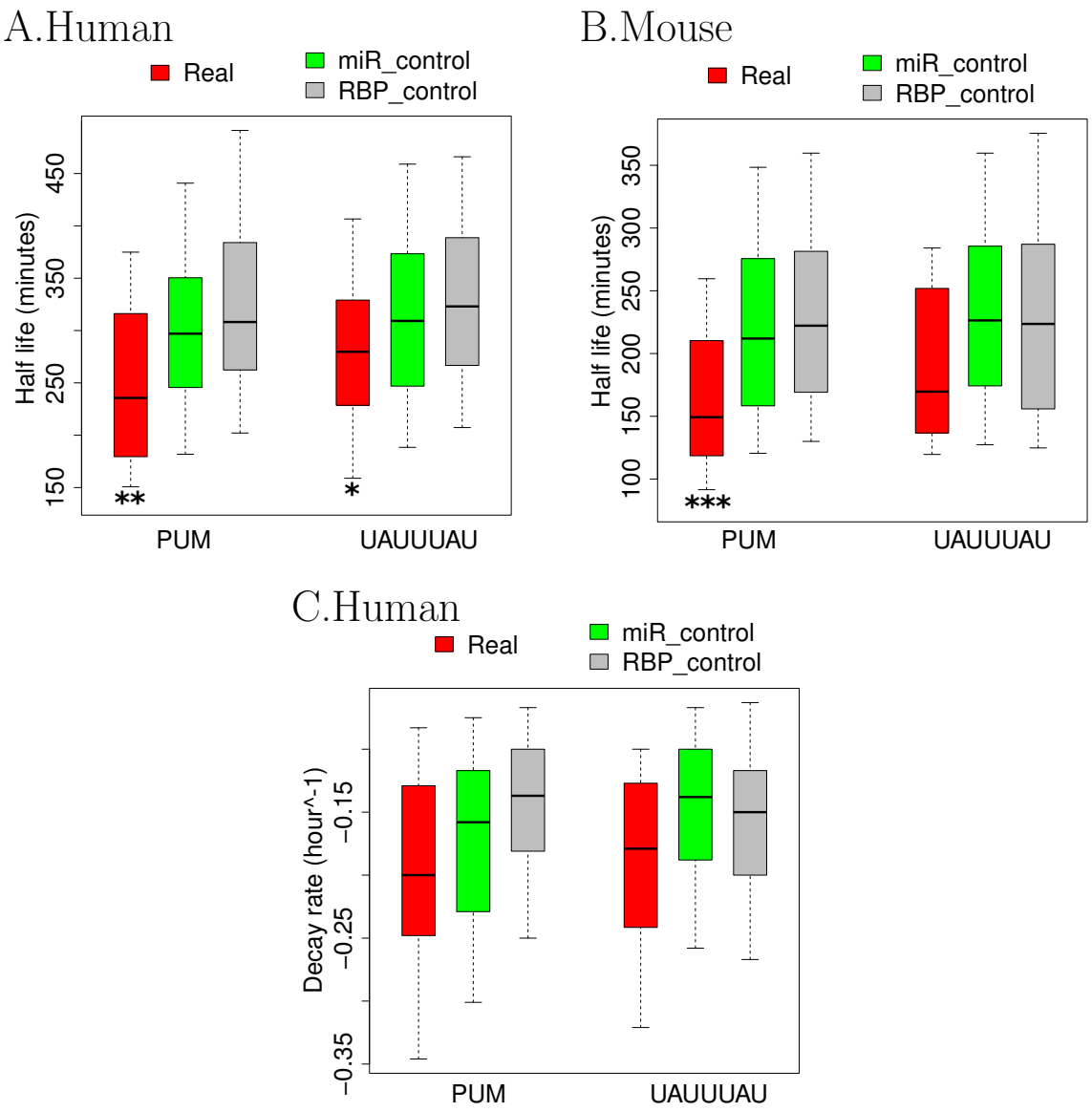



Figure 3.16: More rapid mRNA decay in transcripts in which PUM and interacting miRNA recognition sites colocalize is not only a consequence of AU-content. For each RBP or miRNA recognition motif, the shuffled RBP or miRNA motifs were used as controls for AU content. For each group of transcripts, boxplots of half-lives or decay rates were plotted as described for Figure 3.13AB. Group Real contained transcripts with at least one RBP recognition site and a recognition site for one of the RBP′s interacting miRNA within 50 nts. Group miR control contained transcripts with at least one RBP recognition site and a recognition site of shuffled interacting miRNA motif within 50 nts. Group RBP control contained transcripts with at least one recognition site of a shuffled RBP motif and an associated interacting miRNA recognition site within 50 nts. Asterisks represent comparisons of half-lives between the groups Real and miR control determined by Wilcoxon rank sum tests. (caption continued on next page.)

Figure 3.16: (previous page) One asterisk indicates a $P$-value $\leq 0.05$, two asterisks indicate a $P$-value $\leq 0.01$, and three asterisks indicate a $P$-value $\leq 0.001$. (A, B) Half-life data from Friedel and colleagues [139], (C) Decay rate data from Yang and colleagues [140].

In addition to analyzing mRNA decay, we also extended our observations to evolutionary conservation. We discovered that for both PUM and UAUUUAU, recognition sites that are located within 50 bps of an interacting miRNA are better conserved than recognition sites located more than 50 bps from an interacting miRNA or within 50 bps of a non-interacting miRNA in both human and mouse (Figure 3.13C, D, Figure 3.14B and Table 3.5).

| RBP | miRNA | Conservation BLS | | Number of RBP sites | | P-value |
|---|---|---|---|---|---|---|
| | | Proximal | Distant | Proximal | Distant | |
| PUM | miR-30abcde/384-5p | 0.686 | 0.630 | 147 | 1773 | 7.00E-2 |
| | miR-144 | 0.702 | 0.649 | 227 | 2003 | 7.12E-3 |
| | miR-300 | 0.767 | 0.633 | 236 | 2299 | 9.03E-6 |
| | miR-101 | 0.712 | 0.657 | 183 | 1765 | 7.38E-4 |
| | miR-376c | 0.706 | 0.607 | 130 | 1404 | 3.01E-3 |
| | miR-221/222 | 0.722 | 0.646 | 92 | 1212 | 2.41E-2 |
| | miR-410 | 0.654 | 0.604 | 230 | 1971 | 2.74E-2 |
| UAUUUAU | miR-30abcde/384-5p | 0.612 | 0.462 | 208 | 2258 | 1.39E-4 |
| | miR-144 | 0.626 | 0.475 | 215 | 2365 | 1.42E-5 |
| | miR-494 | 0.511 | 0.453 | 244 | 2865 | 2.81E-1 |
| | miR-26ab/1297 | 0.485 | 0.443 | 167 | 2188 | 2.26E-1 |

(a) Human

| RBP | miRNA | Conservation BLS | | Number of RBP sites | | P-value |
|---|---|---|---|---|---|---|
| | | Proximal | Distant | Proximal | Distant | |
| PUM | miR-30abcde/384-5p | 0.297 | 0.262 | 146 | 1450 | 4.18E-1 |
| | miR-144 | 0.338 | 0.252 | 180 | 1586 | 1.80E-3 |
| | miR-300 | 0.336 | 0.247 | 222 | 1800 | 3.13E-4 |
| | miR-101 | 0.354 | 0.266 | 153 | 1526 | 2.57E-3 |
| | miR-376c | 0.253 | 0.207 | 107 | 1084 | 1.43E-1 |
| UAUUUAU | miR-300 | 0.257 | 0.183 | 189 | 1770 | 1.75E-4 |
| | miR-494 | 0.203 | 0.166 | 165 | 1812 | 6.21E-2 |
| | miR-26ab/1297 | 0.181 | 0.165 | 141 | 1533 | 1.02E-1 |
| | miR-181 | 0.271 | 0.174 | 152 | 1653 | 1.74E-5 |
| | miR-495/1192 | 0.198 | 0.176 | 272 | 2210 | 3.45E-1 |

(b) Mouse

Table 3.5: RBP recognition sites are more conserved when present with interacting miRNAs. For each RBP and its interacting miRNAs, their neighbor recognition sites were classified into either the proximal or distant groups as described for Figure 3.13. The median conservation BLS scores are shown for each combination and $P$-values are calculated based on Wilcoxon rank sum tests as a measure of the difference between the two groups. (A) Human-interacting miRNAs are shown. (B) Mouse-interacting miRNAs are shown.

We also ran Gene Ontology enrichment analysis for human genes with colocalized PUM and interacting miRNA recognition sites in their 3′UTRs. We found that GO categories related to transcriptional regulation were enriched (Table 3.6). Thus, it is possible that the synergistic effects of PUM and miRNAs on mRNA decay rate will subsequently affect the initiation of transcription for genes.

| GO Biological process | #genes | P-value |
|---|---|---|
| protein binding transcription factor activity | 38 | 4.09E-4 |
| transcription factor binding transcription factor activity | 37 | 5.36E-4 |
| transcription cofactor activity | 36 | 5.39E-4 |

Table 3.6: GO enrichments for PUM and its interacting miRNAs. Transcripts were classified into categories as in Figure 3.13. The GO biological process annotations were compared between group Int-proximal (transcripts with at least one RBP site and its interacting miRNA recognition site within 50 nts) and group Int-distant (transcripts with both RBP sites and its interacting miRNA recognition sites, but no pair of recognition sites is within 50 nts). Hypergeometric enrichment was used to calculate $P$-values. We then applied the Benjamini-Hochberg procedure on the $P$-values, and selected enriched GO terms with FDR$\leq$0.05. The number of annotated genes and hypergeometric $P$-values are shown for each significant GO term.

