2014-07-17

Evolutionary Approaches to the Study of Small Noncoding Regulatory RNA Pathways: A Dissertation

Alfred T. Simkin
University of Massachusetts Medical School

Repository Citation

Simkin AT. (2014). Evolutionary Approaches to the Study of Small Noncoding Regulatory RNA Pathways: A Dissertation. GSBS Dissertations and Theses. https://doi.org/10.13028/M2PK6G. Retrieved from https://escholarship.umassmed.edu/gsbs_diss/721

EVOLUTIONARY APPROACHES TO THE STUDY OF SMALL NONCODING REGULATORY RNA PATHWAYS

A Dissertation Presented

By

Alfred T. Simkin

Submitted to the Faculty of the University of Massachusetts Graduate School of Biomedical Sciences, Worcester in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

(July 17, 2014)

(Molecular Evolution)

EVOLUTIONARY APPROACHES TO THE STUDY OF SMALL NONCODING REGULATORY RNA PATHWAYS

A Dissertation Presented
By

Alfred T. Simkin

The signatures of the Dissertation Committee signify
completion and approval as to style and content of the Dissertation

Jeffrey D. Jensen, Fen-Biao Gao, Thesis Advisors

Jeffrey A. Bailey, Member of Committee

William Theurkauf, Member of Committee

Zhiping Weng, Member of Committee

Daniel Weinreich, Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation
meets the requirements of the Dissertation Committee

Victor Ambros, Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences
signifies that the student has met all graduation requirements of the school.

Anthony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical Sciences

Interdisciplinary Graduate Program

July 17, 2014

This work is dedicated to my wife Arshia and to my family, who have encouraged and enabled me to study life, and who constantly remind me of the joy of living

ACKNOWLEDGMENTS

I would like to thank Bill Theurkauf and Zhiping Weng for funding my research and giving me a home during the last year of my PhD, for their patience with my insistence on doing some things the hard way and with my other commitments, and for a large number of lively and useful discussions of piRNA biology and Bioinformatics techniques. I would also like to thank Fen-Biao Gao for a great suggestion of a thesis project, for funding, and for his willingness to explore unfamiliar territory with my miRNA evolution work. Finally, I would like to thank Jeff Jensen for supporting my work financially, for going out of his way to help me achieve my professional aspirations, for opening the door into the field of evolutionary biology, and for being intellectually available and unceasingly interested in my progress, at virtually all hours of the day, year after year.

In addition, nothing in Science is ever done alone, and in many ways the directions I've taken in this thesis represent a consensus of ideas and suggestions from a large number of people as much as they do my own decisions. These people include my thesis research advisory committee, members of the Umass and EPFL communities, attendees of presentations given at conferences and seminars, and intelligent laypeople among my friends and family whose insights, unhampered by scientific dogmas, often lead to entirely new ways of looking at familiar problems.

ABSTRACT

Short noncoding RNAs play roles in regulating nearly every biological process, in nearly every organism, yet the exact function and importance of these molecules remains a subject of some debate. In order to gain a better understanding of the contexts in which these regulators have evolved, I have undertaken a variety of approaches to study the evolutionary history of the components that make up these pathways, in the form of two main research efforts. In the first chapter, I have used a combination of population genetics and molecular evolution techniques to show that proteins involved in the piRNA pathway are rapidly evolving, and that different components of the pathway seem to be evolving rapidly on different timescales. These rapidly evolving piRNA pathway proteins can be loosely separated into two groups. The first group appears to evolve quickly at the species level, perhaps in response to transposons that invade across species lines, while the second group appears to evolve quickly at the level of individual populations, perhaps in response to transposons that are paternally present yet novel to the maternal genome. In the second chapter of my research, I have used molecular evolution techniques and carefully devised controls to show that the binding sites of well-conserved miRNAs are among the most slowly changing short motifs in the genome, consistent with a conserved function for these short RNAs in regulatory pathways that are ancient and extremely slow to change. I have additionally discovered a major flaw in an existing approach to motif turnover calculations, which may lead to systematic biases in the published literature toward the false inference of increased regulatory complexity over time. I have implemented a revised approach to motif turnover that addresses this flaw.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

The chapters of this dissertation are taken mostly verbatim (but with some additional material not included with the published versions) from two previously published works. These works are:

Simkin, A., Wong, A., Poh, Y.-P., Theurkauf, W. E., & Jensen, J. D. (2013). Recurrent and recent selective sweeps in the piRNA pathway. Evolution; International Journal of Organic Evolution, 67(4), 1081–90.

Simkin, A. T., Bailey, J. A, Gao, F.-B., & Jensen, J. D. (2014). Inferring the Evolutionary History of Primate microRNA Binding Sites: Overcoming Motif Counting Biases. Molecular Biology and Evolution. 31(7), 1894-1901

CHAPTER I

Introduction

Noncoding RNAs have traits that make them ideally suited to regulating many aspects of biology. The RNA world hypothesis postulates that all life processes were originally carried out by RNA molecules. In support of this model, several RNAs, called ribozymes, have catalytic functions, allowing them to perform enzymatic reactions and fulfill several of the roles currently performed by proteins. One of the most ancient and conserved enzymatic reactions, protein synthesis, is mediated almost entirely by RNAs, with tRNAs bringing amino acids into the largely RNA-based ribosome for translation of mRNAs into protein (cech 2000). RNAs form the genomes of several viruses, and so are also capable of filling the role of DNA as the genetic blueprint for life.

Because RNAs are so flexible, occupying a role at the interface between the encoded genetic material of the DNA world and the enzymatic reactions catalyzed mainly by proteins, John Mattick and Michael Gagen have argued that the greatest potential of RNA may be its ability to serve as a regulator of the speed and specificity of a wide range of biological processes (Mattick and Gagen 2001). These researchers (and several others) have pointed out that while DNA provides a more stable repository of genetic information, and protein is a more versatile source of enzymatic complexity, RNA is uniquely situated to bind specifically and perfectly to any stretch of DNA or RNA, and translate this sequence specificity into conformational changes in RNA structure whose shapes can be recognized by proteins.

Several short RNA pathways have (relatively) recently been discovered to play regulatory roles in a diverse set of organisms and biological contexts. Among these, the siRNA and CRISPR pathways appear to target ectopic material in Eukaryotes and Prokaryotes, respectively, while the piRNA and miRNA pathways often appear to silence endogenous genes in the form of transposable elements and endogenous host genes, respectively, at the post-transcriptional level in Eukaryotes

(Aalto and Pasquinelli 2012, Sampson and Weiss 2014). Despite these broad classifications, examples are also emerging of short RNAs that act to upregulate genes instead of silencing them (Vasudevan et al. 2007) and target genes at the DNA instead of RNA level (Morris 2009). In addition, some miRNAs appear to target viruses (Lecellier et al. 2005) and some siRNAs appear to target endogenous genes (Ghildiyal et al. 2008). An entire industry is emerging around the recent discovery that the CRISPR system of prokaryotic gene silencing can be repurposed to ablate or insert DNA in a targeted way within Eukaryotes (Sampson and Weiss 2014). Among the implications of this discovery are new approaches to molecular biology and gene therapy. In plants, the miRNA pathway appears to be capable of cleaving invader elements, much the same as the siRNA pathway in animals (Rogers and Chen 2013, Ramesh et al. 2013). In C. elegans, new research is suggesting that a special branch of the siRNA and piRNA pathways (known as epigenetic RNA silencing and RNA activation) is used to distinguish self from non-self within the genome, so that newly inserted copies of ectopically derived genes are stably silenced across multiple generations, while anciently inserted material remains resistant to silencing (Seth et al. 2013).

As exciting as these findings are, there may be a vast sea of new roles for RNA which we have yet to glimpse. The Encyclopedia of DNA Elements (ENCODE) project set out to determine levels of transcription across the genome. It now appears that over 50% of the human genome is transcribed (ENCODE Project Consortium 2012), whereas (as of June 5, 2014) only approximately 1% is annotated as coding for amino acids by the National Center for Bioinformatics' 'Refseq' annotation. After a survey of a large number of organisms, it now appears that amount of protein coding DNA correlates very poorly with organismal complexity (measured as total number of distinct organismal cell types), while proportion of DNA that is noncoding correlates almost perfectly with a logarithmic function of complexity before reaching a plateau, leading to a theory that increasing amounts of noncoding RNA are necessary to encode increasingly complex regulatory circuitry, and that there is some peak level of regulatory complexity that falls at the threshold of a disordered, chaotic system (Liu et al. 2013). If this is true, we might expect the vast majority of noncoding material within genomes to be functional, and there are several studies ascribing essential functions to some recently discovered classes of long intergenic noncoding RNA (reviewed in Bai et al. 2014).

In this environment, these are exciting times to be researching noncoding RNAs, particularly within the field of bioinformatics. Most of the well-known small RNA regulatory molecules appear to be approximately 20 to 30 nucleotides long, and to exert regulatory impacts on target genes via direct basepairing. This sequence length is short enough to permit a large number of regulatory interactions within a single messenger RNA, yet long enough in theory to permit complete specificity of gene targeting within even the largest genomes. As genomes are increasingly becoming readily available, the basepairing property of regulatory RNAs makes them attractive candidates for computational study, as potential targets of any given regulatory RNA can be easily and directly searched within any known genome. Because of this trait, short regulatory RNAs that act by basepairing are currently much better understood than other RNA pathways, and have become the subject of my research.

In the scientific community, there is a fundamental tension between the need to classify biological processes and an interest in accurately describing the full diversity of innovative and unique solutions exploited by different life forms. This appears to be especially true for the field of noncoding RNAs. As exciting as it is to be conducting research in this time period, many of the discoveries being made currently raise as many if not more questions than answers, undermining old restrictions of function and opening the door of putative RNA functions ever wider. I give two examples of recent findings in the piRNA and miRNA pathways below for an illustration of the baffling world we currently occupy.

The piRNA pathway

The piRNA pathway is perhaps best studied in *Drosophila melanogaster*, in which piRNAs approximately 20-30 bp in length bind with Argonaute protein complexes and subsequently silence transposable elements by basepairing with complementary transposon mRNA, at which point some factor in the Argonaute complex is thought to cleave or otherwise destroy the target transcripts (Brennecke et al. 2007). piRNAs appear to be processed from long transcripts of median size 25 kb which can range up to 242 kb in length (Brennecke et al. 2007), consisting of a large number of interspersed fragments of transposable elements. The regions in the genome producing these transcripts are termed “piRNA clusters.” One cluster mutant, known as flamenco, produces lethal phenotypes when it is no longer transcribed, while weaker mutations within the cluster accompany an increase in transposition rates (Mevel-Ninio et al. 2007).

Most piRNAs in the *Drosophila melanogaster* genome appear to map to transposons, yet in chicken and other vertebrates, only 20-30% of piRNAs map to transposons (Li et al. 2013). In *C. elegans*, the piRNA pathway has expanded dramatically, with over 24 of the characteristic piRNA pathway ‘argonaute’ proteins in this species (Wedeles et al. 2013). One argonaute protein, PRG-1, appears to be required for the inherited epigenetic form of gene silencing known as RNAe, while another protein, CSR-1, appears to oppose this function, causing the creation of a distinct class of RNAs that seem to promote epigenetic gene activation (RNAa) (Seth et al. 2013).

In spite of the evidence of causality between piRNA clusters and transposon silencing, many aspects of piRNA biology remain mysterious. Several piRNA pathway proteins have been identified whose removals cause varying levels of increased transposition (See Table II-1), but the exact functions of these components, and relevance to the silencing of novel transposable element threats, remain hidden. In Qin mutants, several transposons increase in expression level despite the continuing presence of piRNAs against these elements (Zhang et al. 2014), while in Rhino and other mutants, piRNA production decreases dramatically without any change in several transposons targeted by these diminished piRNA populations (Klattenhoff et al. 2009). In addition, although translational silencing by cleavage of mRNA transcripts is the most accepted model of transposon silencing, there are strong examples in Drosophila and other systems of what appears to be transcriptional gene silencing of Transposable Elements via alterations in the methylation patterns of transposed material (Sienski et al. 2012).

The miRNA pathway

The miRNA pathway of animals is thought to operate by principles that seem at first glance to be similar to the piRNA pathway. Mature miRNAs of length ~22-24 nt are bound by an Argonaute protein, and target complementary sequence within a transcript in order to mediate repression. Unlike piRNAs, the targets of miRNAs are usually 3’ UTR regions of host genes, and the means of repression is rarely thought to be mediated by cleavage. Two favored mechanisms for target repression are translational inhibition and target mRNA destabilization, both of which are well supported within the literature (Bazzini et al. 2012, Guo et al. 2010). Also unlike piRNAs, miRNAs appear to only require perfect complementarity within a small portion of the target transcript, commonly known as the ‘seed’ (Lewis et al. 2003). In support of a model of miRNA-mediated repression of targets, studies have shown examples of loss of function phenotypes that can be recovered by underexpression of a single target gene or exacerbated with overexpression of the target or mutation of the relevant 3’ UTR target site to no longer bind the regulating miRNA (Brennecke et al. 2003). In further support of an essential role for miRNAs, many appear to be extremely highly conserved across species (Chen and Rajewsky 2006a).

However, as with piRNAs, miRNAs appear to act in a variety of contexts that make it difficult to discover how miRNAs act as an aggregate. Although most of the evidence indicates that miRNAs act by downregulating their targets, there are also studies that show upregulation of target transcripts by miRNAs (Vasudevan et al. 2007), and most of the bona fide targets of miRNAs decrease very subtly in mRNA level, even after the dramatic overexpression of their regulating miRNAs (Guo et al. 2010). Although many miRNAs have knockout phenotypes, there are many more well-conserved miRNAs whose loss causes no phenotype (Miska et al. 2007), and for all the phenotypes that are caused by modifications to miRNA expression, most identify individual target genes that can explain a large portion of the observed phenotypes, implying that many targets are evolutionarily conserved miRNA binding sites and yet appear to make no phenotypic contribution to the organism. Additionally, it is difficult to explain why so many miRNAs are conserved across their entire sequence, when only the seed region appears to be required for target regulation. Finally, although many miRNAs are well-conserved, the effectively targeted genes often appear to differ dramatically between species, suggesting that the functions of miRNAs may be species or clade specific in spite of broadly conserved sequence (Gao 2010, Xu et al. 2013, Tang et al. 2010, Griffiths-Jones et al. 2011, Chen and Rajewsky 2007 Liu et al. 2008).

The advantages of an Evolutionary approach

As seen above, an outstanding challenge in the study of small RNAs (and other fields) is distinguishing exceptional cases from prevalent rules. Molecular biology and genetics can definitively show how individual processes work. With enough experimentation, patterns begin to emerge, and these patterns can be integrated into generalizable models of biology. With deep sequencing techniques and other large-scale approaches, many interactions can be tested simultaneously, making it easier to see patterns that prevail across a whole genome or range of experimental conditions. Still, none of these techniques can on their own explain how the traits we might observe came about, or why some traits have persisted while others have not. For all of modern biology’s advanced techniques, some of the most fundamental questions of ultimate purpose are still unanswerable. We may describe how a signaling pathway occurs, but we can only guess as to whether there is an intrinsic reason for some organism's use of one set of signaling molecules while others use entirely different pathways to regulate similar processes, or what if anything determines which processes are modulated at the transcriptional level and which are modulated post-transcriptionally. For these types of questions, comparisons are needed between a large number of existing organisms, and a theory of how traits change over time. Through evolutionary analysis, we can reconstruct changes that have occurred, and begin to see which aspects of biology are dispensable, mutable, or arbitrarily chosen over time to solve a narrow problem, as well as core components that seem to be absolutely required for life. Through careful analysis, we can sometimes even see biological components currently in search of better solutions to new environments. In this way, by comparing present traits with inferred ancestral states, we can place the dazzling array of solutions biology has taken in extant species in a broader historical context, allowing us to sort through how every trait came to be, and ask questions of ultimate function. By applying the techniques of evolutionary biology to small RNAs, I have attempted to begin to answer some ultimate questions about RNA regulatory pathways, and differentiate the rules from the exceptions.

In my research, I have used a set of evolutionary approaches to ask a very narrow set of ultimate questions about the piRNA and miRNA pathways. Regarding the piRNA pathway, my focus has been on individual piRNA proteins, where I’ve asked if the evolutionary rates of the amino acids in piRNA proteins can tell us anything about the strength and timing of natural selection, and if these results in turn can tell us anything about the functions piRNA proteins have evolved to fill in silencing transposons. In the miRNA pathway, I ask whether interactions between conserved miRNAs and their target genes are changing rapidly or slowly, and use this to ask whether miRNAs are generally used in genomes for the regulation of broad or species-specific processes.

Previous evolutionary findings within the piRNA pathway

Many studies have recently found strong evidence of positive selection within the piRNA pathway. As a precursor to evolutionary studies of the piRNA pathway, Obbard et al. 2006 used *Ka/Ks* calculations (also known as *Dn/Ds*) and McDonald-Kreitman based measures to study RNAi pathway proteins including miRNA and siRNA proteins. While *Ka/Ks* tests compare rates at which observed mutations in the descendants of a common ancestor change amino acids (*Ka* or *Dn*) relative to rates at which mutations do not change amino acids (*Ks* or *Ds*), McDonald-Krietman tests ask whether nonsynonymous polymorphisms fix more or less frequently in divergent species than synonymous polymorphisms. In theory, only *Ka/Ks* rates>1.0 can be used as definitive proof of positive selection, but (*Ka/Ks*) tests rarely exceed 1.0 within functional proteins, as this rate would indicate that on average, nonsynonymous mutations are tolerated to the same extent as synonymous ones, and even the most rapidly evolving proteins must maintain strong purifying selection on some essential domains. Using the *Ka/Ks* technique, this study found that the siRNA proteins Dcr2, R2D2, and Ago2 are evolving among the fastest 3% of the genome, while the miRNA paralogs Dcr1, R3D1, and Ago1 are unexceptional. Similarly, using the McDonald-Kreitman approach, this study found that the proteins Dcr2, R2D2, and Ago2 all show a significant excess of fixed nonsynonymous changes relative to fixed synonymous changes in at least one pairwise species comparison, while none of the miRNA paralogs show these changes. When McDonald-Kreitman tests were repeated using only the functionally annonated or nonannotated domains of the siRNA proteins Ago2, R2D2, and Dcr2, only the nonannotated domains showed evidence of positive selection, consistent with strong purifying selection acting on the annotated functional domains.

In 2009, Obbard and colleagues followed up on this study with two similar publications. Obbard et al. 2009a was written as a review of evolutionary dynamics within RNAi genes and their targeted pathogens, but includes a *Ka/Ks* chart of the observed turnover rates of several piRNA proteins, including Krimper, Aubergine, Armitage, SpindleE, Piwi, Argonaute3, Zucchini, Maelstrom, and Squash. All genes except Piwi and Squash appear to fall in the 75th percentile or higher relative to other genes in the genome having similar length, and the genes Aubergine, Maelstrom, and Krimper appear to reside within the 95th percentile relative to size-adjusted genes from the rest of the genome. Obbard et al. 2009b is a primary research article that examines 136 immunity-related genes and 287 non-immunity genes from adjacent regions. This study used an extension of the McDonald-Kreitman approach to find that immunity-related genes in general are evolving extremely rapidly, and that RNAi-related genes (including those involved in the piRNA pathway) were among the fastest evolving subset.

A more recent study (Kolaczkowski et al. 2011) used both polymorphism-based and divergence-based approaches to examine 23 genes from the piRNA pathway. Polymorphism-based approaches used a program known as Sweepfinder, which determines the number of individuals having a derived SNP to the number having an ancestral SNP. Selective sweeps are expected to produce patterns in which a large number of individuals possess a single derived SNP in the region immediately adjacent to the selected allele, due to the so-called 'hitchhiking' of nearby neutral derived alleles that happen to have resided on the originally selected haplotype. This pattern then tapers off to a more neutral pattern farther away from the selected allele. Sweepfinder assesses the likelihood of a selective sweep relative to a neutral background that uses the observed SNP frequencies in the underlying dataset. Using this method on DPGP population data, Kolaczkowski et al. found that 14/23 piRNA pathway genes show evidence of selective sweeps. The study then continued to a divergence-based approach, which used the phylogenetic analysis by maximum likelihood (PAML) package under the 'branch-sites model' to assign one branch of a phylogenetic tree as being allowed to evolve under positive selection while constraining the rest of the phylogenetic tree to evolve neutrally or under selective constraint. For cases in which such a model fit the data substantially better than a model in which all branches were constrained to evolve neutrally or under selective constraint, the branch-sites model then assigned probabilities to individual amino acids that might be under positive selection, using a bayesian procedure. When this approach was implemented with data from the 12 genomes of the *Drosophila* 12 genomes project, 16/23 piRNA pathway proteins showed some branches with evidence of positive selection.

In order to attempt to categorize piRNA genes by the patterns of natural selection that act on them, Kolaczkowski et al. used a novel 'kmer approach' to re-analyze portions of their proteins, and investigate how branch-site results changed as the length of the portion increased. Finally, the sites identified by the branch-sites model as potentially positively selected were mapped onto individual domains of the 23 piRNA pathway proteins. These sites were found in both functional annotated domains and unstructured regions of the proteins surveyed.

Previous evolutionary findings within the miRNA pathway

In 2000, Pasquinelli et al. conducted an analysis of a founding miRNA, let-7, which used Northern probes to find that this miRNA is expressed in 14 species that include mammals, fish, molluscs, annelids, nematodes, and insects. This study also identified a target that had been identified previously in *C. elegans*, lin-41, as having conserved complementary sequence to let-7 in *D. rerio* and *D. melanogaster*. This work demonstrated that miRNAs are evolutionarily conserved.

In 2003, Lewis et al. published an algorithm known as TargetScan, which used a metric of basepairing energy, in combination with conservation across several species, to find targets whose baseparing energy to real miRNAs is substantially better than that of 'targets' of randomized control sequences. Using this process, 451 targets were found for 79 miRNAs conserved across their entire length in human and mouse. When conservation of targeting was required across broader species, real miRNAs performed substantially better than control sequences, even as the number of total predicted targets diminished, demonstrating that miRNAs tend to have more instances of highly conserved target sites than random sequences. 15 of these target 3' UTRs were tested by cloning downstream of a luciferase reporter. In 11 cases, expression of these targets increased when the suspected target site was mutated to no longer bind its cognate miRNA. This work demonstrated that miRNAs regulate many conserved targets, established a standard for miRNA target discovery that used evolutionary conservation as a metric of functional validation, and determined that the 7 nt seed of bp 2-7 was the most potent predictor of miRNA targeting.

Several papers were published in 2005 investigating the evolution of miRNA-targeting interactions. These included Farh et al. 2005, Lewis et al. 2005, Xie et al. 2005, and Brennecke et al. 2005. In addition to several experimental validations of the principles of miRNA targeting, Brennecke et al. 2005 demonstrated that 7mers and 8mers corresponding to miRNA binding sites (positions complementary to nucleotides 1-8 of mature miRNAs) are much more conserved than control sequences in 3' UTRs, that using this conservation as the only signal performed as well if not better than original TargetScan algorithms, and that as few as 4 nucleotides matching within the seed may be enough to confer 3' UTR regulation in experimental systems. This paper posited a mechanism by which novel targeting interactions may evolve, with initial targeting governed by a short seed match, followed by more extended basepairing in the remainder of the seed or 3' compensatory basepairing to modulate the strength of target interactions.

Lewis et al. 2005 reported an improved TargetScan algorithm that explicitly used conservation within seed regions of miRNAs as its only criteria for predicting targeting. Using this approach, over 5,300 genes were predicted as conserved, biologically relevant targets of miRNAs. This algorithm found, notably, that an 'A' nucleotide is often highly conserved opposite the first nucleotide of conserved miRNAs within 3' UTR alignments, and that by requiring an 'A' at this position of target sites, the signal of conservation of miRNA binding sites could be improved substantially relative to randomized control sequences and relative to sites requiring basepairing to the first nucleotide of miRNAs.

Xie et al. 2005 produced an investigation of all conserved motifs in 3' UTRs and promoters. This approach rediscovered many known transcription factor binding sites, as well as finding a large number of known miRNA binding sites and predicting the existence of a large set of previously unknown miRNAs.

Farh et al. 2005 extended previous methods to the analysis of the spatiotemporal expression of miRNAs and their targets. In this work, the authors noted that conserved target sites make up approximately 1/10th of all seed matches to any given miRNA. Reporter assays were therefore constructed to test the activity of these nonconserved sites. Both conserved and nonconserved sites were effectively repressed by miRNAs, indicating that even nonconserved miRNA binding sites may be functional in-vivo. However, unlike conserved sites, nonconserved sites were often observed in genes not normally co-expressed with their predicted regulatory RNA. Additionally, genes that were tissue-specific and highly expressed were found to be depleted for miRNA binding sites, suggesting the possibility that these genes co-evolved to be specifically missing these binding sites. These results showed that nonconserved sites are functional and can be strongly selected against, being tolerated preferentially in circumstances in which they are not co-expressed with their regulatory miRNA. In combination with work by Brennecke et al. 2005 showing that a short seed match is often sufficient for target regulation, these results also opened the possibility that newly evolved beneficial miRNA-target interactions may occur and be selected for with relative ease.

In 2006, Chen and Rajewsky published two papers dealing with the evolution of miRNAs and their targets. Chen and Rajewsky 2006a applied the methods of Xie et al. 2005 to nematodes, flies, and vertebrates, finding highly conserved motifs present across all three clades. This study also examined individual miRNA-target relationships, and asked how many were conserved in two out of three clades, or in all three clades. Although many miRNAs are highly conserved in all three clades, this study found the surprising result that there are only slightly more target interactions conserved across pairs of clades and across all three clades than would be predicted by random chance, leading the authors to conclude that miRNA-target interactions have undergone extensive rewiring during metazoan evolution. These results are somewhat conflicted, as the authors note that target prediction algorithms are effective in part because of deep conservation of miRNA targets within clades, yet the results between clades support little or no targeting conservation across deeper evolutionary time relative to random chance. The authors go on to suggest future work quantifying the relative evolutionary rates of transcription factor and miRNA regulatory networks.

Chen and Rajewsky 2006b used human SNP data from the HapMap and Perlegen projects to examine selective pressures operating on miRNA binding sites. While conserved miRNA binding sites appeared to have lower SNP densities, an excess of rare SNPs, and more polymorphisms than divergent events relative to conserved control regions of 3' UTRs not associated with miRNA binding sites, nonconserved miRNA binding sites also exhibited significant evidence of negative selection against loss. The authors interpreted these results as supporting the conclusion that ~85% of conserved miRNA binding sites show evidence of signficant negative selection against loss, while 30-50% of nonconserved miRNA binding sites may also incur negative selection against loss when co-expressed with their regulatory miRNA. These results therefore reinforce prior findings of the conserved nature of miRNA binding sites while continuing to support functionality and selection for the maintenance of nonconserved binding sites.