### 3.2.5 PUM rescues recognition site accessibility for PUM-interacting miRNAs

We further investigated why a specific group of miRNA recognition sites tend to be localized proximal to PUM recognition sites and promote decay. Previous studies have reported that for miR-221/222 and miR-410, PUM can alleviate the constraints of RNA secondary structure and make miRNA binding sites more accessible to the RISC complex [28, 52]. We hypothesized that the genome-wide co-occurrence of PUM and a specific set of miRNAs is related to the ability of PUM to rescue miRNA recognition site accessibility.

To address this issue on a genome-wide scale, we used a computational approach to estimate the frequency of RBP regulation of local 3′UTR secondary structure. For

each pair of neighboring RBP-miRNA recognition sites, we determined the number of base pairs of miRNA recognition site that RBP binding can rescue from pairing with other nucleotides within the 3′UTR as estimated by RNAfold (Methods) [52, 144, 145]. As an example, when we used RNAfold to determine the secondary structure for the p27Kip1 3′UTR, we discovered that 6 out of 7 base pairs of the miR-221/222 recognition seed site were hybridized to other nucleotides and therefore inaccessible due to the sequence′s secondary structure. When we simulated PUM binding by converting all of the bases in the PUM recognition sites to N's, and thus made them unavailable to hybridize to other bases in the sequence, 0 base pairs of the miR-221/222 seed site were blocked. We calculate the amount of miRNA site rescue as 6 - 0 = 6.

For each RBP, we plotted the histogram of miRNA site rescue counts for RBP sites in close proximity to recognition sites for interacting miRNAs, and non-interacting miRNAs (Figure 3.17A, B and Figure 3.18A, B). When we performed this analysis for PUM, interacting miRNAs produced significantly higher rescue counts than non-interacting miRNAs in both human and mouse (Figure 3.17A, B, Wilcoxon $P$-value $<$ 1E-10 for both human and mouse). We also performed a control in which, for interacting miRNAs, the sequence of the miRNA and RBP recognition sites were shuffled while preserving mono and di-nucleotide frequency [146, 147]. For PUM, the proportion of RBP recognition sites with large rescue counts was consistently higher in the true histogram than in the background model, while the proportion of smaller rescue counts was depleted (Figure 3.17C, D and Figure 3.19, Wilcoxon $P$-value $<$ 1E-10 for both human and mouse). For UAUUUAU, we did not observe any enrichment of high rescue counts for interacting miRNAs compared with controls (Figure 3.18). In summary, miRNA recognition sites located near a PUM site have a significantly increased frequency of high recognition site rescue by simulated PUM binding than expected by chance.

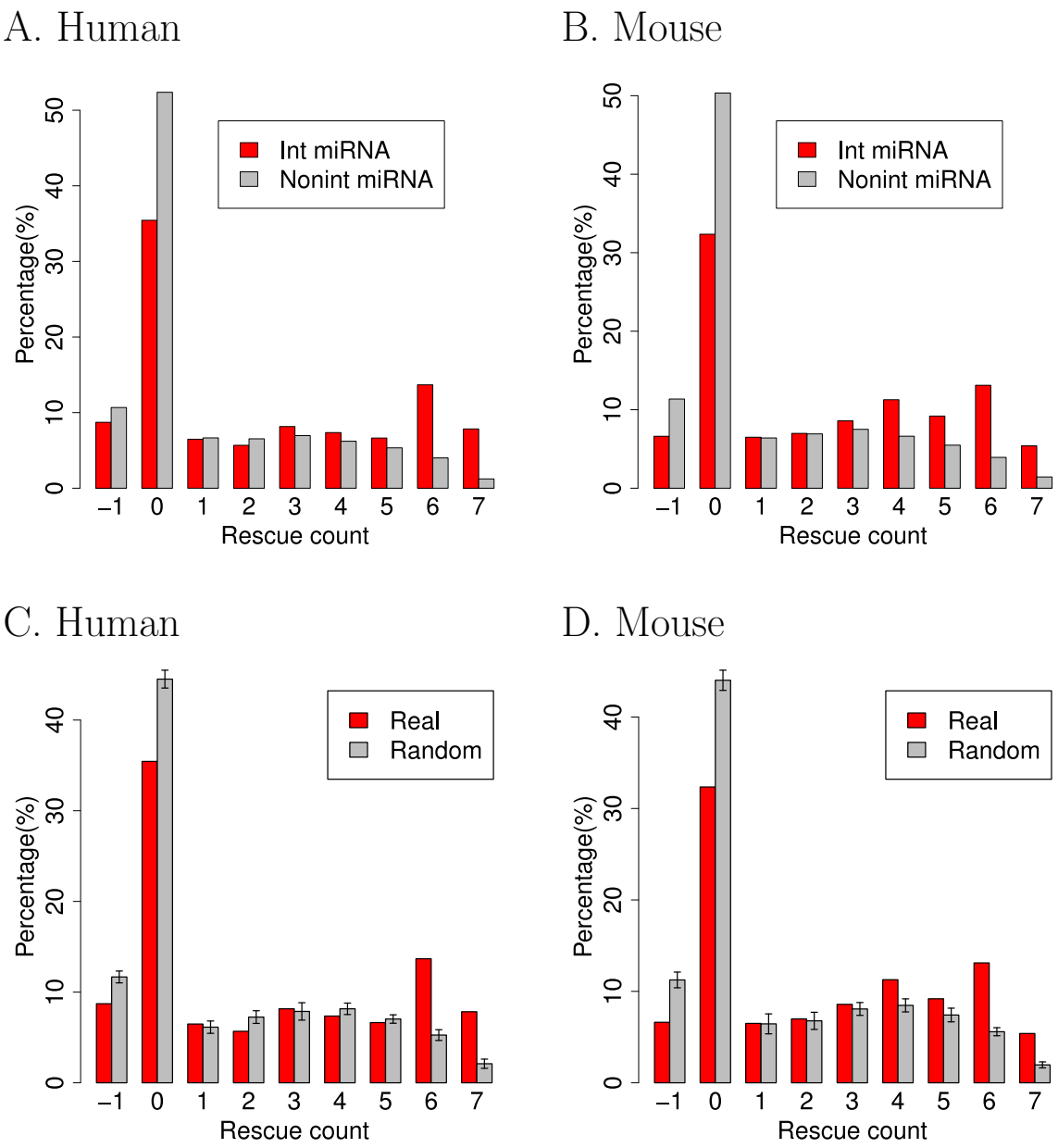



Figure 3.17: PUM rescues nucleotides in neighboring interacting miRNA recognition sites. (A, B) For each PUM recognition site with a neighboring miRNA recognition site within 50 nts, the rescue count was computationally estimated as the number of nucleotides in the miRNA recognition site that PUM binding frees from hybridization with other nucleotides. The distributions of miRNA site rescue counts are shown in histograms for interacting miRNAs (red) and non-interacting miRNAs (gray) in human and mouse. (C, D) For all interacting miRNAs of PUM, the background model (Random) represents the histogram generated when RBP-miRNA paired site sequences were randomly shuffled while preserving mono and di-nucleotide frequency. Standard deviations were estimated from 10 randomizations.