In 2007, Chen and Rajewsky published a review article, entitled 'the evolution of gene regulation by transcription factors and microRNAs'. This article pointed out that very few highly conserved targets of miRNAs had been identified as of 2007, and cited several studies indicating that the rate of miRNA-target conservation might be relatively low, even within relatively short time spans. The article continued to discuss rates of creation of novel miRNAs vs. transcription factors, noting that while transcription factors can tolerate mismatches and still function, miRNAs do not tolerate mismatches, leading to an effectively larger binding site needed for miRNA repression of a target, and making it much easier to lose a given binding site over evolutionary time than to gain one. The authors argue that perhaps the creation of novel miRNAs utilizing existing binding sites, followed by the selective removal of deleterious regulatory sites, is easier than the selective accumulation of novel binding sites for an existing miRNA, using results from an earlier study showing that evolution of an eight nucleotide binding site by chance point mutation would take approximately 650 million years on average, while a seven nucleotide site (presumably to a transcription factor) would only take 60,000 years. The article then proposes a model of miRNA evolution in which newly evolved miRNAs may be initially expressed at very low levels that do not create strong detrimental interactions, and then later (after weak negative selection has removed these slightly detrimental interactions) acquire higher expression. In support of this idea, Chen and Rajewsky cite a study reporting that ancient miRNAs are expressed at higher levels than newly evolved ones, concluding that miRNAs may evolve more readily than transcription factors, allowing them to play roles in more recently evolved gene regulation than transcription factors. This idea is intriguing, particularly from the viewpoint of miRNAs as rapidly evolving regulators of genomic processes, but doesn't explain why so many miRNAs are well conserved, or how species-specific and clade-specific targets of conserved miRNAs appear with relatively high frequency.

For a counterpoint to the idea that miRNAs fill essential evolutionary niches, I read Lynch 2007, a paper which noted that multicellular organisms have more genomic space for transcription factor and miRNA binding sites to acccumulate within, and that multicellular organisms tend to have smaller effective population sizes than prokaryotes, decreasing the effective strength of selection in these species and allowing weakly deleterious regulatory systems to persist where they would be purged in prokaryotes. The paper goes on to state that the large number of transcription factors present in many Eukaryotes virtually guarantees the neutral evolution of novel regulatory interactions by neutral processes, challenging the notion that functional redundancy and distributed robustness of genetic networks can be acted on by natural selection, and concludes that regulatory complexity and redundancy may evolve as a byproduct of large genome sizes and weak selective constraints that dominate in more complex organisms, rather than being fixed by natural selection.

In 2009, two publications examined the evolution of genetic robustness from the perspective of miRNAs. Genetic robustness is the process by which phenotypes encoded by genetic networks are stabilized against environmental fluctuations and genetic perturbations elsewhere in the genome. In theory, robustness has an interesting relationship with natural selection, in that natural selection may act to confer robustness to systems that would otherwise be frequently misregulated, but at the same time, by shielding phenotypes from variations in genotype, robustness may serve to modulate the ability of natural selection to operate on heritable variation. The first paper to examine miRNAs in this context was Wu et al. 2009. This paper postulates that, in addition to roles in 'tuning' gene expression, miRNAs may serve to impart genetic robustness against genetic and environmental perturbations through a variety of mechanisms, 'buffering' gene expression to remain constant in spite of environmental fluctuations. In one of these mechanisms, stimuli that normally would influence expression of some target gene are tied to the activation or repression of a miRNA that inhibits this same target gene, with the net effect that expression of the target gene is stabilized even as the initial gene fluctuates in expression level. In another, a gene upregulates a miRNA that acts to repress the same gene. Examples of a large number and diversity of these buffering type interactions are cited from the literature. During the same year, Li et al. 2009 present evidence that the removal of miR-7 results in instability in embryonic development upon subjection to oscillating warm and cool temperatures. In 2013, Cassidy et al. followed up this work with a study of miR-9, which influences cell fate by repressing the target *senseless*. This study found that when senseless was rendered insensitive to miR-9 through a modified 3' UTR target site, underlying genomic differences exerted a significantly larger effect on cell fate outcomes than in the presence of miR-9, thus demonstrating that in this instance, miR-9 promotes the robustness of this network against genetic differences, and masks the ability of natural selection to act on these differences.

In 2010, Loh et al. reported a study of cichlid fish from Lake Malawi, in Malawi, Africa. The cichlids in this lake have speciated into hundreds of species with diverse morphologies, behaviors, and ecological niches, yet share a common ancestor less than one million years ago, and share most identified segregating SNPs across species boundaries. In this study, higher SNP density was found within miRNA binding sites than in other regions of 3' UTRs and the rest of the genome as a whole, and divergent SNP expression showed evidence of positive selection and biologically relevant correlations with observed phenotypic differences between species, leading to the conclusion that shifts in miRNA binding sites may have recurrently driven phenotypic adaptation to a variety of ecological niches.

The studies described above explore several aspects of miRNA and target site evolution. Initial studies found that some miRNAs have targets that are highly conserved (Pasquinelli et al. 2000), that miRNAs themselves are deeply conserved (Chen and Rajewsky 2006a), and that miRNAs have many more conserved targets than similar random sequences (Lewis et al. 2003, Lewis et al. 2005, Chen and Rajewsky 2006a).

There is also evidence that 9/10 of potential targets of miRNAs are not conserved (Farh et al. 2005), that many of these nonconserved sites are likely to be functional, both through molecular evidence (Farh et al. 2005) and analysis of signatures of selection operating on SNPs (Chen and Rajewsky 2006b), that very few miRNA-target interactions appear to be deeply conserved in the distantly related clades that miRNAs themselves are conserved in (Chen and Rajewsky 2006a), and that recently speciated genuses may have exploited these nonconserved sites to explore novel phenotypic space (Loh et al. 2010).

Theoretically, the field has seen several very different arguments that attempt to encompass various traits of miRNAs. While some claim that miRNAs acquire novel targets very slowly through mutagenesis, and argue that novel miRNAs are more likely to evolve new functions than existing ones (Chen and Rajewsky 2007), others use the short sequence specificity of miRNAs to argue that seed matches to beneficial targets may be selected for relatively easily with as little as four basepairs of complementarity (Brennecke et al. 2003) or to claim that the number of novel targets of miRNAs accumulate by neutral drift (Lynch 2007). Finally, there is the argument made by Cassidy et al. 2013, Li et al. 2009, and Wu et al. 2009 that miRNAs are naturally selected to buffer gene expression, implying that the plasticity of targeting may itself be dependent on whether maintaining a stable phenotype or selecting a new one is beneficial.

To me, it seems possible that dynamics of conserved and rapidly evolving function may both operate simultaneously. If the results of Brennecke et al. 2003 can be believed, acquiring a novel function for an existing miRNA may be relatively simple, and positive selection or some other factor may accelerate the evolution of longer miRNA binding sites. If this is the case, miRNAs may be conserved through the utility of their targeting interactions. The results of Farh et al. 2005 strongly suggest the presence of 'antitargets' and selective avoidance of detrimental miRNA binding sites in highly expressed genes. Taken in combination with Chen and Rajewsky's 2007 argument for the birth of novel miRNAs by the slow purging of detrimental targets, and the hundreds of targets predicted for each individual miRNA in Lewis et al. 2005, it may be the case that conserved miRNAs are conserved simply because having a regulatory molecule that has already undergone the process of purging detrimental targets, and whose expression can be linked to hundreds of genes with only a seed match to its targets, is useful and easier to maintain over time than evolving a novel miRNA for every new regulatory interaction. If such a regulator evolves early enough, and is preserved for a long enough time span, some targets may accumulate essential roles that prove useful over long periods of time, leading to a strong signature of selection on these target interactions. The conservation of target interactions may therefore be seen as a neutral consequence of the conservation of miRNAs: given enough time, some conserved interactions will develop, and conserved interactions will accumulate as time progresses, until, after enough time, a substantial fraction of the targets of conserved miRNAs will also appear to be conserved. But these processes do not rule out the functionality of miRNAs in newly evolved interactions, and this model still does not say enough about whether miRNAs fill some essential evolutionary role not addressed by transcription factors, or accumulated their functions through drift or even a failure to purge weakly deleterious and overly complicated regulatory systems. It is for these reasons that I set out to develop a better method to quantify the expected and actual rates at which the targets of miRNAs change over time.

Chapter II

Recurrent and recent selective sweeps in the piRNA pathway.

This chapter derives from the article of the same name, and was part of a collaboration between Alex Wong, who supervised the correct parameterization of the PAML package, Yu-Ping Poh, who procured syntenic alignments of the relevant piRNA proteins within species, William Theurkauf, who assisted with the biological interpretations of the results, and Jeffrey D. Jensen, who suggested the appropriate tests to run in order to detect selection at varying timescales, and assisted in troubleshooting the required programs and interpreting results. All authors contributed extensively to manuscript preparation. This work appeared in the journal Evolution; International Journal of Organic Evolution, (2013) 67(4), 1081–90.

Chapter II Summary

Uncontrolled transposable element (TE) insertions and excisions can cause chromosome breaks and mutations with dramatic deleterious effects. The PIWI interacting RNA (piRNA) pathway functions as an adaptive TE silencing system during germline development. Several essential piRNA pathway proteins appear to be rapidly evolving, suggesting that TEs and the silencing machinery may be engaged in a classical “evolutionary arms race.” Using a variety of molecular evolutionary and population genetic approaches, we find that the piRNA pathway genes *rhino, krimpe*r, and *aubergine* show patterns suggestive of extensive recurrent positive selection across *Drosophila* species. We speculate that selection on these proteins reflects crucial roles in silencing unfamiliar elements during vertical and horizontal transmission of TEs into naïve populations and species, respectively.

Chapter II Introduction

piRNAs have been identified as the primary germline silencing agents for TEs (Brennecke et al. 2007). In the current model for TE silencing, 23-30nt piRNAs, in complex with PIWI clade Argonaute proteins, recognize and cleave complementary TE mRNAs. In *Drosophila*, piRNA silencing begins with a pool of pre-existing piRNAs, termed primary piRNAs. piRNAs can replenish themselves (see below), but appear to require pre-existing maternally inherited piRNAs to prime the system. piRNAs are encoded by specialized 25-240 kb heterochromatic loci composed of nested TE fragments, termed piRNA clusters (Brennecke et al. 2007). The primary piRNAs which are complementary to TE sequences, termed "anti-sense stranded piRNAs" are bound by the PIWI protein Aubergine and cleave sense stranded TE transcripts, silencing expression and generating the precursors of so-called "sense strand" piRNAs that associate with the PIWI protein Argonaute 3 (Ago3). The Ago3-sense strand piRNA complexes cleave cluster transcripts, producing precursors for antisense stranded piRNAs (Gunawardane et al. 2007). The piRNA pathway is therefore composed of genetically defined proteins and clusters which require epigenetically inherited small RNAs to amplify and transmit silencing activity. The *Drosophila* RNAi and miRNA pathways, by contrast, do not appear to generate epigenetically heritable silencing activity.

Several recent studies have examined the evolution of small silencing RNA pathway proteins, including some with roles in piRNA silencing. Utilizing McDonald-Kreitman based approaches (Obbard et al. 2006), polymorphism-based composite likelihood tests of selection (e.g., CLSW (Kim and Stephan 2002) and Sweepfinder (Nielsen et al 2005)), and divergence-based analyses (e.g., Phylogenetic Analysis by Maximum Likelihood (PAML); Yang 2007), these studies have shown that adaptive evolution is frequent within the RNAi pathway, which provides anti-viral activity and is involved in transposon-silencing. These observations suggest that viral infection and TE activity drive evolution of these small RNA based silencing pathways. Numerous studies have demonstrated that certain classes of TEs have exhibited bursts of activity in the recent past (e.g. Yang et al. 2006, Diaz Gonzalez et al. 2010). However, Castillo et al. (2011) have used divergence estimates derived from PAML to show that positive selection in the piRNA pathway does not correlate with either the number of TE families or amount of TE sequence in the genome. This suggests that positive selection of piRNA pathway proteins is limited by purifying selection maintaining the core functionality of the piRNA pathway. Using both Sweepfinder and PAML, Kolaczkowski et al. (2011) find pervasive evidence of selection in both piRNA pathway and RNAi pathway genes, and note that RNAi pathway proteins with the strongest evidence of selection tend to be those that interact most directly with double stranded RNA, consistent with the idea that selection is strongest at the interface between target RNA molecules and silencing machinery. Others have noted that there is strong evidence of recent positive selection centered on Argonaute2 in *D. melanogaster, D. simulans*, and *D. yakuba*, which the authors attribute to a longstanding arms race with viral and/or TE antagonists (Obbard et al. 2011).

We extend existing evolutionary approaches at higher resolution using multiple timescales, through an examination of the evolutionary pressures operating on ten piRNA pathway proteins (Table 2-1) using a combination of divergence-based and polymorphism-based methods, and find that selection is surprisingly non-uniform. However, we find strong evidence of positive selection within a core set of piRNA proteins in both divergence and polymorphism datasets, leading us to propose a model in which TEs recurrently exploit the same proteins as they first invade and then spread through natural populations.

Chapter II Materials and Methods

We investigated the evolutionary history of ten piRNA pathway genes, chosen by dramatically increased transposition rates in double or single knockout individuals (see Table II-1) using two main approaches: divergence-based statistics were used to detect recurrent selection across multiple lineages, and polymorphism-based statistics were used to detect recently completed species-specific selective sweeps.