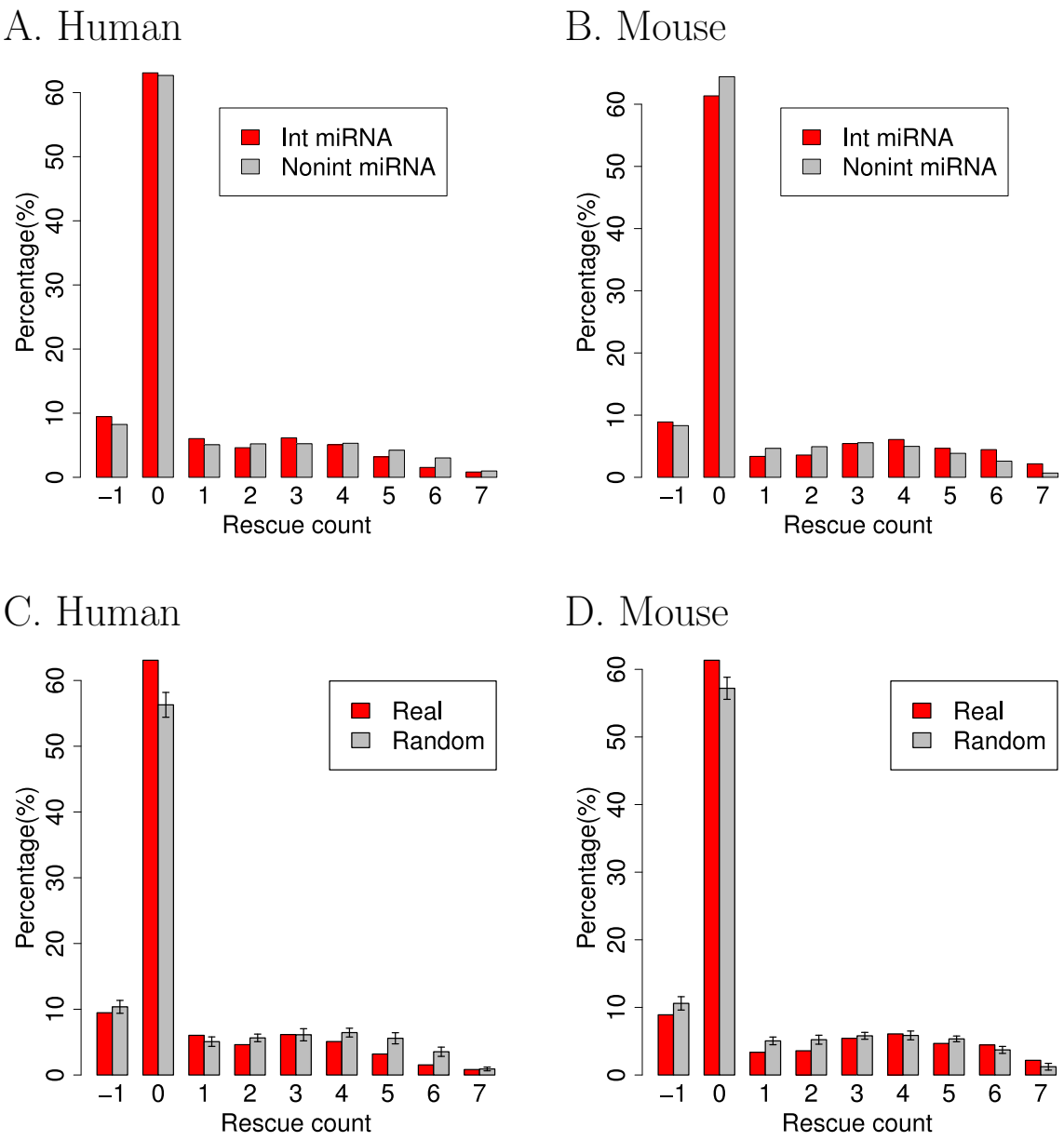



Figure 3.18: Histograms of rescue counts for miRNA recognition sites upon UAUUUAU binding. Histograms of site rescue were calculated for both UAUUUAU-interacting miRNAs and non-interacting miRNAs as described in Figure 3.17. (A, B) Comparison of histograms between interacting miRNAs and non-interacting miRNAs are shown separately for human and mouse. (C, D) Comparison of histograms between real rescue counts and random rescue counts determined based on a background model are shown for human and mouse.

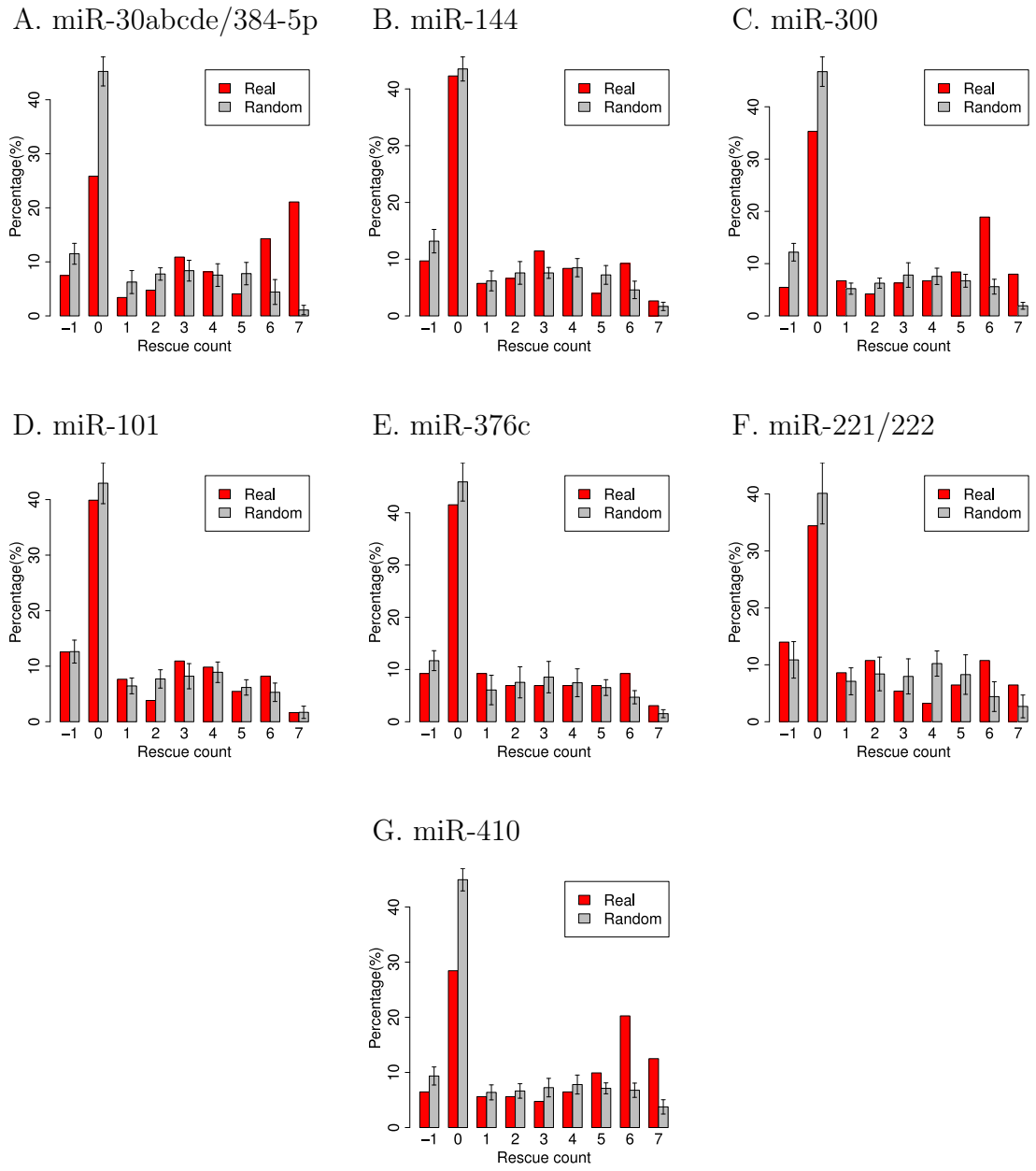



Figure 3.19: PUM rescues nucleotides in neighboring interacting miRNA recognition sites. The histogram of miRNA seed rescue was compared with rescue values calculated with the background model of Figure 3.17C, D. Data for each of the seven interacting miRNAs of PUM are shown separately for the human genome. (A) miR-30abcde/384-5p. (B) miR-144. (C) miR-300. (D) miR-101. (E) miR-376c. (F) miR-221/222. (G) miR-410.

### 3.2.6 miRNAs that interact with PUM have recognition seeds reverse complementary to the PUM recognition motif

We derived a score to measure the ability of miRNA recognition seed sequences to hybridize with the reverse PUM recognition motif, an association that would result in RNA hairpin loop structures in the target mRNA, based on sequence alignment [148] (Figure 3.20A). A larger score indicates that there is more nucleotide complementarity between the miRNA seed sequence and the reverse of the PUM recognition motif. We found that PUM-interacting miRNAs have significantly higher alignment scores than non-interacting miRNAs in both human and mouse (Figure 3.20B, C). Thus, if a miRNA co-occurs with PUM recognition sites, it has a higher potential to pair up with the reverse PUM sequence. For UAUUUAU, there was no difference in the alignment scores for interacting versus non-interacting miRNAs (Figure 3.20B, C).

By comparing real and shuffled PUM motifs, we found that the reverse recognition motifs for PUM tend to have larger alignment scores with interacting miRNA seed sequences than shuffled PUM motifs (Figure 3.21). Thus, the observed enrichment of higher miRNA rescue counts for PUM is likely to derive from its reverse complementarity with a specific group of miRNA seeds that also have recognition sites preferentially co-localized with PUM recognition sites. For all interacting miRNAs in human or mouse, we diagrammed their 8mer seed (7mer+1A [11]) sequences aligned to the reverse PUM motif (Figure 3.20A).

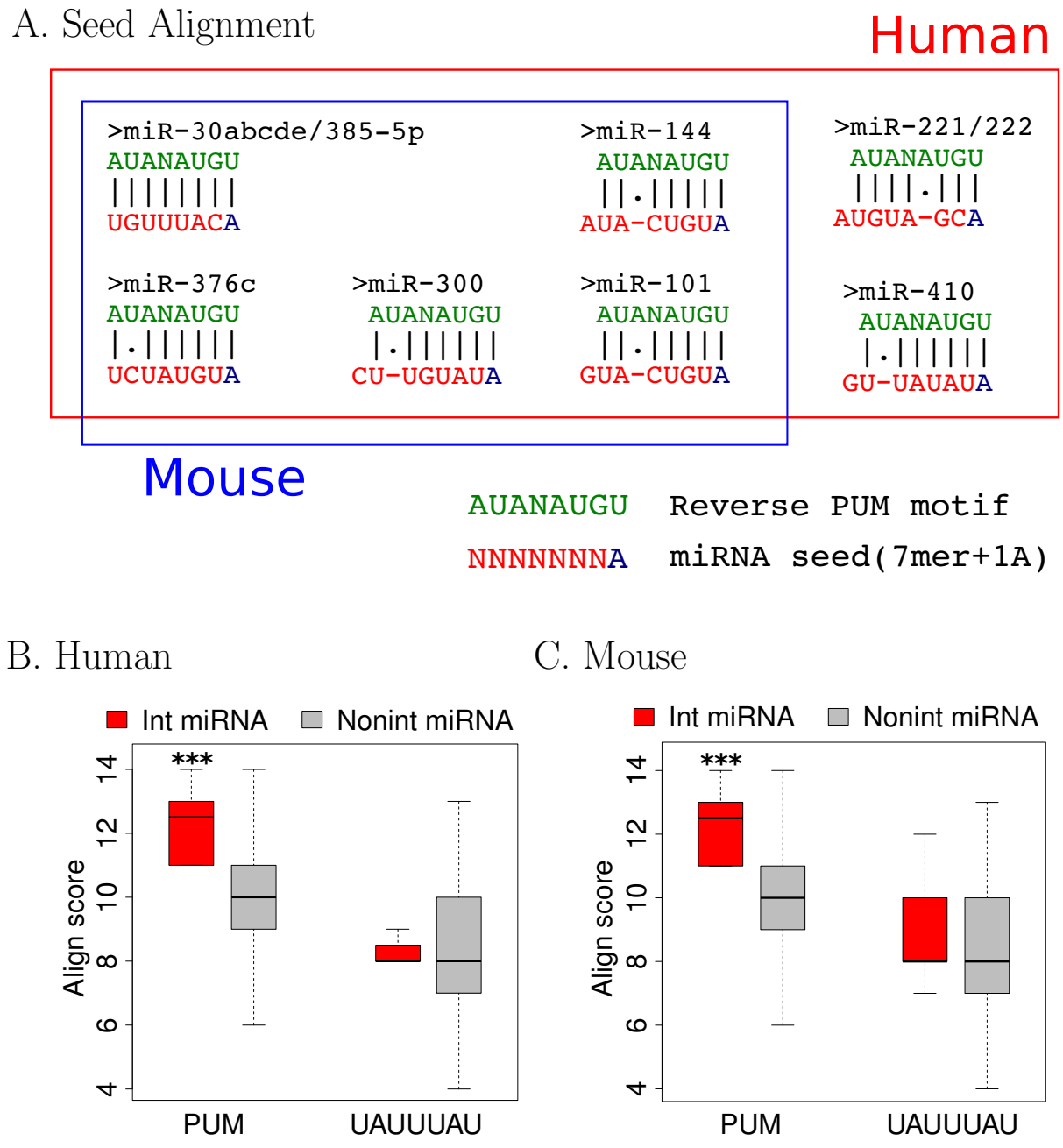



Figure 3.20: PUM-interacting miRNAs have seed sequence complementarity to the reverse PUM recognition motif. (caption continued on next page.)

Figure 3.20: (previous page) (A) Optimal complementary alignments between miRNA recognition seed sequences and the reverse PUM motif are shown for interacting miRNAs in both human and mouse (Figure 3.8). Nucleotides 2 to 8 of the miRNA seed site sequence are highlighted in red and the first adenosine position is colored in blue [11]. The reverse PUM recognition motif is colored in green. (B, C) For each miRNA seed, a score was determined based on the extent of complementary base pairing with the reverse RBP recognition motifs [148]. Higher scores indicate better matches with the reverse complementary RBP motif. For each RBP, box-plots of alignment scores are shown for interacting miRNA seeds and non-interacting miRNA seeds. The bottom and top of the box are the 25th and 75th percentiles (the inter-quartile range). Whiskers on the top and bottom represent the maximum and minimum data points within the range represented by 1.5 times the inter-quartile range. For each RBP, asterisks represent comparisons of alignment scores between Int miRNA and Nonint miRNA by Wilcoxon rank sum test. One asterisk indicates a $P$-value $\leq 0.05$, two asterisks indicate a $P$-value $\leq 0.01$, and three asterisks indicate a $P$-value $\leq 0.001$.

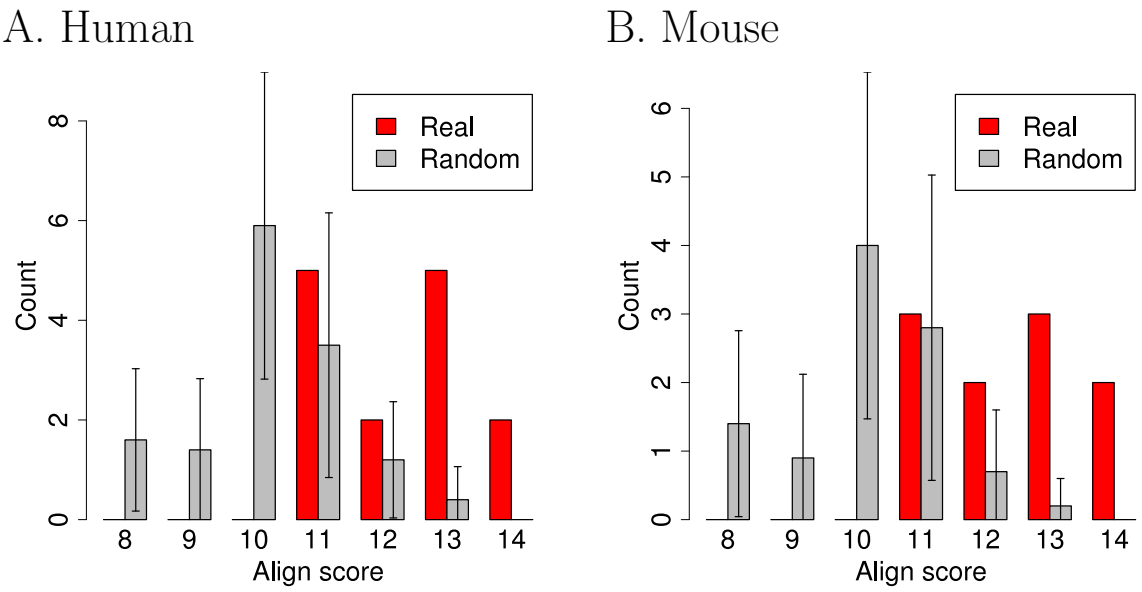



Figure 3.21: Histograms of alignment scores between miRNA seeds and PUM recognition motifs. For all interacting miRNAs of PUM, their seed alignment scores with the reverse PUM motif were determined for real PUM motif and shuffled PUM motifs. Histograms for all shuffled PUM motifs were merged into average values and standard deviations, and are shown for PUM interacting miRNAs identified in (A) Human and (B) Mouse.

## 3.3 Discussion

### 3.3.1 Prevalent models for RBP-miRNA interactions

Some previous studies on RBP-miRNA interactions have experimentally demonstrated specific instances in which RBPs and miRNAs compete with each other, sometimes for the same binding site [106, 109, 110]. In this model, the presence of a RBP recognition site would protect the associated transcript from miRNA-mediated decay and stabilize it. However, this mode of interaction does not seem to be prevalent among the RBP-miRNA interactions we uncovered from our transcriptome-wide analysis as the presence of both a recognition site for a RBP and a miRNA did not result in a global shift toward more stable transcripts using the methodology we employed.

Another model for RBP-miRNA interactions involves RBPs binding closely to miRNA sites and altering the local secondary structure to make miRNA sites more accessible to the RISC complex [28, 52]. When PUM was computationally folded with nearby miRNA recognition sites, the presence of the RBP resulted in increased availability of the miRNA recognition sites. For PUM, the rescue counts were higher for the interacting miRNA sites than for non-interacting miRNAs and background models (Figure 3.17).

For the PUM recognition site, we are able to develop a computational model to explain the miRNA-specific interactions based on reverse complementarity between the recognition seeds for miRNAs and the PUM recognition motif, which is advantageous for formation of hairpin loops (Figure 3.20). A previous report described a case study in which miR-221/222 pairs up with the PUM recognition sequence to achieve condition-specific miRNA-mediated decay of the p27Kip1, based on PUM expression and its phosphorylation state [28]. Our analysis indicates that the mechanism described for this particular case may also occur for other miRNAs. Further, selective

pressure for this mechanism may have shaped the localization pattern of a group of miRNAs by enriching them to be close to PUM recognition sites. Regulation of the levels or activity of RBPs represents a previously unappreciated mechanism for increasing or decreasing the efficiency of many miRNA binding sites simultaneously.

For the ARE element UAUUUAU, we identified a group of miRNAs that are enriched in their co-localization with its recognition sites (Figure 3.8). These sites may have a function because UAUUUAU motifs are more evolutionarily conserved if they are proximal to an interacting miRNA recognition site (Figure 3.13C, D) and they did promote more rapid decay, although the effect was not as significant as the effect observed for PUM (Figure 3.13A, B). However, the presence of UAUUUAU demonstrated no capacity to rescue miRNA binding sites from secondary structure. Our data suggest that UAUUUAU may cooperate with nearby miRNAs to affect transcript decay through a different mechanism, but more studies will be needed to clarify whether there is an effect of proximal UAUUUAU-interacting miRNA sites on transcript decay and its mechanistic basis.

In sum, our results estimated the prevalence of synergistic interactions between PUM and miRNAs. Some previous observations about miRNA targeting, including the efficiency of miRNA recognition sites in 3′UTRs, in AU-rich regions and at the beginning and end of the 3′UTR may be partially explained by synergistic interactions with RBPs [11, 149]. Currently, 829 human proteins are annotated as having RNA binding capacity by Gene Ontology [150] and we have only investigated the small fraction of them for which recognition site information is available. Other RBPs may also mediate the accessibility of miRNA recognition sites. A more comprehensive understanding of the interactions between miRNAs and RBPs could improve our ability to predict their targets and physiological functions, and provide insight into the mechanistic basis for their action.

## 3.4 Methods

### 3.4.1 Multiple genome alignment and 3′UTR annotation

Gene annotations for human, mouse, fly and worm were downloaded from the UCSC genome browser (http://genome.ucsc.edu/). Multiple genome alignments for the human genome (hg19) aligned with 32 placental mammals and mouse genome (mm9) with 29 vertebrates were also downloaded from the UCSC genome browser. 3′UTR regions were extracted for further analysis. Branch lengths for the associated phylogenetic tree were also downloaded.

The 3′UTR was defined as the region between the last stop codon and the 3′ end of the spliced mRNA. In some unusual cases, 3′UTRs are formed by the union of several distinct exons and cannot be mapped to a single continuous region on the genome assembly. For ease of analysis, we considered the last spliced exon, excluding any overlap with the protein-coding region, to be the 3′UTR. We also required that each 3′UTR was longer than 10 nts. In total, we analyzed 18,854 human 3′UTRs with average length 1,292.3 +/- 1,480.3 as standard deviation. Among these 3′UTRs, 17,766 were initiated at the stop codon and had no overlap with any coding region, and thus 94.2% of the 3′UTR annotations are complete.