Divergence based evolutionary analysis

Our divergence-based analyses used three basic aspects of the PAML package (Yang 2007), referred to here as the sites test of selection, the branch test of selection, and the branch-sites test of selection. All of these divergence-based approaches give estimates of the ratio of nonsynonymous changes (*dN*) to synonymous changes (*dS*), with each test implemented to detect selective pressures under a different, narrow set of assumptions. Ratios of *dN/dS* greater than 1 can be attributed to positive selection driving nonsynonymous fixation, ratios near 1 are consistent with neutrality, and ratios smaller than 1 may be attributed to the action of purifying selection on non-synonymous sites. In all cases, a likelihood ratio test (LRT) is used to compare a neutral null model with an alternative model allowing positive selection.

Divergence tests for positive selection were carried out among six closely related *Drosophila* species (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba, D. erecta,* and *D. ananassae)* using the PAML analysis package (Table 2-1, Yang 2007). Only those sequences no more divergent from *D. melanogaster* than *D. ananassae* were considered (roughly corresponding to a per site substitution rate of 1.0) to avoid the issue of saturation of synonymous sites which has been shown to cause dS to appear artificially small in highly divergent species, thus overestimating *dN/dS* in divergent clades (Stark et al. 2007, Clark et al. 2007). The protein-coding cDNA of all proteins among these six species was obtained from the flybase (release 5.29) website (Tweedie et al. 2009) and aligned using PRANK alignment software with the codon alignment option (Löytynoja and Goldman 2005). Where a single ortholog was not annotated among all six species, best reciprocal blast annotated transcripts were chosen for analysis. Finally, where no best reciprocal blast hits were returned, syntenic alignments were collected from the UCSC genome browser as recommended by B. Kolaczkowski and A. Kern (personal communication). The Rhino protein sequences were obtained directly from Genbank annotations provided in the supplementary methods of Vermaak et al. (2005).

In order to examine positive selection on individual amino acids within a background of purifying selection, we utilized the sites model of PAML. If it is assumed that selective pressures do not vary across the phylogeny, PAML can use this model to estimate the distribution of selective constraints across the length of a protein. The sites model uses three pairs of null and alternative models, termed M1a vs. M2a, M7 vs. M8, and M8a vs. M8 (see Supplemental Online Material). In all three comparisons, the null model assumes all codons evolve only under purifying selection or neutrality, while the alternative allows the possibility of an additional class of codons which evolve under positive selection.

The branch-based method relies on a user-input phylogeny, and compares a model in which *dN/dS* values of each lineage are fixed at the same value to a model in which *dN/dS* scores vary across lineages. In addition to log likelihood scores, the null model produces an estimate of the fixed maximum likelihood *dN/dS* across all lineages, while the alternative estimates individual maximum likelihood estimates for each branch.

The branch-sites model separates all possible lineages into groups, such that one group is designated as “background” and constrained to evolve neutrally or under selective constraint, while the other is designated “foreground” and allowed to contain, in addition to neutral or constrained sites, sites with *dN/dS* values greater than one. This alternative model is measured against a more constrained model in which the foreground branch is also constrained to be neutral or negatively selected (*dN/dS*<=1).

To test the significance of the above methods within the context of the *Drosophila* genome, we obtained a full list of protein-coding annotated orthologs from Flybase. Only *D. melanogaster* genes with exactly one annotated ortholog in the species *D. simulans, D. sechellia, D. yakuba, D. erecta*, and *D. ananassae* were included. cDNA of these orthologs was aligned using PRANK alignment software as above, and only sequences with greater than 20 alignable amino acids across all 6 species were retained. Sequences that did not complete in any PAML tests were removed from all analyses, and the resulting 9,005 genes were used as a rough proxy to estimate the genomic distribution of selective pressures.

In permutation tests drawing random sets of genes from this genomic set in order to assess significance, the following criteria were used to assign individual genes to a putatively positively selected category. For the branch model, only genes which met the 3 criteria of rejecting equal *dN/dS* values across all lineages, containing one or more lineages with elevated *dN/dS* >0.5, and possessing *dS* values >0.02 for these elevated *dN/dS* values were considered to be putatively positively selected (*i.e.*, in order to avoid false inference owing to low *dS* values (Supplementary Figure II-1)). For the sites model, genes drawn from the permutation which rejected neutrality under at least one of the model pairs M1a vs. M2a, M7 vs. M8, and M8a vs. M8 were examined for presence of individual amino acids with high probabilities of selection. For the branch-sites model, lineages that rejected neutrality were considered to be under putative positive selection, and if these lineages also possessed amino acids with a high probability (>0.95) of selection, these amino acids were also considered to be putatively positively selected.

Polymorphism based evolutionary analysis

As a complement to the purely divergence-based analyses, we investigated variation at the population level using 1) site frequency spectrum based tests of selection (e.g., CLSW (Kim and Stephan 2002)), 2) site frequency spectrum based tests of neutrality (e.g., the *H* statistic (Fay and Wu 2000), as well as 3) the McDonald Kreitman (1991) test.

The alignments studied here consist of syntenic assemblies from an average of 34 *D. melanogaster* individuals from North Carolina, an average of 6 *D. melanogaster* individuals from Malawi, and 6 *D. simulans* individuals from several inbred stocks, analyzed separately for each of the ten piRNA pathway proteins. We used polymorphism data from the Drosophila Population Genomics Project (DPGP). Ambiguous nucleotides were conservatively analyzed by changing them to the most common nucleotide at the site among the population. These alignments were converted into ms format using *D. yakuba* as an outgroup ancestral state using an online recombination rate estimator derived from mapping studies (Fiston-Lavier et al. 2010) and an assumed population size of 1,000,000 individuals.

Using the frequency of polymorphisms across the sequence, CLSW assigns a likelihood to models of selection and neutrality, and performs a ratio test. The likelihood ratio scores from CLSW were compared to 1000 neutral simulations of identical θ ($=4N_e\mu$), ρ($=4N_er$), and length (in basepairs) parameters generated using the ssw simulation program (Kim and Stephan 2002) – simulations which were also used to assess significance of Fay and Wu's *H* test (2002). Significant regions ($p<0.05$) were further compared to 1000 selection simulations generated using ssw assuming a beneficially fixed mutation immediately prior to sampling ($\tau=0.000001$ 2N generations). The goodness-of-fit statistic (GOF: Jensen et al. 2005) was used to determine whether selection alone was sufficient to explain the data.

The McDonald Kreitman test is a comparison of nonsynonymous polymorphism to divergence compared with synonymous polymorphism to divergence. The expectation under neutrality (for which synonymous sites here serve as a proxy) is that the rate of fixation is simply given by the neutral mutation rate. We utilized two population samples of *D. melanogaster* (deriving from Malawi and North Carolina), and determined divergence as compared to *D. simulans*.

Chapter II Results

Species-level analyses

In the sites model of PAML, two of three tests of Ago3 rejected neutrality in favor of positive selection, but failed to localize this selection to any individual amino acid, while Piwi was identified as containing amino acids with a high posterior probability of positive selection but rejected neutrality in only one out of three model comparisons (M7 vs. M8). Only Rhino was significant across all three model comparisons and identified individual amino acids with a high posterior probability of positive selection (Table II-2 columns 3 and 5-7), with approximately 0.011% of all genes in our genomic set showing a similar or more extreme pattern (Table II-2).

In the branch model allowing *dN/dS* to vary among lineages, estimates of *dN/dS* are produced for each branch. Nearly all piRNA pathway proteins fit the varying *dN/dS* model significantly better than the one constraining selection to be identical along all lineages, suggesting that selective pressures vary across the phylogeny. Few proteins were inferred to have *dN/dS*>1.0. However, even proteins experiencing positive selection on key amino acids can be expected to evolve under strong purifying selection along most of their length, producing average *dN/dS* values less than 1. The Rhino protein is one such example, showing clear evidence of positive selection within individual amino acids and rejecting neutrality in all three sites model comparisons, yet evolving generally under constraint, with *dN/dS*<1.0 in the branch model in all lineages. We noted that the piRNA pathway proteins Rhino and Aubergine each had several lineages with *dN/dS*>0.5, higher than we would expect for proteins with sterile loss of function phenotypes, which we would expect to evolve under strong selective constraint with *dN/dS* near 0. The other six piRNA proteins had no lineages with such elevated *dN/dS* (Figure II-2A; Table II-2).

In order to generate an empirical distribution, we compared these values to *dN/dS* estimates performed across 9,005 proteins with 1:1 annotated orthologs among all six species, and found that only 7% of proteins possessed 1 or more lineages with *dN/dS*>0.5. Performing a permutation test (10 proteins randomly chosen from our genomic distribution with 10,000 replicates), 0.4% of permutations contain 4 or more proteins with 1 or more lineages having *dN/dS* values >0.5. Because only a subset of genomic proteins with elevated *dN/dS* are likely to be essential, our permutation results present an overestimate of the number of essential genomic proteins with similarly elevated *dN/dS*. Therefore, while we cannot rule out relaxed selective constraint as a cause, it is plausible that the statistically enriched elevations in *dN/dS* among our protein set are attributable to positive selection in a subset of amino acids (see also Swanson et al. 2004).

It is notable that the sites model assumes no variation in selective pressures across branches, an assumption which the branch model suggests is not valid, while the branch model cannot assign a probability of positive selection to a given branch or amino acid. Therefore, in order to distinguish between lineage-specific positive selection among individual amino acids and relaxed selective constraint, the branch-sites test of selection (Zhang et al. 2005) was implemented to allow for positive selection along an individual lineage while constraining the rest of the phylogeny to evolve neutrally or under negative selection. The probability of such a scenario is assessed relative to a nearly identical model that assumes no positive selection on the lineage under examination or the rest of the phylogeny. By assuming selection along a single lineage, the probability of positive selection along a particular lineage can be directly measured in a way that is not possible in the branch model. The branch-sites model thus gains power relative to the branch model to detect the probability of lineage-specific selection at the expense of a loss of power to detect positive selection operating across more than one lineage, making these two tests, to some extent, complements of one another.

In contrast to the earlier analysis of sites positively selected across all lineages, which failed to find individual amino acids under selection in most piRNA proteins, the branch-specific tests were able to reject neutrality in favor of positive selection in several proteins, and identified individual amino acids with a high posterior probability of positive selection in Krimper and Aubergine. This suggests that the elevations in *dN/dS* seen in the branch model in these proteins could be attributable to site and lineage-specific positive selection. Branch-sites tests consistently revealed the lineages *D. melanogaster*, *D. simulans,* and *D. sechellia* to be undergoing some combination of recurrent positive selection within these proteins. Other piRNA proteins also were identified as putatively positively selected, but without localization to any individual amino acids. SpindleE and Rhino showed evidence of positive selection within *D. ananassae*, Zucchini showed evidence of some recurrent positive selection in the *D. melanogaster*, *D. simulans*, and *D. sechellia* lineages, and Rhino showed evidence of positive selection in the ancestor of *D. yakuba* and *D. erecta*. Notably, Rhino—for which individual amino acids were identified as under recurrent positive selection in all three sites models—did not show evidence for positive selection on these amino acids in any individual lineage under the branch-sites model. This lack of overlap illustrates the respective power of the branch and branch-sites models to detect recurrent and lineage-specific positive selection acting on individual amino acids.

Within Aubergine, 24 sites were estimated in the branch-sites model to have a probability greater than 95% of being under positive selection. These amino acids do not appear to be centralized within one domain, but rather are dispersed across the length of the protein, similar to the findings of Kolaczkowski et al. (2011). Krimper had five sites estimated to be similarly positively selected within an annotated tudor domain and a non-significant Pfam match to a second unannotated tudor domain, as well as two additional amino acids outside of either domain.

When the branch-sites and branch results are summarized, it is notable that Rhino, Aubergine, and Krimper all have lineages with *dN/dS* >0.5 and appear to have strong evidence of positively selected amino acids in the branch-sites or sites tests. Based on a genome-wide permutation test, the probability of such an observation in a set of ten random proteins = 0.007. These observations suggest that a large portion of the piRNA pathway, as defined by the ten proteins examined here, is shaped by recurrent positive selection.

Population-level analyses

In CLSW tests with the GOF correction, the North Carolina *D. melanogaster* group rejected neutrality in favor of selection in Aubergine, SpindleE, and Rhino (Table II-3). Within *D. simulans*, Armitage, Vasa, and Zucchini all rejected neutrality in favor of selection (Table II-3). In McDonald Kreitman tests, Armitage, Aubergine, Krimper, and SpindleE all rejected neutrality in both *D. melanogaster* populations, while Vasa rejected neutrality in the Malawi population but not the North Carolina population. Performing Fay and Wu's *H* test separately for each piRNA gene within each population, twelve such tests rejected equilibrium neutrality (Table II-S1).

Chapter II Discussion

We find evidence of pervasive positive selection operating in the piRNA pathway – most notably within Rhino, Aubergine and Krimper, which have strong divergence-based and polymorphism-based evidence for both recurrent and recent strong positive selection. In earlier studies, these three proteins do not stand out relative to the rest of the piRNA pathway.

Here, we utilize a new alignment algorithm, PRANK, which has been shown recently (Fletcher and Yang 2010) to have a dramatically lower false positive rate in detecting positive selection in PAML branch-sites simulations while incurring only modest sacrifices in the detection of true positives. Additionally, our use of the branch, sites, and branch-sites models allows for the detection of positive selection operating across the entire phylogeny. Also, using recent genomic resources from the DPGP we are able to evaluate two distinct population samples of *D. melanogaster* (Malawi and North Carolina). Finally, in order to characterize the relative mode and tempo of selection as compared to other coding regions, we examine a dataset of 9,005 proteins in divergence analyses. This allows for a genomic distribution of selective pressures relative to which we can compare the results from the piRNA pathway.