### 3.4.2 RBP and miRNA recognition motif selection

All RBP binding motifs were first converted to consensus sequences. We expressed the consensus sequences as regular expressions, and used the regular expressions to search for recognition sites within the genomic sequence. We searched the 3′UTRs of the transcribed strands in multiple genome alignments for RBP recognition motifs in all aligned sequences. For each motif hit in the reference genome, we determined whether the same recognition motif was present within 10 nts in either direction in each of the other genomes. Based on the presence or absence of the motif within the

investigated genomes, we calculated the branch length score (BLS) by defining the minimum phylogenetic sub-tree that includes all conserved instances of the motif. The BLS is the branch length of this sub-tree as a fraction of the entire tree. Using this method, the BLS of individual motif hits can be inflated by a single hit to a distant species. To avoid this, we assigned no score to the most distant hit if there was a gap to that species that included more than one-third of the number of aligned genomes, and if the evolutionary distance from the most distant genome to the reference genome was more than twice as large as the distance to the second-most distant genome.

In order to assess the extent of conservation for each RBP motif, we generated 200 shuffled motifs by randomly swapping the nucleotides within the recognition motif. To remove redundant shuffled motifs, we profiled the similarity of each pair of motifs using Tomtom to generate $P$-values among the canonical and shuffled recognition motifs [151]. We ranked the $P$-values and determined the 10% threshold for the pairs with the most similar $P$-values. We eliminated shuffled motifs if they fell within this range and thus were considered too similar to any previously generated shuffled motif or the canonical motif. From the remaining motifs, we selected ten or the maximum possible number of shuffled motifs. When possible, we selected shuffled motifs that had a similar number of hits (+/- 20%) to the canonical motif from the reference genome. Through these criteria, we largely corrected for differences in the frequencies of di and tri-nucleotides [152]. For certain recognition motifs, nearly all shuffled motifs were associated with significantly fewer hits than the canonical motif. In this case, we selected 10 or the maximum possible number of arbitrary shuffled motifs as controls. For some RBP motifs with low complexity, for instance, sequences that were represented by a string of U's, we could not create three distinct shuffled motifs. These motifs were eliminated from further analysis.

We compared the conservation BLS scores for the canonical motif and each shuffled motif within the genome. We then determined the number of occurrences of hits to

the genome for the canonical motifs and the average among shuffled motifs for 100 different BLS thresholds from 0 to 1 with increments of 0.01. For each BLS threshold, we determined the precision as 1-(the average number of matches of shuffled motifs)/ (the number of matches of the canonical motif). We selected for further analysis RBPs for which the recognition motif contained more than 10 motif hits above a precision threshold of 0.6.

Mature human, mouse, fly and worm miRNA annotations were downloaded from Targetscan (http://targetscan.org). Two types of miRNA seeds were used: miRNA nucleotides 2-8 (m8) and nucleotides 2-7 with an adenosine opposite miRNA position 1 (1A) [11]. miRNA recognition seed motifs were defined as the complements of the miRNA seed and only conserved miRNA families were considered in further analyses.

### 3.4.3 RBP-miRNA motif site interaction

For each pair of miRNA and RBP, we defined the position of the RBP recognition motifs and identified the locations of the neighboring miRNA recognition sites in either direction. We generated histograms to depict the frequency with which the closest miRNA recognition motifs were present at ten different 50 nt windows 5' and 3' of the RBP motif. To generate a background model, we shuffled the identities of the miRNAs within each chromosome while keeping their positions intact. By shuffling the miRNA identities, we specifically tested the importance of co-localization with that particular miRNA. This approach eliminates any bias introduced by the fact that miRNA binding sites tend to be present together. Ten thousand shuffles were generated.

For each RBP, in each 50-nt-window, we compared the number of miRNA recognition sites for the real distribution versus the number derived from shuffled distributions. For each miRNA seed, the empirical $P$-value was calculated as the proportion of times that the number of miRNA sites was equal to or larger than the real

number of miRNA sites when 10,000 shuffles were performed. We then applied the Benjamini-Hochberg procedure on the $P$-values, and selected interacting miRNAs for each RBP with a FDR$\leq$0.05 [138]. Since for each miRNA there are two possible types of miRNA recognition seeds (1A and m8) [11], we required both of them to pass the FDR threshold of 0.05 to be included as an interacting miRNA.

Both RBPs and miRNAs tend to have recognition sites located at the beginning or end of 3′UTRs (Figure 3.4) [129, 130]. The miRNAs are more effective when localized in AU-rich regions [11, 104, 34]. Further, several of the RBP recognition motifs investigated have high AU content, with the most extreme instance being UAUUUAU. Thus, it is possible that the RBP-miRNA site colocalization we observed is a reflection of the similar positional preference of RBPs and miRNAs or their similarity with respect to the AU-richness of both types of motifs. In order to control for positional preference and AU content, we derived additional miRNA site identity shuffling procedures. To control for positional preference, all 3′UTRs were equally divided into 10 deciles and miRNA recognition sites were grouped according to the 3′UTR decile to which they belong. Then the identities of the miRNAs were shuffled among each 3′UTR decile group; in this way, miRNA sites located at the very end (or beginning) of 3′UTRs were swapped exclusively with other miRNA sites located at the very end (or beginning) of 3′UTRs.

To control for AU content, miRNA recognition seeds of 1A and m8 were classified into 3 groups based on the number of nucleotides that are an A or U out of the seven base pairs in the seed sequence. Category 1 contained miRNAs with a high (6 or 7) number of A/U nucleotides; category 2 contained miRNAs with a medium (3-5) number of A/U nucleotides; and category 3 contained miRNAs with a low (0-2) number of A/U nucleotides. The identities of miRNAs were shuffled with other miRNAs within the same category. In this way, AU-rich miRNA recognition sites

were swapped exclusively with other AU rich miRNA sites. Empirical $P$-values and Benjamini-Hochberg correction were performed as described above.

We only accepted miRNAs identified by the intersection of all three methods as interacting miRNAs for each RBP in each window. We found the window of 50 nts closest to the RBP site contained the largest number of interacting miRNAs, while windows that were more distant contained fewer or none (Table 3.1). To compile our final set of interacting miRNAs for each RBP, we only considered the miRNAs identified in the first 50 nt window. For each 50 nt window, an enrichment ratio was defined and visualized (Figure 3.8A) as the minimum ratio of (number of miRNA sites)/(expected number) among the three different shuffle methods.

Detailed statistics for all possible pairs of RBP and miRNA recognition motifs are available on the webpage http://cat.princeton.edu/miRNA_RBP/.

### 3.4.4 Test for the effect of AU content on RBP-miRNA interaction

When defining interacting miRNAs for each RBP, we explicitly accounted for the AU content of miRNAs by only shuffling across miRNAs with similar recognition seed AU content. In order to test the impact of the AU content of the RBP motifs, we generated window plots comparing the distribution of true RBP motifs to the distribution of shuffled RBP motifs (generated in the RBP recognition motif selection section). We created ten 50 nt windows upstream and downstream of each RBP motif and shuffled RBP motifs and counted the number of real miRNA recognition sites. For each window, the enrichment ratio was defined as ([Number of pairs for real RBP site and real miRNA site]/[Number of pairs for shuffled RBP site and real miRNA site]) normalized by an overall ratio of ([Number of pairs for real RBP site and real miRNA site across all windows]/[Number of pairs for shuffled RBP site and real miRNA site

across all windows]). For each RBP and its interacting miRNAs, enrichment ratios were visualized with heatmaps (Figure 3.11).

### 3.4.5 Cell-type specific expression profiles of miRNAs

For the cell lines included in our analysis, we identified companion, published small RNA sequencing experiments. For mRNA half-life measurements in human B cells (BL41) [139], miRNA sequencing reads were analyzed from the dataset generated by Landgraf and colleagues [141]. For mRNA half-life measurements in mouse fibroblasts (NIH-3T3) [139], miRNA reads were analyzed from the dataset generated by Zhu and colleagues [142]. For mRNA decay rate measurements in human HepG2 [140], miRNA reads were analyzed from the dataset generated by the ENCODE project [143]. For Par-CLIP datasets of PUM2 and AGO binding, small RNA sequencing data was analyzed from the dataset generated by Hafner and colleagues [118].

For each sequencing experiment, all conserved miRNAs annotated in TargetScan were ranked by the number of their mapped sequence reads. The most highly expressed 25% of the miRNAs were defined as expressed and the rest were defined as non-expressed.

### 3.4.6 Effects of RBPs on miRNA site accessibility

To test whether the binding of RBPs makes miRNA recognition sites more accessible, we analyzed sequences that contain RBP and miRNA recognition sites within 50 nts of each other, and included an extra 5 nts upstream of the 5' most site and 5 nts downstream of the 3' most site. We computationally folded these sequences using RNAfold [144] and determined the count C1 as the number of base pairs within the miRNA recognition seed site that are paired. Then, we converted all of the nucleotides within the RBP recognition site and one flanking nucleotide on each side to an 'N' to mask them from pairing [52, 144, 145], and reran RNAfold to determine the number

of base pairs within the miRNA site that were paired as count C2. For RNAfold predicted structure with folding energy larger than -1 kcal/mol, we considered this structure to be totally open and ignored any base pairing predicted. Finally, we calculated the rescue count C = C1 - C2.

In order to generate a background distribution of rescue counts, for each pair of neighboring miRNA and RBP sites localized within 50 nts, we randomized the sequence itself, but preserved the relative positions of the miRNA and RBP recognition sites. For this randomization, we preserved the mono and di-nucleotide frequency [146], as RNA folding energy is known to depend on di-nucleotide base stacking energies, and certain known RNA structures, such as tRNAs, have indistinguishable folding energy from di-nucleotide-preserving shuffles [147]. After randomizing the sequence for all miRNA-RBP neighboring sites, we repeated the rescue count calculation again. Ten randomizations were generated to estimate the average and standard deviation of the rescue counts in the background model.

To score the ability of miRNA recognition seed sequences to hybridize with the reverse RBP recognition motif, we used a simple scoring scheme in which A-U, G-C and G-U base pair matches were scored as 2, mismatches were scored as 0, and insertions and deletions were penalized with -1. The Smith-Waterman algorithm was then applied to find the best alignment [148].

# Chapter 4

# Conclusion

We developed a pipeline CCAT for predicting regulators that work together. We applied CCAT to find TF interactions based upon known TF binding motifs and DNaseI data, and to estimate the prevalence of synergistic interactions between RBPs and miRNAs based on their recognition motifs. The basic input to CCAT is just the genomic location of binding sites and a specific regional context (e.g., DNaseI chromatin open regions or gene 3′UTRs). Thus, CCAT is a general computational framework for finding combinatorial colocalization.

The CCAT predicted interaction pairs, both amongst TFs and between RBPs and miRNAs, recovered many known cases of cooperativity. Further, systematic quality assessments showed CCAT predicted pairs were coherent with other genomic datasets. Thus for specific biological condition of interest, without requiring hundreds of ChIP experiments (or RBP and miRNA binding profiling experiments), our pipeline enabled us to profile the genomic landscape of regulatory cooperativity.

In addition to uncovering regulatory cooperativity, the CCAT pipeline also provides tools for manipulating regulatory motifs, such as clustering similar motifs and searching for conserved motif instances using multiple genome alignments. All source code is released in a Unix software package. We also built an accompanying web-

site to assist in the use of the CCAT package and in further downstream analyses (`http://cat.princeton.edu`).

With the rapid development of DNaseI experiments and other related technologies, global regulatory landscapes are being actively profiled for an increasing number of biological processes. Knowledge of the binding specificities for eukaryotic TFs is also growing rapidly. Thus, the rapidly increasing availability of input datasets will enable CCAT to uncover cooperativity in many different biological processes, and thereby help obtain a more thorough understanding of the nature of combinatorial regulation.

# Bibliography


[1] Jiang P, Singh M, Coller HA: **Computational Assessment of the Cooperativity between RNA Binding Proteins and MicroRNAs in Transcript Decay**. *PLoS Comput Biol* 2013, **9**(5):e1003075.

[2] Spitz F, Furlong EE: **Transcription factors: from enhancer binding to developmental control**. *Nat Rev Genet* 2012, **13**(9):613–626.

[3] Mata J, Marguerat S, Bhler J: **Post-transcriptional control of gene expression: a genome-wide perspective**. *Trends Biochem Sci* 2005, **30**(9):506–14.

[4] Arnosti DN, Kulkarni MM: **Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?** *J Cell Biochem* 2005, **94**(5):890–898.

[5] Lagha M, Bothma JP, Levine M: **Mechanisms of transcriptional precision in animal development**. *Trends Genet* 2012, **28**(8):409–416.

[6] Juven-Gershon T, Kadonaga JT: **Regulation of gene expression via the core promoter and the basal transcriptional machinery**. *Dev Biol* 2010, **339**(2):225–229.

[7] Houseley J, Tollervey D: **The many pathways of RNA degradation**. *Cell* 2009, **136**(4):763–76.

[8] Kong J, Lasko P: **Translational control in cellular and developmental processes**. *Nat Rev Genet* 2012, **13**(6):383–94.

[9] Lai EC: **Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation**. *Nat Genet* 2002, **30**(4):363–4.

[10] Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function**. *Cell* 2004, **116**(2):281–97.

[11] Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP: **MicroRNA targeting specificity in mammals: determinants beyond seed pairing**. *Mol Cell* 2007, **27**:91–105.

[12] Bartel DP: **MicroRNAs: target recognition and regulatory functions**. *Cell* 2009, **136**(2):215–33.

[13] Barreau C, Paillard L, Osborne HB: **AU-rich elements and associated factors: are there unifying principles?** *Nucleic Acids Res* 2005, **33**(22):7138–50.

[14] von Roretz C, Gallouzi IE: **Decoding ARE-mediated decay: is microRNA part of the equation?** *J Cell Biol* 2008, **181**(2):189–94.

[15] Brewer G: **An A + U-rich element RNA-binding factor regulates c-myc mRNA stability in vitro**. *Mol Cell Biol* 1991, **11**(5):2460–6.

[16] Hudson BP, Martinez-Yamout MA, Dyson HJ, Wright PE: **Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d**. *Nat Struct Mol Biol* 2004, **11**(3):257–64.

[17] Chen CY, Gherzi R, Ong SE, Chan EL, Raijmakers R, Pruijn GJ, Stoecklin G, Moroni C, Mann M, Karin M: **AU binding proteins recruit the exosome to degrade ARE-containing mRNAs**. *Cell* 2001, **107**(4):451–64.

[18] Wickens M, Bernstein DS, Kimble J, Parker R: **A PUF family portrait: 3'UTR regulation as a way of life**. *Trends Genet* 2002, **18**(3):150–7.

[19] Olivas W, Parker R: **The Puf3 protein is a transcript-specific regulator of mRNA degradation in yeast**. *EMBO J* 2000, **19**(23):6602–11.

[20] Goldstrohm AC, Hook BA, Seay DJ, Wickens M: **PUF proteins bind Pop2p to regulate messenger RNAs**. *Nat Struct Mol Biol* 2006, **13**(6):533–9.

[21] Brennan CM, Steitz JA: **HuR and mRNA stability**. *Cell Mol Life Sci* 2001, **58**(2):266–77.

[22] Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo HCL, Chu HC, Ogawa N, Inwood W, Sementchenko V, Beaton A, Weiszmann R, Celniker SE, Knowles DW, Gingeras T, Speed TP, Eisen MB, Biggin MD: **Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm**. *PLoS Biol* 2008, **6**(2):e27.

[23] MacArthur S, Li XY, Li J, Brown JB, Chu HC, Zeng L, Grondona BP, Hechmer A, Simirenko L, Kernen SV, Knowles DW, Stapleton M, Bickel P, Biggin MD, Eisen MB: **Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions**. *Genome Biol* 2009, **10**(7):R80.

[24] Gerstein MB, Kundaje A, Hariharan M, Landt SG, et al.: **Architecture of the human regulatory network derived from ENCODE data**. *Nature* 2012, **489**(7414):91–100.

[25] Rada-Iglesias A, Bajpai R, Prescott S, Brugmann SA, Swigut T, Wysocka J: **Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest**. *Cell Stem Cell* 2012, **11**(5):633–648.

[26] Xue Y, Ouyang K, Huang J, Zhou Y, Ouyang H, Li H, Wang G, Wu Q, Wei C, Bi Y, Jiang L, Cai Z, Sun H, Zhang K, Zhang Y, Chen J, Fu XD: **Direct Conversion of Fibroblasts to Neurons by Reprogramming PTB-Regulated MicroRNA Circuits**. *Cell* 2013, **152**(1-2):82–96.

[27] Yez-Cuna JO, Kvon EZ, Stark A: **Deciphering the transcriptional cis-regulatory code**. *Trends Genet* 2012, **S0168-9525**(12):00147–3.

[28] Kedde M, van Kouwenhove M, Zwart W, Oude Vrielink JA, Elkon R, Agami R: **A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility**. *Nat Cell Biol* 2010, **12**(10):1014–20.

[29] Erives A, Levine M: **Coordinate enhancers share common organizational features in the Drosophila genome**. *Proc Natl Acad Sci U S A* 2004, **101**(11):3851–3856.

[30] Papatsenko D, Levine M: **A rationale for the enhanceosome and other evolutionarily constrained enhancers**. *Curr Biol* 2007, **17**(22):R955–7.

[31] Zinzen RP, Senger K, Levine M, Papatsenko D: **Computational models for neurogenic gene expression in the Drosophila embryo**. *Curr Biol* 2006, **16**(13):1358–1365.

[32] Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD: **The evolution of combinatorial gene regulation in fungi**. *PLoS Biol* 2008, **6**(2):e38.

[33] Senger K, Armstrong GW, Rowell WJ, Kwan JM, Markstein M, Levine M: **Immunity regulatory DNAs share common organizational features in Drosophila**. *Mol Cell* 2004, **13**:19–32.

[34] Sun G, Li H, Rossi JJ: **Sequence context outside the target region influences the effectiveness of miR-223 target sites in the RhoB 3'UTR**. *Nucleic Acids Res* 2010, **38**:239–52.

[35] Kim HH, Kuwano Y, Srikantan S, Lee EK, Martindale JL, Gorospe M: **HuR recruits let-7/RISC to repress c-Myc expression**. *Genes Dev* 2009, **23**(15):1743–8.

[36] Nolde MJ, Saka N, Reinert KL, Slack FJ: **The Caenorhabditis elegans pumilio homolog, puf-9, is required for the 3'UTR-mediated repression of the let-7 microRNA target gene, hbl-1**. *Dev Biol* 2007, **305**(2):551–63.

[37] Negre N, Brown CD, Ma L, Bristow CA, et al.: **A cis-regulatory map of the Drosophila genome**. *Nature* 2011, **471**(7339):527–531.

[38] Dunham I, Kundaje A, Aldred SF, Collins PJ, et al.: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**(7414):57–74.

[39] Yez-Cuna JO, Dinh HQ, Kvon EZ, Shlyueva D, Stark A: **Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding**. *Genome Res* 2012, **22**(10):2018–2030.