While we see strong evidence of positive selection in piRNA proteins consistent with an evolutionary “arms race”, it is difficult to account for the mechanism by which natural selection operates through a high substitution rate in TEs alone. Because the units conferring resistance to TEs are genetically inherited cluster insertions and epigenetically inherited mature piRNA pools that target TE transcripts, we would expect these variants to sweep through populations to confer resistance to a novel threat. Modifications to piRNA proteins that allow for these variants to occur, by contrast, would not be under selective pressure. Furthermore, because clusters have extensive sequence complementarity to the TEs they regulate, it is unlikely that a silenced TE could evade the piRNA pathway through mutation without destroying functionality. In order to explain the recurrent fixations we observe consistently in Rhino, Aubergine, and Krimper, we therefore speculate that the adaptive silencing mediated by the piRNA pathway is responding to a hypothetical class of TE encoded inhibitors.

All 10 piRNA pathway proteins we examined have key roles in TE silencing and germline development. The clear prominence of only Krimper, Aubergine, and Rhino within our analyses is therefore quite unexpected, and is consistent with specific roles for these proteins in the adaptive response to novel elements. TEs are thought to spread through populations predominantly through direct inheritance. Therefore most TEs will be transmitted with their silencing clusters. However, piRNAs are epigenetically inherited maternally, and amplification of the silencing RNA pool requires pre-existing piRNAs. Paternally inherited TEs thus escape silencing in hybrids, leading to genetic instability and sterility. As hybrids age, however, piRNAs are produced *de novo* from paternal clusters, TEs are silenced and fertility is recovered (Khurana et al. 2011). If the initial failure to generate primary piRNAs from inherited clusters is mediated in part by inhibitors encoded by the invading element, vertical TE spread may impose strong, recurrent selection within the host for genetic variants that evade these inhibitors and thus enhance *de novo* production of primary piRNAs from existing cluster transcripts, allowing for the observed re-establishment of fertility.

The evidence we see of positive selection acting on Krimper and Aubergine is consistent with a previous analysis of *dN/dS* conducted in *D. melanogaster* and *D. simulans* which found that these two proteins have the highest rates of amino acid substitution in the piRNA pathway (Obbard et al. 2009a). These proteins may therefore represent promising candidates for further study of the adaptive response to new TEs. Aubergine is a PIWI protein that binds to mature piRNAs and has a direct role in the cleavage of piRNA precursors needed to amplify the primary piRNA pool (Brennecke et al. 2007). Krimper has a Tudor domain, and many Tudor domain proteins appear to directly bind dimethylated PIWI proteins (Siomi et al. 2010). This observation, the observed positive selection in both divergence based (Figure II-2, table II-2) and polymorphism based (Table 2-3, Table 2-4) analyses, and the co-localization of Krimper and Aubergine in the nuage complex (Figure 2-1), open the possibility that they may directly interact to process novel TEs into mature piRNAs in dysgenic scenarios even in the absence of pre-existing guide RNAs, allowing them to activate their inherited clusters.

During horizontal transfer of TEs between species, by contrast, silencing appears to require insertion of invading elements into clusters, which generates piRNA precursors capable of initiating the amplification and silencing cycle. These occurrences may be surprisingly frequent, as evidenced by the introduction and fixation of the P-element within *D. melanogaster* over the course of the last

40 years, likely from *D. willistoni* (Anxolabéhére et al. 1988, Daniels et al. 1990), as well as the great diversity of TEs within and between Drosophila species (Yang et al. 2006, Clark et al. 2007, Diaz Gonzalez et al. 2010).

Rhino localizes to heterochromatin and is necessary for the production of piRNAs from dual strand clusters (Klattenhoff et al. 2009). Rhino could therefore play some role in directing transposition into clusters during horizontal transfer, perhaps through interaction with the transposition machinery. Alternatively, TEs may encode proteins that inhibit transposition into clusters to avoid silencing. Both models predict that Rhino will be under selection during horizontal transfer, but not vertical transfer, and are consistent with recent studies indicating that clusters, piRNA populations, and siRNA populations change dramatically and often globally on very short evolutionary timescales in response to changes in TE composition (Khurana et al. 2011, Rozhkov et al. 2010, and Rozhkov et al. 2011).

These evolutionary insights should help guide studies on the epigenetic and genetic functions of rapidly evolving piRNA pathway proteins in TE silencing within naïve populations and species. The sterile phenotype of piRNA mutants, the rapid accumulation of TEs in the Drosophila phylogeny, and the increasing number of studies demonstrating hybrid dysgenesis suggest that these events may be a strong and perhaps surprisingly frequent contributor to the complex interplay between piRNA pathway function and TE propagation.

Acknowledgments

We would like to thank Andrew Kern, Bryan Kolaczkowski, Nadia Singh and members of the Jensen and Theurkauf Labs for numerous intellectual contributions.

Funding

This work was funded by a grant from the National Science Foundation [DEB-1002785], a faculty award from the Worcester Foundation to JDJ, a grant from the National Institutes of Health to Fen-Biao Gao [R01 NS066586], and a grant from the National Institutes of Health to William Theurkauf [RO1 HD049116.

Figure II-1: Transposon control by the piRNA pathway. Black arrows represent steps in pathways, while blue and red colored arrows represent sense and antisense RNA transcripts. Bi-directional cluster transcripts are produced from both strands of loci with homology to transposons. During hybrid dysgenic scenarios, these transcripts are cleaved by unknown mechanisms, which we speculate may be associated with modifications to Piwi and Krimper (?a) that produce primary "seed" piRNAs. A large number of piRNA pathway proteins appear to localize to the nuage complex (green), an assembly of proteins situated between the nucleus and cytoplasm. piRNAs corresponding to sense strands (dark blue) are preferentially loaded onto Argonaute3, while piRNAs corresponding to antisense strands (red) are loaded onto Piwi and Aubergine. Once loaded, these Piwi family proteins are thought to mediate cleavage of complementary sequence from both transposons and clusters, silencing transposons and further amplifying both sense-stranded and antisense-stranded piRNA pools in a self-sustaining process which may be maintained across generations if these piRNAs are heritable. During horizontal transfer scenarios or activation of older transposons lacking cluster silencing, complementarity of clusters to transposons must be established, which we speculate might be associated with modifications to Rhino that favor the integration of these transposon transcripts into clusters (?b), thus expanding the repertoire of effective piRNAs.

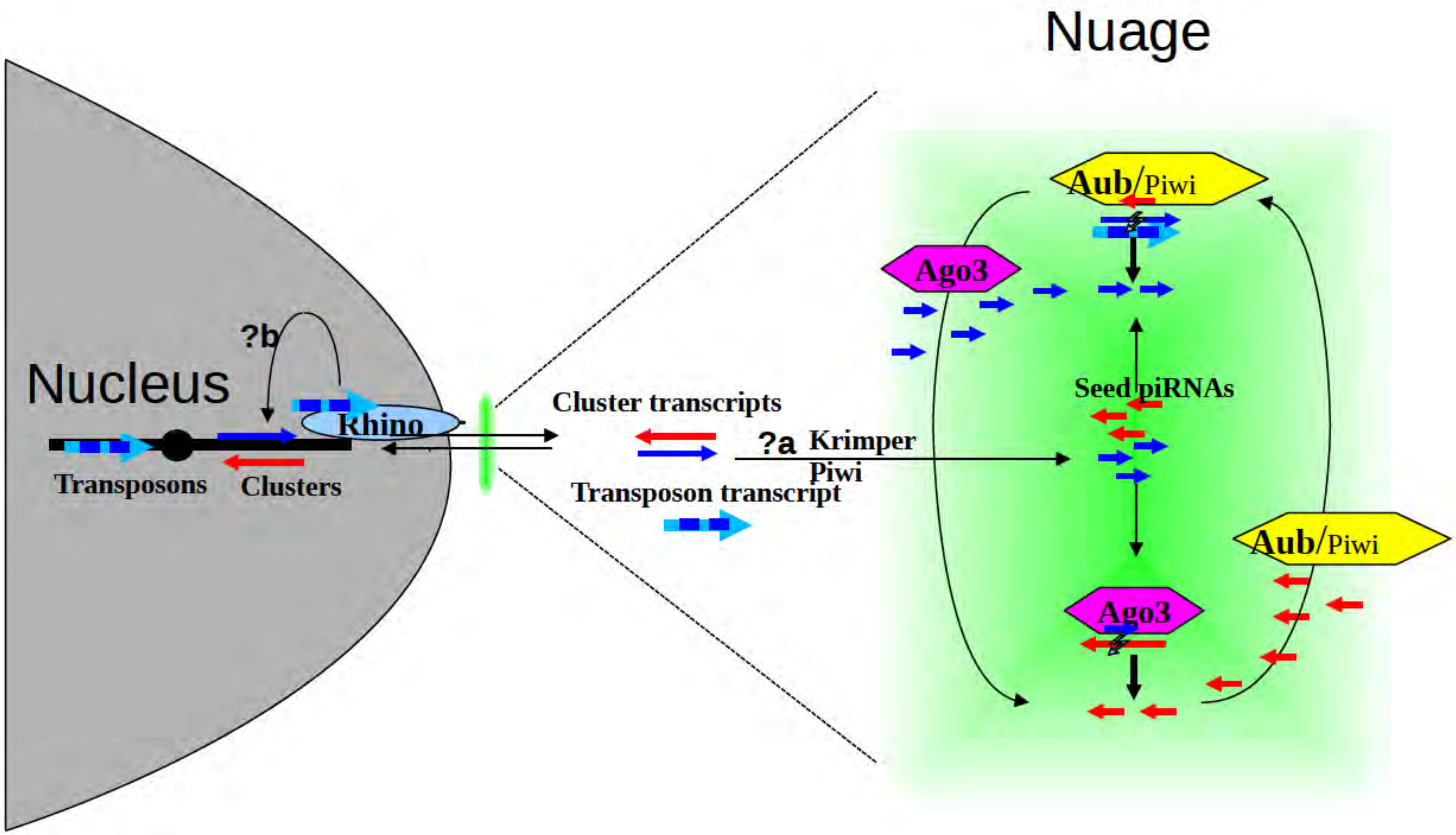


Figure II-2A: Branch model of PAML lineages under positive selection. Because this model cannot assign probabilities to individual lineages under selection, *dN/dS* cut-offs of 0.5 and 1.0 were used as proxies for positive selection or relaxed selective constraint, with an additional filter for dS>0.02. When analyzed in this way, 3/9 lineages in Rhino, 1/9 lineages in Krimper, 2/9 lineages in Aubergine, and 1/9 lineages in Armitage show evidence of recurrent positive selection or relaxed selective constraint (dN/dS >0.5, gray font, dN/dS >1.0, black font). (Abbreviations: Krimper (Kri), Aubergine (Aub), Armitage (Armi), Rhino (Rhi), Spindle-E (SpnE), Zucchini (Zuc)).

Figure II-2B: Branch-sites model of PAML lineages under positive selection. When imposing the criteria of a subset of branches allowing positive selection compared against a background of the remaining branches constrained to be under negative or neutral selection, the result is a group of piRNA proteins that, by definition, are not recurrently positively selected across the entire phylogeny. piRNA protein lineages rejecting neutrality with p-values<0.05 are shown, and are most prominently represented by Aubergine. All lineages rejecting neutrality under these criteria also have *dN/dS* values >1.0. Black represents single selective events, while red represents recurrent selection along an internal branch and all of its descendant lineages.