[40] Mullen AC, Orlando DA, Newman JJ, Lovn J: **Master transcription factors determine cell-type-specific responses to TGF-beta signaling**. *Cell* 2011, **147**(3):565–756.

[41] He Q, Bardet AF, Patton B, Purvis J: **High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species**. *Nat Genet* 2011, **43**(5):414–420.

[42] Yu X, Lin J, Masuda T, Esumi N, Zack DJ, Qian J: **Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae**. *Nucleic Acids Res* 2006, **34**(3):917–927.

[43] Yu X, Lin J, Zack DJ, Qian J: **Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues**. *Nucleic Acids Res* 2006, **34**(17):4925–4936.

[44] Kranz AL, Eils R, Knig R: **Enhancers regulate progression of development in mammalian cells**. *Nucleic Acids Res* 2011, **39**(20):8689–8702.

[45] Adryan B, Teichmann SA: **FlyTF: a systematic review of site-specific transcription factors in the fruit fly Drosophila melanogaster**. *Bioinformatics* 2006, **22**(12):1532–1533.

[46] Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM: **A census of human transcription factors: function, expression and evolution**. *Nat Rev Genet* 2009, **10**(4):252–263.

[47] Bulger M, Groudine M: **Functional and mechanistic diversity of distal transcription enhancers**. *Cell* 2011, **144**(3):327–339.

[48] Calhoun VC, Levine M: **Long-range enhancer-promoter interactions in the Scr-Antp interval of the Drosophila Antennapedia complex**. *Proc Natl Acad Sci U S A* 2003, **100**(17):9878–9883.

[49] Amano T, Sagai T, Tanabe H, Mizushina Y, Nakazawa H, Shiroishi T: **Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription**. *Dev Cell* 2009, **16**:47–57.

[50] Neph S, Vierstra J, Stergachis AB, Reynolds AP, Stamatoyannopoulos JA, et al.: **An expansive human regulatory lexicon encoded in transcription factor footprints**. *Nature* 2012, **489**(7414):83–90.

[51] Galgano A, Forrer M, Jaskiewicz L, Kanitz A, Zavolan M, Gerber AP: **Comparative analysis of mRNA targets for human PUF-family proteins suggests extensive interaction with the miRNA regulatory system**. *PLoS One* 2008, **3**(9):e3164.

[52] Leibovich L, Mandel-Gutfreund Y, Yakhini Z: **A structural-based statistical approach suggests a cooperative activity of PUM1 and miR-410 in human 3'-untranslated regions**. *Silence* 2010, **1**:17.

[53] Johnson AD: **Molecular mechanisms of cell-type determination in budding yeast**. *Curr Opin Genet Dev* 1995, **5**(5):552–558.

[54] Miller JA, Widom J: **Collaborative competition mechanism for gene activation in vivo**. *Mol Cell Biol* 2003, **23**(5):1623–1632.

[55] Falvo JV, Thanos D, Maniatis T: **Reversal of intrinsic DNA bends in the IFN beta gene enhancer by transcription factors and the architectural protein HMG I(Y)**. *Cell* 1995, **83**(7):1101–1111.

[56] Panne D, Maniatis T, Harrison SC: **An atomic model of the interferon-beta enhanceosome**. *Cell* 2007, **129**(6):1111–1123.

[57] Slattery M, Riley T, Liu P, Abe N: **Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins**. *Cell* 2011, **147**(6):1270–1282.

[58] Joshi R, Passner JM, Rohs R, Jain R: **Functional specificity of a Hox protein mediated by the recognition of minor groove structure**. *Cell* 2007, **131**(3):530–543.

[59] Garvie CW, Hagman J, Wolberger C: **Structural studies of Ets-1/Pax5 complex formation on DNA**. *Mol Cell* 2001, **8**(6):1267–1276.

[60] Siggers T, Duyzend MH, Reddy J, Khan S, Bulyk ML: **Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex**. *Mol Syst Biol* 2011, **7**:555.

[61] Kheradpour P, Stark A, Roy S, Kellis M: **Reliable prediction of regulator targets using 12 Drosophila genomes**. *Genome Res* 2007, **17**(12):1919–1931.

[62] Marbach D, Roy S, Ay F, Meyer PE, Candeias R, Kahveci T, Bristow CA, Kellis M: **Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks**. *Genome Res* 2012, **22**(7):1334–1349.

[63] Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U: **Predicting expression patterns from regulatory sequence in Drosophila segmentation**. *Nature* 2008, **451**(7178):535–540.

[64] He X, Samee MA, Blatti C, Sinha S: **Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression**. *PLoS Comput Biol* 2010, **6**(9).

[65] Hesselberth JR, Chen X, Zhang Z, Sabo PJ: **Global mapping of protein-DNA interactions in vivo by digital genomic footprinting**. *Nat Methods* 2009, **6**(4):283–289.

[66] Li XY, Thomas S, Sabo PJ, Eisen MB, Stamatoyannopoulos JA, Biggin MD: **The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding**. *Genome Biol* 2011, **12**(4):R34.

[67] Thomas S, Li XY, Sabo PJ, Sandstrom R: **Dynamic reprogramming of chromatin accessibility during Drosophila embryo development**. *Genome Biol* 2011, **12**(5):R43.

[68] Thurman RE, Rynes E, Humbert R, Vierstra J: **The accessible chromatin landscape of the human genome**. *Nature* 2012, **489**(7414):75–82.

[69] Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA: **Circuitry and dynamics of human transcription factor regulatory networks**. *Cell* 2012, **150**(6):1274–1286.

[70] He HH, Meyer CA, Chen MW, Jordan VC, Brown M, Liu XS: **Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics**. *Genome Res* 2012, **22**(6):1015–1025.

[71] **FlyFactorSurvey** `[http://pgfe.umassmed.edu/TFDBS/]`.

[72] Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA: **Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites**. *Cell* 2008, **133**(7):1277–1289.

[73] Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, Wolfe SA: **A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system**. *Nucleic Acids Res* 2008, **36**(8):2547–2560.

[74] Bergman CM, Carlson JW, Celniker SE: **Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaste**. *Bioinformatics* 2005, **21**(8):1747–1749.

[75] Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A: **JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update**. *Nucleic Acids Res* 2008, **36**:D102–106.

[76] Matys V, Fricke E, Geffers R, Gssling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Mnch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles**. *Nucleic Acids Res* 2003, **31**:374–378.

[77] Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE: **Combinatorial binding predicts spatio-temporal cis-regulatory activity**. *Nature* 2009, **462**(7269):65–70.

[78] Schwartz YB, Linder-Basso D, Kharchenko PV, Tolstorukov MY: **Nature and function of insulator protein binding sites in the Drosophila genome**. *Genome Res* 2012, **22**(11):2188–2198.

[79] Ni X, Zhang YE, Ngre N, Chen S, Long M, White KP: **Adaptive Evolution and the Birth of CTCF Binding Sites in the Drosophila Genome**. *PLoS Biol* 2012, **10**(11):e1001420.

[80] Grant CE, Bailey TL, Noble WS: **FIMO: Scanning for occurrences of a given motif**. *Bioinformatics* 2011, **Epub ahead of print**.

[81] **The MEME Suite: Motif-based sequence analysis tools** `[http://meme.nbcr.net]`.

[82] Stark A, Lin MF, Kheradpour P, Pedersen JS: **Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures**. *Nature* 2007, **450**(7167):219–232.

[83] Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS: **Quantifying similarity between motifs**. *Genome Biol* 2007, **8**(2):R24.

[84] Roy S, Ernst J, Kharchenko PV, Kheradpour P, et al.: **Identification of functional elements and regulatory circuits by Drosophila modENCODE**. *Science* 2010, **330**(6012):1787–1797.

[85] Ngre N, Brown CD, Shah PK, Kheradpour P, et al.: **A comprehensive map of insulator elements for the Drosophila genome**. *PLoS Genet* 2010, **6**:e1000814.

[86] Bushey AM, Ramos E, Corces VG: **Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions**. *Genes Dev* 2009, **23**(11):1338–1350.

[87] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**:25–29.

[88] **Regulatory Element Database for Drosophila** [http://redfly.ccr.buffalo.edu/].

[89] Milo R, Kashtan N, Itzkovitz S, Newman MEJ, Alon U: **On the uniform generation of random graphs with prescribed degree sequences**. *arXiv* 2004.

[90] Harrison MM, Li XY, Kaplan T, Botchan MR, Eisen MB: **Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition**. *PLoS Genet* 2011, **7**(10):e1002266.

[91] Driever W, Nsslein-Volhard C: **The bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner**. *Cell* 1988, **54**:95–104.

[92] Zinzen RP, Senger K, Levine M, Papatsenko D: **Computational models for neurogenic gene expression in the Drosophila embryo**. *Curr Biol* 2006, **16**(13):1358–1365.

[93] Mason PBJ, Lis JT: **Cooperative and competitive protein interactions at the hsp70 promoter**. *J Biol Chem* 1997, **272**(52):33227–33233.

[94] Mohan M, Bartkuhn M, Herold M, Philippen A, et al.: **The Drosophila insulator proteins CTCF and CP190 link enhancer blocking to body patterning**. *EMBO J* 2007, **26**(19):4203–4214.

[95] Gke J, Jung M, Behrens S, Chavez L: **Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development**. *PLoS Comput Biol* 2011, **7**(12):e1002304.

[96] Papatsenko D, Goltsev Y, Levine M: **Organization of developmental enhancers in the Drosophila embryo**. *Nucleic Acids Res* 2009, **37**(17):5665–5677.

[97] Junion G, Spivakov M, Girardot C, Braun M, Gustafson EH, Birney E, Furlong EE: **A transcription factor collective defines cardiac cell fate and reflects lineage history**. *Cell* 2012, **148**(3):473–486.

[98] Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK: **Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data**. *Genome Res* 2011, **21**(3):447–455.

[99] **UCSC Genome Bioinformatics Site** [http://genome.ucsc.edu].

[100] **A Database of Drosophila Genes & Genomes** [http://flybase.org/].

[101] Gordn R, Murphy KF, McCord RP, Zhu C, Vedenko A, Bulyk ML: **Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights**. *Genome Biol* 2011, **12**(12):R125.

[102] Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society* 1995, **57**:289–300.

[103] Gregory RI, Chendrimada TP, Cooch N, Shiekhattar R: **Human RISC couples microRNA biogenesis and posttranscriptional gene silencing**. *Cell* 2005, **123**(4):631–40.

[104] Didiano D, Hobert O: **Molecular architecture of a miRNA-regulated 3' UTR**. *RNA* 2008, **14**(7):1297–317.

[105] Jiang P, Coller HA: **Functional Interactions Between microRNAs and RNA Binding Proteins**. *MicroRNA* 2012, **1**:70–79.

[106] Mukherjee N, Corcoran DL, Nusbaum JD, Reid DW, Georgiev S, Hafner M, Ascano J M, Tuschl T, Ohler U, Keene JD: **Integrative Regulatory Mapping Indicates that the RNA-Binding Protein HuR Couples Pre-mRNA Processing and mRNA Stability**. *Mol Cell* 2011, **43**(3):327–39.

[107] Jacobsen A, Wen J, Marks DS, Krogh A: **Signatures of RNA binding proteins globally coupled to effective microRNA target sites**. *Genome Res* 2010, **20**(8):1010–9.

[108] Bhattacharyya SN, Habermacher R, Martine U, Closs EI, Filipowicz W: **Relief of microRNA-mediated translational repression in human cells subjected to stress**. *Cell* 2006, **125**(6):1111–24.

[109] Kedde M, Strasser MJ, Boldajipour B, Oude Vrielink JA, Slanchev K, le Sage C, Nagel R, Voorhoeve PM, van Duijse J, Orom UA, Lund AH, Perrakis A, Raz E, Agami R: **RNA-binding protein Dnd1 inhibits microRNA access to target mRNA**. *Cell* 2007, **131**(7):1273–86.

[110] Elcheva I, Goswami S, Noubissi FK, Spiegelman VS: **CRD-BP protects the coding region of betaTrCP1 mRNA from miR-183-mediated degradation**. *Mol Cell* 2009, **35**(2):240–6.

[111] Caudy AA, Myers M, Hannon GJ, Hammond SM: **Fragile X-related protein and VIG associate with the RNA interference machinery**. *Genes Dev* 2002, **16**(19):2491–6.

[112] Miles WO, Tschop K, Herr A, Ji JY, Dyson NJ: **Pumilio facilitates miRNA regulation of the E2F3 oncogene**. *Genes Dev* 2012, **26**(4):356–68.

[113] Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO: **Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system**. *PLoS Biol* 2008, **6**(10):e255.

[114] Auweter SD, Oberstrass FC, Allain FH: **Sequence-specific binding of single-stranded RNA: is there a code for recognition?** *Nucleic Acids Res* 2006, **34**(17):4943–59.

[115] Ray D, Kazan H, Chan ET, Pena Castillo L, Chaudhry S, Talukder S, Blencowe BJ, Morris Q, Hughes TR: **Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins**. *Nat Biotechnol* 2009, **27**(7):667–70.

[116] Wang X, McLachlan J, Zamore PD, Hall TM: **Modular recognition of RNA by a human pumilio-homology domain**. *Cell* 2002, **110**(4):501–12.

[117] Zhang CL, Darnell RB: **Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data**. *Nat Biotechnol* 2011, **29**(7):607–U86.

[118] Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano J M, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T: **Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP**. *Cell* 2010, **141**:129–41.

[119] Kheradpour P, Stark A, Roy S, Kellis M: **Reliable prediction of regulator targets using 12 Drosophila genomes**. *Genome Res* 2007, **17**(12):1919–31.

[120] Auweter SD, Fasan R, Reymond L, Underwood JG, Black DL, Pitsch S, Allain FH: **Molecular basis of RNA recognition by the human alternative splicing factor Fox-1**. *EMBO J* 2006, **25**:163–73.

[121] Ponthier JL, Schluepen C, Chen W, Lersch RA, Gee SL, Hou VC, Lo AJ, Short SA, Chasis JA, Winkelmann JC, Conboy JG: **Fox-2 splicing factor binds to a conserved intron motif to promote inclusion of protein 4.1R alternative exon 16**. *J Biol Chem* 2006, **281**(18):12468–74.

[122] Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470–6.

[123] Boelens WC, Jansen EJ, van Venrooij WJ, Stripecke R, Mattaj IW, Gunderson SI: **The human U1 snRNP-specific U1A protein inhibits polyadenylation of its own pre-mRNA**. *Cell* 1993, **72**(6):881–92.

[124] Oubridge C, Ito N, Evans PR, Teo CH, Nagai K: **Crystal structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin**. *Nature* 1994, **372**(6505):432–8.

[125] Brewer BY, Malicka J, Blackshear PJ, Wilson GM: **RNA sequence elements required for high affinity binding by the zinc finger domain of tristetraprolin: conformational changes coupled to the bipartite nature of Au-rich MRNA-destabilizing motifs**. *J Biol Chem* 2004, **279**(27):27870–7.

[126] Yeo GW, Coufal NG, Liang TY, Peng GE, Fu XD, Gage FH: **An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells**. *Nat Struct Mol Biol* 2009, **16**(2):130–7.

[127] Kuroyanagi H: **Fox-1 family of RNA-binding proteins**. *Cell Mol Life Sci* 2009, **66**(24):3895–907.

[128] Price SR, Evans PR, Nagai K: **Crystal structure of the spliceosomal U2B"-U2A' protein complex bound to a fragment of U2 small nuclear RNA**. *Nature* 1998, **394**(6694):645–50.

[129] Majoros WH, Ohler U: **Spatial preferences of microRNA targets in 3' untranslated regions**. *BMC Genomics* 2007, **8**:152.

[130] Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M: **Inference of miRNA targets using evolutionary conservation and pathway analysis**. *BMC Bioinformatics* 2007, **8**:69.

[131] Spassov DS, Jurecic R: **The PUF family of RNA-binding proteins: does evolutionarily conserved structure equal conserved function?** *IUBMB Life* 2003, **55**(7):359–66.

[132] White EK, Moore-Jarrett T, Ruley HE: **PUM2, a novel murine puf protein, and its consensus RNA-binding site**. *RNA* 2001, **7**(12):1855–66.

[133] Gerber AP, Luschnig S, Krasnow MA, Brown PO, Herschlag D: **Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in Drosophila melanogaster**. *Proc Natl Acad Sci U S A* 2006, **103**(12):4487–92.

[134] Bernstein D, Hook B, Hajarnavis A, Opperman L, Wickens M: **Binding specificity and mRNA targets of a C. elegans PUF protein, FBF-1**. *RNA* 2005, **11**(4):447–58.

[135] Lai WS, Parker JS, Grissom SF, Stumpo DJ, Blackshear PJ: **Novel mRNA targets for tristetraprolin (TTP) identified by global analysis of stabilized transcripts in TTP-deficient fibroblasts**. *Mol Cell Biol* 2006, **26**(24):9196–208.

[136] DeRenzis S, Elemento O, Tavazoie S, Wieschaus EF: **Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo**. *PLoS Biol* 2007, **5**(5):e117.

[137] Wu J, Duggan A, Chalfie M: **Inhibition of touch cell fate by egl-44 and egl-46 in C. elegans**. *Genes Dev* 2001, **15**(6):789–802.

[138] Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society Series B-Methodological* 1995, **57**:289–300.

[139] Friedel CC, Dolken L, Ruzsics Z, Koszinowski UH, Zimmer R: **Conserved principles of mammalian transcriptional regulation revealed by RNA half-life**. *Nucleic Acids Res* 2009, **37**(17):e115.

[140] Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, Darnell J J E: **Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes**. *Genome Res* 2003, **13**(8):1863–72.

[141] Landgraf P, Rusu M, Sheridan R, Sewer A, et al.: **A mammalian microRNA expression atlas based on small RNA library sequencing**. *Cell* 2007, **129**(7):1401–14.

[142] Zhu JY, Strehle M, Frohn A, Kremmer E, Hofig KP, Meister G, Adler H: **Identification and analysis of expression of novel microRNAs of murine gammaherpesvirus 68**. *Journal of Virology* 2010, **84**(19):10266–75.

[143] Project AET: **Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs**. *Nature* 2009, **457**(7232):1028–32.

[144] Hofacker IL, Fontana W, Stadle PF, Bonhoeffer LS, Tacker M, Schuster P: **Fast Folding and Comparison of RNA Secondary Structures**. *Monatshefte fr Chemie* 1994, **125**(2):167–188.

[145] Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E: **The role of site accessibility in microRNA target recognition**. *Nat Genet* 2007, **39**(10):1278–84.

[146] Altschul SF, Erickson BW: **Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage**. *Mol Biol Evol* 1985, **2**(6):526–38.

[147] Krogh A, Workman C: **No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution**. *Nucleic Acids Res* 1999, **27**(24):4816–4822.

[148] Smith TF, Waterman MS: **Identification of common molecular subsequences**. *J Mol Biol* 1981, **147**:195–7.

[149] Forman JJ, Coller HA: **The code within the code: microRNAs target coding regions**. *Cell Cycle* 2010, **9**(8):1533–41.

[150] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**:25–9.

[151] Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS: **Quantifying similarity between motifs**. *Genome Biol* 2007, **8**(2):R24.

[152] Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB: **Prediction of mammalian microRNA targets**. *Cell* 2003, **115**(7):787–98.