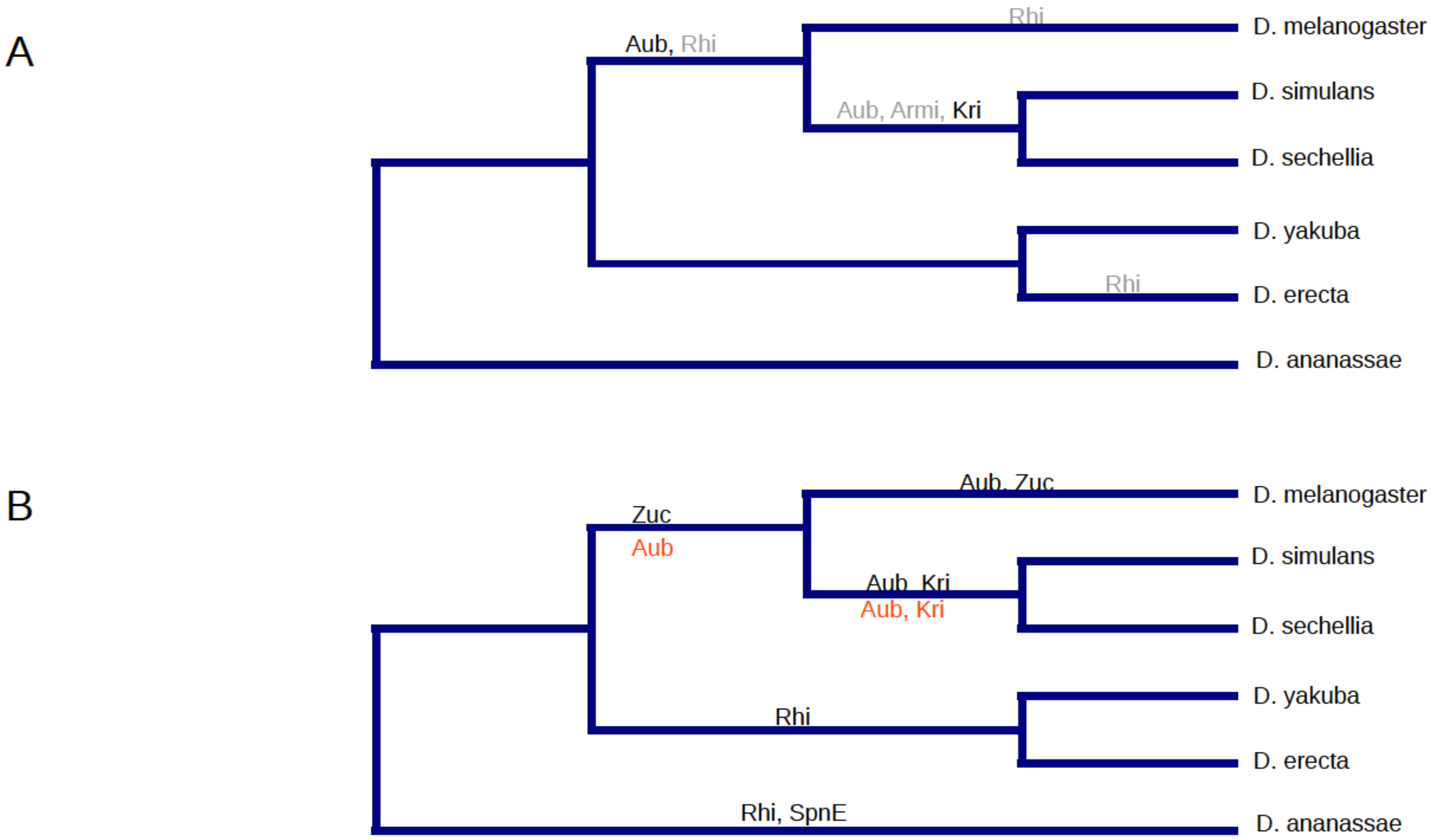


Figure II-S1: Genomic Distribution of dS vs. dN/dS. Although there is a weak negative correlation between dS and dN/dS, piRNA pathway proteins show dN/dS levels elevated beyond the genomic background. Top equation: best fit linear regression (on 73 datapoints) of piRNA pathway proteins (light blue). Bottom equation: best fit linear regression (on 68,218 datapoints) of Flybase Drosophila proteins with 1:1 orthologs (orange). Outlier dS values greater than 0.4 and dN/dS values greater than 2 were excluded to show greater detail.

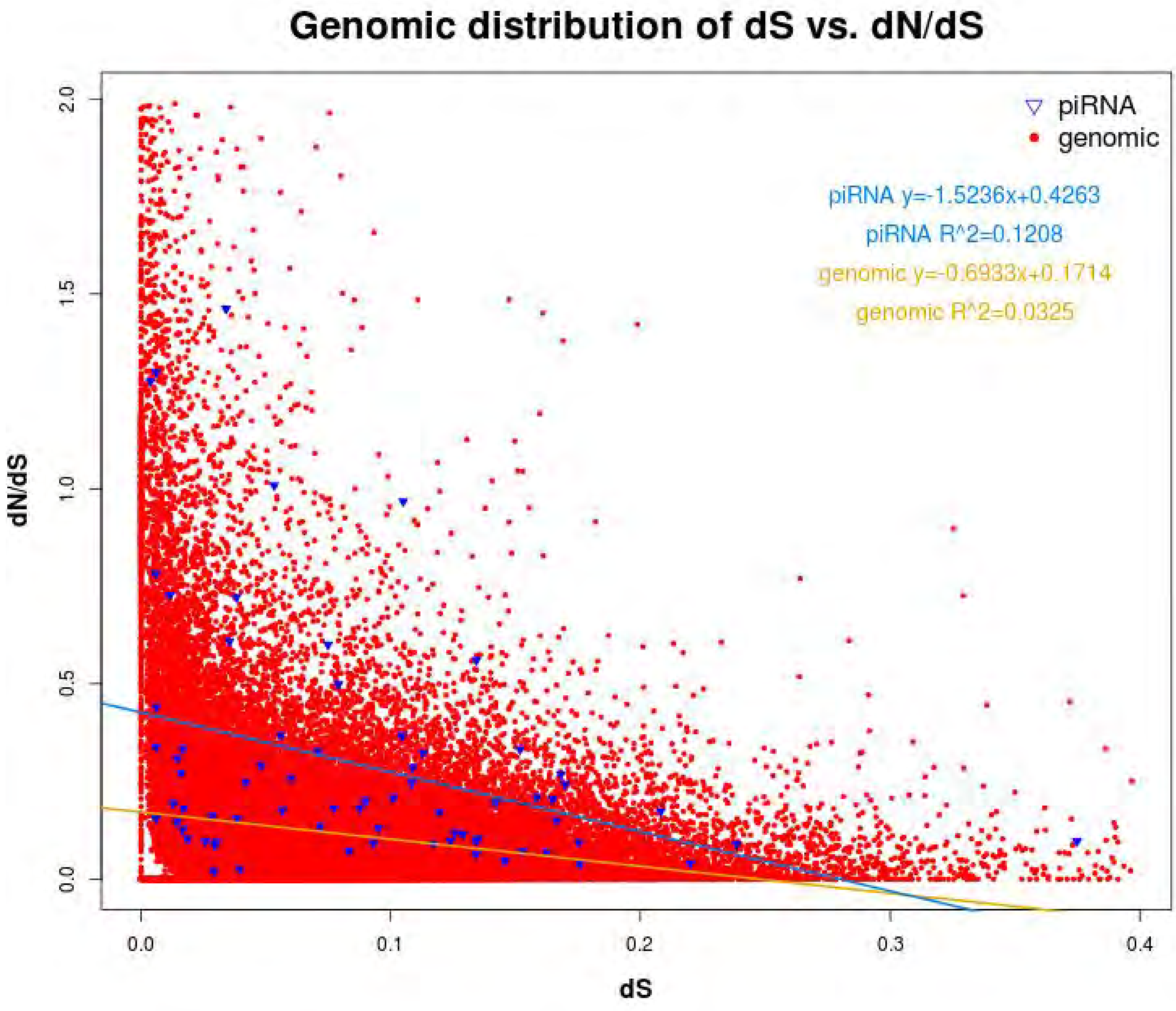


Table II-1: piRNA proteins studied and putative functions. # annotated sp. describes the number of species correctly annotated on flybase (v. 5.29) as protein-coding orthologs.

Name	Annotation symbol	function	# annotated sp.	Citation
Ago3	CG40300	piRNA binding/target cleavage	1	(LI et al. 2009)
Armitage	CG11513	helicase	10	(VAGIN et al. 2006)
Aubergine	CG6137	piRNA binding/target cleavage	10	(BRENNECKE et al. 2007)
Krimper	CG15707	tudor domain/nuage	12	(LIM and KAI 2007)
Piwi	CG6122	piRNA binding/target cleavage	10	(BRENNECKE et al. 2007)
Rhino	CG10683	Chromatin assembly	1	(KLATTENHOFF et al. 2009)
SpnE	CG3158	helicase	12	(VAGIN et al. 2006)
Squash	CG4711	nuclease	12	(PANE, WEHR, and SCHÜPBACH 2007)
Vasa	CG3506	helicase	4	(MALONE et al. 2009)
Zucchini	CG12314	nuclease	12	(PANE, WEHR, and SCHÜPBACH 2007)

Table II-2 Divergence-based analyses. Rhino, Krimper, and Aubergine show the largest number of lineages under positive selection (1st and 2nd columns, out of 15 total pairwise comparisons and 9 total branches, respectively). Some piRNA pathway proteins were predicted to have individual amino acids under significant positive selection across the entire phylogeny[1]. Others showed significant evidence of lineage-specific positive selection (columns 5-6). Bonferroni corrected p-values are shown in parentheses, and comparisons to genomic distributions are shown in square brackets.

Name	Pairwise dN/dS >0.5	Branch model branches >0.5	branch vs. equal	Sites M7 vs. M8[2]	Branch-sites branches[3]	Branch-sites lineage p-values
Ago3	1/15	0/9	P<0.001	P<0.004[0.02]	none	Not significant
Armitage	0/15	1/9[0.07]	P<0.001	Not significant	none	Not significant
Aubergine	1/15	2/9[0.02]	P<0.001	P<0.035[0.06]	mel sim sec all, sim sec, sim sec all, mel[0.009]	0.04(0.47), 0.0008(0.009), 0.0009(0.01), 0.007(0.08)
Krimper	3/15	1/9[0.07]	P<0.001	Not significant	sim sec, sim sec all[0.1]	0.00005(0.0006), 0.0001(0.002)
Piwi	0/15	0/9	P<0.002	P<0.037[0.06]	none	Not significant
Rhino	6/15	3/9[0.009]	P<0.001	P<0.017[0.04]	yak ere, ana[0.1]	0.04(0.49), 0.006(0.07)
SpindleE	0/15	0/9	P<0.001	Not significant	ana	0.05(0.60)
Squash	0/15	0/9	P<0.019	Not significant	none	Not significant
Vasa	0/15	0/9	P<0.001	Not significant	none	Not significant
Zucchini	0/15	0/9	P<0.001	Not significant	mel sim sec, mel[0.1]	0.003(0.03), 0.03(0.36)

[1] Only Piwi and Rhino identified individual amino acids under positive selection. Piwi identified 56L as under positive selection under M8. Rhino identified 46S as under positive selection in M2a and M8.

[2] Two other sites tests were also performed(see methods). For M1a vs. M2a, Rhino was the only significant protein (P<0.048 with 1% of the genomic dataset more significant), while for M8a vs. M8, Rhino and Argonaute3 were both significant (P<0.006 and P<0.022, with 1% and 2% of the genomic dataset more significant, respectively).

[3] Only Krimper and Aubergine localized positive selection to individual amino acids. Krimper analysis found 6 amino acids: 122T, 130S, 144E, 326E, 411S, and 416T, while Aubergine analysis found over 20 amino acids throughout the protein.

Table II-3 Polymorphism-based analyses. Only genes which reject neutrality in favor of positive selection are shown.

gene	LR[4] value	LR p value	GOF[5] score	GOF p -value	α[6]	X[7]
SpnE_NC	18.502	0	14.589	0.139	3913.87	6621
Aubergine_NC	10.145	0	-77.615	0.573	670.20	1865
Vasa_sim	5.478	0.001	99.155	0.234	1817.54	268
Armi_sim	4.563	0.001	93.992	0.731	964.41	2821
Zucchini_sim	1.950	0.007	26.697	0.223	162.90	394
Rhino_NC	4.616	0.022	-93.566	0.719	169.57	302

[4] LR denotes the natural log likelihood ratio of selection vs. Neutrality, as calculated in Kim and Stephan 2002. Because demographic parameters can affect these scores, neutral simulation is performed to assign empirical p-values, and selection is accepted as an alternative to neutrality when p-values are <0.05.

[5] GOF is a measure of the goodness of fit of the data to selection, as calculated in Jensen et al. 2005. Simulation under selection is performed to estimate empirical p-values, and selection is accepted as a viable alternative to demographic processes when p-values are >0.1

[6] α denotes the selection strength, and is given by 2 Ns

[7] X is the maximum likelihood location of the beneficial mutation in nucleotides

Table II-4 McDonald Kreitman tests (only significant genes are shown).

Gene	Fixed nonsyn	Fixed syn	Poly nonsyn	Poly syn	Fisher table p	Fisher marg p	Chi sq p	G test p	G test williams corr. p
Armi_MW	87	70	33	57	0.00545409	0.00188761	0.00455326	0.00436815	0.00450088
Armi_NC	83	70	38	64	0.0103282	0.00296076	0.00776147	0.0075214	0.00770929
Aubergine_MW	101	59	8	20	0.000814674	0.000537145	0.000632438	0.000638369	0.00071739
Aubergine_NC	99	58	11	28	0.0001194	6.78798e-05	8.65505e-05	8.03173e-05	8.97841e-05
Krimper_MW	119	46	57	48	0.00382677	0.00122766	0.00270824	0.00283003	0.00292413
Krimper_NC	123	47	58	55	0.000390862	0.000160545	0.000308591	0.000321512	0.000335574
SpnE_MW	87	95	24	54	0.0136117	0.00421318	0.0109429	0.0100884	0.0103694
SpnE_NC	88	96	30	56	0.0492191	0.0143502	0.0457688	0.044437	0.0451641
Vasa_MW	114	100	15	38	0.00121451	0.000571227	0.0011274	0.000944243	0.00100057

Table II-S1 Fay and Wu's H tests. Significant scores, assessed relative to 1000 standard neutral simulations of otherwise identical parameters (see methods), are in bold.

Population	H value	H p value
Ago3_NC	0.427807	.508
Armi_MW	1.3	.695
Armi_NC	**-5.537815**	**.032**
Armi_sim	-0.5	.381
Aubergine_MW	-1.333333	.156
Aubergine_NC	**-10.689076**	**0**
Aubergine_sim	1	.620
Krimper_MW	**-11.466667**	**.005**
Krimper_NC	-4.352941	.094
Krimper_sim	-0.6	.316
Piwi_MW	**-8**	**0**
Piwi_NC	**-9.305882**	**.002**
Piwi_sim	**-7**	**.009**
Rhino_MW	-2.133333	.130
Rhino_NC	**-3.08254**	**.048**
Rhino_sim	**-3.333333**	**0**
SpnE_MW	1.4	.718
SpnE_NC	**-16.232955**	**0**
SpnE_sim	7	.975
Squash_MW	-0.8	.134
Squash_NC	**-3.729055**	**.008**
Squash_sim	0	1
Vasa_MW	-1.333333	.073
Vasa_NC	-2.015152	.063
Vasa_sim	**-6**	**.015**
Zucchini_MW	-0.8	.128
Zucchini_NC	-1.240642	.094
Zucchini_sim	**-2.333333**	**.039**

Chapter III

Inferring the evolutionary history of primate microRNA binding sites: overcoming motif counting biases

This chapter derives from the article of the same name, and was part of a collaboration between Jeffrey Bailey, who assisted extensively with troubleshooting the scripts and suggested an innovative method for calculating turnover probabilities of individual miRNAs (which will hopefully be correctly implemented in the near future), Fen-Biao Gao, who initially suggested an evolutionary analysis of miRNA binding site turnover rates and provided funding, and Jeffrey D. Jensen, who provided valuable feedback in interpreting results, suggested prioritization of experiments, and provided funding. All authors contributed extensively to manuscript preparation. This work appeared in the journal Molecular Biology and Evolution (2014) 31(7), 1894-1901.

Chapter III Summary

The first microRNAs (miRNAs) were identified as essential, conserved regulators of gene expression, targeting the same genes across nearly all bilaterians. However, there are also prominent examples of conserved miRNAs whose functions appear to have shifted dramatically, sometimes over very brief periods of evolutionary time. In order to determine whether the functions of conserved miRNAs are stable or dynamic over evolutionary time scales, we have here defined the neutral turnover rates of short sequence motifs in predicted primate 3' UTRs.

We find that commonly used approaches to quantify motif turnover rates, which use a presence/absence scoring in extant lineages to infer ancestral states, are inherently biased to infer the accumulation of new motifs, leading to the false inference of continually increasing regulatory complexity over time. Using a maximum likelihood approach to reconstruct individual ancestral

nucleotides, we observe that binding sites of conserved miRNAs in fact have roughly equal numbers of gain and loss events relative to ancestral states, and turnover extremely slowly relative to nearly identical permutations of the same motif. Contrary to case studies showing examples of functional turnover, our systematic study of miRNA binding sites suggests that in primates, the regulatory roles of conserved miRNAs are strongly-conserved. Our revised methodology may be used to quantify the mechanism by which regulatory networks evolve.

Chapter III Introduction

Prominent studies have shown experimentally that some conserved miRNAs can regulate the same targets over deep evolutionary time (Pasquinelli et al. 2000, Moss and Tang 2003, Kucherenko et al. 2012, LaTorre and Giorgi 2013), yet most conserved miRNAs have no experimentally identified conserved target genes and are dispensable in *Caenorhabditis elegans* under laboratory conditions, suggesting that the function of these conserved miRNAs may be changing or hidden (Chen and Rajewsky 2007, Miska et al. 2007). Of the 21-23 nucleotides of a mature miRNA that might potentially basepair to regulate target mRNAs, only the first seven to eight nucleotides are essential (Lewis et al. 2003, Brennecke et al. 2005). This region has come to be known as the “seed” of miRNA target interactions, and has been used as a basis for computationally predicting potential interactions between miRNAs and their target mRNAs. In support of a model of the conserved roles of miRNAs, Farh et al. 2005 have used these computational predictions to show that miRNA binding sites are conserved relative to random sequence within 3’ UTRs. This signature of relative conservation has served as a basis to improve miRNA targeting prediction, implemented through the popular miRNA target predictor TargetScan (Lewis et al. 2003, Lewis et al. 2005). In addition to this work, evolutionary analyses have shown that SNPs in miRNA binding sites are rare relative to other SNP classes, and that miRNAs themselves appear to be deeply conserved as a class (Chen and Rajewsky 2006a, Chen and

Rajewsky 2006b). Together, this literature suggests that miRNAs are conserved regulators of key developmental processes.

Simultaneously, many of these same studies and others also show evidence that the processes miRNAs participate in are changing. Since the first miRNAs were discovered, very few appear to be similarly essential to developmental processes (Miska et al. 2007) or to participate in targeting interactions that are well conserved, even for miRNAs that are themselves extremely well-conserved (Chen and Rajewsky 2007, Gao 2010). In previous studies, which used a comparative approach to examine 3' UTRs derived from human, dog, mouse, rat, and chicken to find the conservation levels of miRNA targets (as defined by miRNA seed complementarity), ninety percent of potential miRNA binding sites are estimated to be non-conserved (Lewis et al. 2005, Farh et al. 2005, Xie et al. 2005, as summarized in Hiard et al. 2010 Figure 2A). Within these non-conserved target sites, many appear to be functional in vitro (Farh et al. 2005), and evolutionary studies have estimated that thirty to fifty percent of nonconserved miRNA binding sites may be functional (Chen and Rajewsky 2006b).

Some of the most compelling evidence that miRNAs may play non-conserved roles comes from a study of the members of a recently speciated clade of cichlids with dramatically divergent behavior, morphology, and diet, but nearly identical genomes (so much so that the vast majority of SNPs within species are also shared between them). Genotyping of these species revealed both a significantly elevated occurrence and divergence of SNPs in predicted miRNA binding sites relative to the surrounding 3’ UTR and the rest of the genome. These results suggest that the modifications underlying speciation in these fish may be driven in part by changes in miRNA regulation (Loh et al. 2010).

Similar restructuring of miRNA regulatory networks has been seen in insects, through an evolutionary and experimental study of the miR10/100 family of miRNAs, which found that the strand from which the mature form of this conserved miRNA hairpin is derived has shifted at least 3 times during insect evolution, completely changing the functional targets of this conserved gene (Griffiths-Jones et al. 2011).

One explanation for the conservation of miRNAs despite shifts in the target genes invokes a model in which miRNAs may improve the robustness of gene regulatory networks by participating in feedback loops that buffer the genome against perturbations (Wu et al. 2009). According to this model, removing miRNAs does not necessarily produce any particular phenotype, but makes phenotypes less stable in environments that fluctuate, a process known as canalization. miR-7 has been demonstrated to participate in this type of role in the neuronal specification of *D. melanogaster*, as flies lacking miR-7 develop abnormally only at fluctuating temperatures (Li et al. 2009). Interestingly, the robustness of these networks has been shown to contribute to the overall evolvability of systems by permitting phenotypes to vary and thus to be exposed to natural selection - allowing miRNAs in these roles to indirectly modulate the pace of evolution (Wu et al. 2009). Indeed, through multigenerational selection experiments, it is shown that miR-9a, another miRNA that ensures the precise neuronal specification in *Drosophila* (Li et al., 2006), dampens the impact of genomic diversity on variability of cell behavior (Cassidy et al., 2013). This model does not predict the pace at which miRNAs change their targets, but only suggests that traits that vary will be less heavily regulated by miRNAs than canalized traits. Under this model, the loss of individual binding sites reduces the overall canalization of pathways, making them less robust and more heritably variable, while the gain of binding sites confers the opposite effect. In this way, strong positive or negative selection may operate on binding sites without any immediately observable phenotypes.

As an alternative explanation for changes in regulatory networks, and the increase in regulatory complexity inferred from reconstructed ancestral states, it has been proposed that regulatory complexity evolved not out of increased selective pressure but due to its absence (Lynch 2007). Under this model, increases in regulatory complexity are compensations for weakly deleterious alleles that can not be purged effectively by purifying selection. This model leads to the hypothesis that changes in miRNA targeting will be expected to be quite frequent, and to lend themselves to the gradual, constant accumulation of new miRNA binding sites, with little selective pressure for the loss of existing interactions.

Quantifying the rate at which miRNA binding sites evolve has important implications for discovering the purpose of miRNA regulatory networks, yet even individual studies such as those described above show contradictory evidence as to whether fast or slow turnover rates predominate in the targeting of mRNAs by miRNAs. Although we know of only one other study that has undertaken a systematic survey of miRNA binding site turnover rates (Xu et al. 2013), several studies have examined the turnover rate of miRNAs themselves by scoring extant species as having or not having particular miRNAs, and extrapolating ancestral states accordingly (Nozawa et al. 2010, Nozawa et al. 2012, Meunier et al. 2013, Xiao et al. 2013). These motif-based ancestral reconstruction studies show a systematic increase in the number of miRNAs over evolutionary time.

We have here applied existing approaches and devised alternative methods to more accurately survey the rates at which primate 3' UTRs gain and lose the binding sites necessary for miRNA targeting. In this way we can differentiate between models proposing that miRNAs are changing their functions rapidly via positive selection, are strongly conserved via purifying selection, or are evolving neutrally via drift.

Chapter III Results and Discussion

Approaches and terminology

Initially, we applied an ancestral reconstruction model which we have termed the 'motif-based' approach. For these purposes, we defined the presence of a motif in a given lineage as an exact match to a given sequence (an arbitrarily chosen eight basepair stretch of nucleotides with no initial reference to function) at a given position in a 3' UTR alignment. Using this definition, we assigned each species as either having or not having the motif of interest, and used these values to infer ancestral states most parsimonious with the observed distribution of the motif in extant lineages. Ancestors inferred not to have a motif whose descendants possessed a motif at the site in question were considered as 'gain' events in the descendants, while ancestors inferred to have a motif whose descendants did not possess the motif were considered to have 'lost' this motif. This process was repeated across all possible eight nucleotide motifs and all positions of all 3' UTRs. Note that in this model, only the presence or absence of a motif is inferred in the ancestor, and not the ancestral sequence itself. Therefore it is possible in these cases for the descendants of a common ancestor to both be inferred as having 'gained' their current states without inferring a pre-existing ancestor (or any losses from this ancestor).

We also implemented a second approach, termed a 'nucleotide-based' ancestral reconstruction, which uses a per-nucleotide ancestral motif reconstruction approach to infer ancestral motif states. Note that unlike the motif-based ancestral reconstruction, with this method ancestral sequences can be queried directly, and every loss of a particular motif by point mutation in an ancestral sequence can be alternatively defined as the gain of the new descendant motif.

The undercounting of losses by the motif based Parsimony approach

The motif-based approach which we initially used to determine the turnover rates of miRNA binding sites was designed in a similar way to previous studies conducted on miRNAs themselves (Nozawa et al. 2010, Nozawa et al. 2012, Meunier et al. 2013, Xiao et al. 2013). Although there are variations in methodology, all of these approaches use the presence or absence of a gene in extant species to infer presence or absence of the gene in ancestral lineages. Our motif-based approach yielded results suggesting that miRNA binding sites, like miRNA genes, are gained much more frequently than they are lost, leading to a net gain of miRNA binding sites over time. However, when this work was repeated with all motifs of length 8, including those not thought to have any function, the bias remained (Table III-1, Figure III-1A). These results were quite unexpected; in the absence of insertions and deletions (which we excluded from analysis), every motif lost from an ancestral sequence should be replaced with a different motif in a descendant created by a point substitution, such that although individual motif types might be selected for or against, gains and losses should be identical overall.

In order to further examine these results, we simulated neutral data whose ancestry could be directly queried, with speciation times based on those estimated for primates. Using our initial approach, we found that in this neutral dataset, the apparent rate of miRNA binding site gain still dramatically outpaced the rate of binding site loss (Figure III-2A). When we used the known ancestors to determine actual turnover rates, it was immediately apparent that our initial method dramatically undercounts loss events (Figure III-2B). A more detailed analysis revealed that the inferred discrepancy between binding site loss events and gain events is due to a saturation of mutations in the underlying sequence, and inherent differences in the effects of this saturation on gain and loss events. Although it has been recognized that independent convergent gains of a trait are inherently less likely than independent losses of a trait (Figure III-3), and there are several techniques that have been used to attempt to account for this bias by counting 'gain' type substitutions as less likely than 'loss' substitutions, there is a second factor that makes a bias toward gains unavoidable. Specifically, when multiple mutations occur within the descendants of a common ancestral motif, the probability of correctly identifying the ancestral motif decreases. In the most extreme circumstances, there are so many mutations that no ancestral motif can be inferred at all, making all subsequent mutations appear to be gain events (Figure III-5). In less extreme circumstances, ancestral conditions are often still misinferred in favor of gains. For a comprehensive case study of the flaws of the motif-based approach, let us examine the primate phylogeny in greater detail. Suppose an ancestral motif that has mutated into a descendant motif in a primate phylogeny is represented by a string of five numbers, where each number in the string is a motif state in one of the five primate species (Human, Chimpanzee, Gorilla, Orangutan, and Gibbon). If we assume that two independent mutation events on an ancestral motif will almost always create two different descendant motifs, and let a 0 represent a motif state shared in some subset of species, a 1 represent a second motif state also present in subset of species, a 2 represent a third type of motif, a 3 represent a fourth type of motif, and a 4 represent a fifth type of motif, we can summarize all possible combinations of species groups sharing motifs in a compact format (see Supplemental File III-4). In each case, a trait-based approach has been used that gives first precedence to finding a minimum number of mutations needed to explain a given binding site pattern, and then attempts to maximize and minimize loss events (to see the extent to which the known bias toward gains relative to losses is correctable). As an example, let us take the pattern 00102. This pattern represents a motif present in one state in Human, Chimpanzee, and Orangutan, (0's) which exists in a second state in Gorilla, (the 1 at position 3), and a third state in Gibbon (the 2 at position 5). When examining the first motif (labeled '0' in 00102) the observed pattern is yes, yes, no, yes, no (which can be abbreviated 11010) in the species Human, Chimp, Gorilla, Orangutan, and Gibbon, respectively, and the minimum number of mutations needed to explain this pattern is two, with two options for how these mutations could have occurred. The first possibility is that the motif is ancestral, and that Gorilla and Gibbon have lost the motif in two independent events. The second possibility is that the motif is derived, and that the ancestor of Human and Chimp gained the motif, followed by an independent gain in Orangutan. The maximum number of losses for this '0' motif is therefore two, with zero gains, while the minimum number of losses for this motif is zero, with two gains.

Moving on to the second motif present at this site ('1' allele present only in Gorilla), we find the following pattern: no, no, yes, no, no (00100). There is only one most parsimonious explanation for this pattern, which is a species-specific gain of the motif in question in gorillas. The same is true for the third motif found only in Gibbon (labeled '2' in the pattern 00102), which appears as a species-specific gain in Gibbon. Therefore, the final gains and losses inferred for the pattern 00102 is a maximum number of losses corresponding to two gains and two losses, and a minimum number of losses of four gains and no losses. By continuing this analysis across all possible single, double, triple, quadruple, and quintuple mutations (Supplemental File III-4), we see that 7/8 single mutations have the potential to accurately count the true number of gains and losses and 1/8 has the potential to overcount or undercount losses, while 20/23 patterns that can be explained by two mutations have the potential to accurately count gains and losses, and 3/10 are guaranteed to undercount losses while either accurately counting or overcounting gains. In total, 5/33 patterns have the potential to overcount losses relative to gains, and 13/33 patterns have the potential to overcount gains relative to losses. When examining all patterns that can be explained by triple mutations, every possible series of 3 species to have mutations within them leads to an undercounting of loss events, and the same is true of all quadruple and quintuple mutations. As stated before, this is because as the number of mutations increases within a short interval of sequence within an aligned set of species, the power to correctly infer ancestral sequences decreases.

In addition to this inability to infer ancestors as the mutation rate rises, there is the already alluded to issue of an inherently higher likelihood of convergent loss events than convergent gain events. This is because convergent loss events in two descendant lineages can occur through any two mutations anywhere within the existing motif, while convergent gain events from the same inherited ancestral sequence require the same mutation to occur twice at the same site (Figure III-3). Because this probability of independent convergent losses is inherently much greater than the probability of independent convergent gains, any method that initially scores a motif only as present or absent and then uses parsimony to infer the smallest number of mutations will mis-infer multiple loss events as a single loss much more often than it mis-infers multiple gain events as a single gain, in a manner that becomes increasingly evident as the sequence divergence between species increases and the likelihood of convergent substitutions within a motif increases accordingly. This correlation of levels of bias with evolutionary divergence can be observed both experimentally (see Table 1) and via simulation (data not shown). Although several existing approaches attempt to correct these types of problems by giving higher probabilities to loss events than gain events, it is impossible to know with a motif-based approach to phylogenies which motifs in the extant species might be ancestral, and which might be derived. (Figure III-5, Supplemental File III-4).

Correctly counting gains and losses

To correct for the bias induced by the motif-based approach, we implement a nucleotide-based model (using the maximum likelihood based program dnaml) to reconstruct ancestral sequence states, and directly measure gain and loss events relative to these ancestral sequences (Felsenstein 2013). By implementing an explicit model of nucleotide mutation likelihoods, the maximum likelihood model partially corrects for the undercounting of loss and gain events due to multiple substitutions at the same site. By reconstructing individual nucleotides rather than only scoring for absence of a particular binding site, this approach is able to explicitly define ancestral motifs, forcing every gain of a new motif to come about through the loss of an ancestral sequence. This approach is also able to count different nucleotide mutations causing the loss of the same motif as independent events, thus avoiding the mis-inference of a single ancestral loss in these cases, and normalizing the undercounting of loss events to be the same as that of gain events. With this modified approach, we revisited our simulated dataset to reconstruct ancestral nucleotides, and observed turnover rates that closely reproduced those found when using the known ancestors (Figure III-2C). Using this approach, gains and losses were correctly inferred to occur in equal proportion (with minor discrepancies due to the misinference of ambiguous nucleotides in some of the ancestral sequences), and when this revised approach was applied to our experimental dataset without considering sites containing ambiguous nucleotides, gains and losses were also inferred in equal proportions (Figure III-1B)

Inferring the evolutionary rate of miRNA binding sites

When short sequences corresponding to miRNA binding sites were empirically ranked relative to other sequences one mutational step away – in order to mirror the underlying evolutionary pressures operating on nucleotide composition as closely as possible (see Methods) –we observed that miRNA binding sites corresponding to a large number of well-conserved miRNAs had both lower loss rates and lower gain rates than any other members of their cohorts of nearly identical sequences. When searching for this pattern across all short sequences, we found that both 7mers and 8mers corresponding to miRNAs have a significantly greater proportion of slow ranking gain and loss rates than that of the overall pool of short sequences. This result was most pronounced in the case of 8mers, in which, out of the set of 93 well-conserved miRNAs, seven had both gain rates and loss rates slower than all sequences one mutational step away, while only 60 out of all 65,536 8mers met this criteria in the full pool (empirical binomial p-value $2.46X10^{-12}$) as shown in Table III-2. When presented graphically, it is immediately apparent that 8mers corresponding to conserved miRNAs undergo gain and loss events much less frequently than the overall pool of 8mers (Figure III-4). We also examined various other metrics of binding site turnover rates (gain and loss rates slower than the median, gain rates slower and loss rates faster than the median, gain rates faster and loss rates slower than the median, gain and loss rates both faster than the median, and gain and loss rates faster than all sequences one mutational step away) and found that binding sites of conserved miRNAs were much more likely than other sequences to be slow evolving and less likely to be fast evolving. Consistent with experimental evidence that 8mers are the biologically relevant determinants of miRNA-mediated gene regulation (Brennecke et al. 2005), and that the 9th nucleotide does not contribute to seed binding, 9mer results were markedly less significant than those of 7mers and 8mers, with 8mers showing the most significant results.

Biological implications

Having found a strong signal of slower overall turnover rates for the binding sites of strongly conserved miRNAs relative to our dataset as a whole, it is notable that although crystal structures only show evidence for interactions between nucleotides 2-8 of mature miRNAs and target mRNAs (Faehnle et al. 2013), binding sites defined as the reverse complement of nucleotides 1-8 had a stronger signal than any other seed examined. These results cannot be explained by any neutral byproduct of the 8th nucleotide at position 1, as adding a nucleotide at position 9 abolishes significance completely.

Previous work has found increased signatures of conservation for mRNAs containing an 'A' opposite the first nucleotide of the mature miRNAs (Brennecke et al. 2005, Lewis et al. 2005), but as there is a strong bias toward 'U' at this position in the mature miRNA, it is unclear whether this signature of conservation is due to base-pairing or some other factor. In order to understand the nature of this interaction, we examined the subset of conserved miRNA 8mer seeds whose first nucleotide is not a 'U'. We found that out of 32 such miRNAs, 21 of the binding sites for these miRNAs had reduced turnover rates when modified to contain an 'A' opposite the first nucleotide relative to the unmodified binding sites. While individual reductions in turnover rate were only modest and non-significant, the binomial probability of 21 reduced turnover rates in 32 trials occurring by chance is 6.16*10^-7. (Table III-S1, Supplemental File III-1). These results lend support to the notion that an 'A' opposite the first nucleotide has a general stabilizing effect on the interaction between miRNAs and mRNA targets through mechanisms other than base-pairing.

Our dataset contains well-conserved miRNAs whose binding sites are turning over faster than most sequences one mutational step away, as well as some whose binding sites exhibit unbalanced gain and loss rates. Strong skews in the turnover rates of particular well-conserved miRNAs may represent potential candidates for miRNAs with newly evolved functions. Some of the sixty 8mers in the overall dataset with slower gain and loss ranks than all other sequences one mutational step away may likewise correspond to the motifs of undiscovered 3' UTR regulators with strongly constrained function.

Contrary to theoretical arguments advocating the neutral expansion of complex regulatory networks, and case studies of newly evolved regulatory circuits that may underlie adaptive changes in miRNA targeting, our empirical results on a large number of well-conserved miRNAs show that in general, the regulatory networks of well-established miRNAs are neither expanding nor contracting within primates, and there is no significant enrichment for miRNA binding sites with rapid turnover. Instead, we find evidence of strong purifying selection against both gains of new sites and loss of existing ones. Taken as a class, the binding sites of well-conserved miRNAs change extremely slowly, suggesting that the evolutionary niche played by miRNAs in primates is not largely driven by positive selection, but by the maintenance of conserved essential biological functions, many of which may be as yet undiscovered. These conserved functions may serve to canalize or modify gene expression levels. Although our study cannot directly address the evolutionary dynamics that play out in other taxa, the accumulation of more experimentally annotated 3' UTRs across closely related species groups should make this possible in the near future.

Conclusions

Our results suggest that the 8th nucleotide at position 1 of miRNA seeds may impart specificity to miRNA targeting, that the binding sites of well-conserved miRNAs are governed by strong purifying selection in primates, and that the functions of well-conserved miRNAs are therefore also likely to be strongly conserved.

Methodologically, our study makes several improvements over previous approaches to the quantification of motif turnover rate dynamics. By examining every nucleotide for maximum likelihood reconstruction rather than applying a binary gain/loss condition, we effectively correct for inherently biased gain/loss ratios of motif based analyses that have been previously interpreted in the literature as indicative of a general accumulation of regulatory complexity. By correcting turnover rates for the number of occurrences of each short sequence, and ranking each sequence relative to its nearest mutational neighbors, our method corrects not only for differences in mononucleotide and dinucleotide composition, but higher order effects out to the full length of the sequence under investigation.

Although we here report overall turnover rates of individual well-conserved miRNAs, our analysis may productively be re-examined to derive the turnover rates of particular species or categories of target genes, or expanded to include the turnover rates of poorly conserved miRNAs. With slight modification, our methods may be extended to transcription factor binding site turnover rates and those of other motifs. We caution against the application of these methodologies to sequences predicted by synteny to widely diverged species, as alignable material decreases rapidly with phylogenetic distance even within primates.

Chapter III Materials and Methods

Datasets

Our experimental dataset was curated from primate genomes using MAF alignments to the gorilla genome (to take advantage of the most current primate genomes) curated on the UCSC genome browser (Kent et al. 2002). We compared these alignments to annotated human genes, and used the alignment program LASTZ (Harris 2007) to filter out human proteins with low coverage or duplicated sequence (see Supplemental Methods in Supplemental File 2). The 3' UTRs of the resulting set of genes were aligned with ambiguous nucleotides inserted to separate discontinuous regions in the MAF alignment, and regions with gaps or ambiguous nucleotides in some species were masked to ambiguous nucleotides in all species.

A second primate dataset was simulated under a neutral model of evolution to evaluate whether underlying biases exist in different methods of counting miRNA binding site turnover events. This dataset was created with the program SFS_CODE (Hernandez 2008, see Supplemental Methods Supplemental File III-2 and Table III-S2 in Supplemental File III-1), and consisted of 30,000 nucleotides of simulated sequence, with speciation events added using primate divergence times estimated by TimeTree (Hedges et al. 2006) to approximate the experimental primate data.

Defining the miRNA 'seeds' and control motifs

We analyzed the turnover rates of all seven basepair, eight basepair, and nine basepair sequences within 3' UTRs (which we have termed 7mers, 8mers, and 9mers respectively, or *k*-mers as a general term for these sequences of length k). In the case of 7mers, we defined a subset of putatively functional miRNA 'seed' binding sites as those 7mers corresponding to the reverse complement of nucleotides 2-8 in the mature miRNA. The reverse complement of nucleotides 1-8 was defined as putatively functional for 8mers, and the reverse complement of nucleotides 1-9 was used for 9mers. In all cases, well-conserved mature miRNA seeds were defined as those annotated in miRbase 18 as a miRNA seed in human, mouse, and zebrafish (Kozomara and Griffiths-Jones 2014).

Inferring number of gain and loss events

Motif-based Parsimony approach

We implemented a parsimony approach modeled on previous motif turnover studies (Nozawa et al. 2010, Nozawa et al. 2012, Meunier et al. 2013, Xiao et al. 2013) in which each location in an aligned 3' UTR having at least one instance of a given short 'seed' sequence – excluding those regions of 3' UTR with gaps in any of the species – was examined to determine which species contained the full length seed sequence at that location (coded as a '1') and which did not ('0'). Subsequently, we fit a most parsimonious interpretation of the ancestral gain and loss substitutions needed to fit the observed presence/absence values to the known species. Locations in the 3' UTR in which multiple types of substitutions led to the same overall level of parsimony were not analyzed. When examining all 3' UTRs in aggregate for a given 7mer, 8mer, or 9mer, each branch in the phylogeny had a cumulative number of inferred gain and loss events. By adding together the gains and losses across all branches, we were able to infer a total number of gain and loss events for a given *k*-mer across the phylogeny. In this way, the overall turnover rates of miRNA binding sites could be compared to the turnover rates of other 7mers, 8mers, and 9mers.

Nucleotide-based Maximum likelihood approach

As a separate approach, we inferred the ancestral states of each nucleotide using the maximum likelihood implementation of dnaml from the phylip phylogeny package (Felsenstein

2013). As before, seed sequences corresponding to miRNA binding sites were analyzed, as well as other sequences of identical length, for 7mers, 8mers, and 9mers. Gain and loss events are assigned directly by comparing the aligned inferred ancestral sequences to the descendants.

Because frequently occurring motifs in 3' UTRs have a higher probability of observing turnover events than rare ones (see Supplemental Figures, Supplemental File III-3), we calculated a normalized turnover rate by dividing the number of gain or loss events of each sequence by the number of total sites observed for that sequence. To control for the effects of nucleotide composition on turnover rate, we compared the normalized turnover rates of every short sequence to those of a cohort of short sequences within one base substitution of the sequence of interest, and assigned each short sequence a gain rank, loss rank, and total number of occurrences rank relative to this cohort.

Chapter III Acknowledgments

We would like to thank Matt Rasmussen, Andrew Grimson, Jaaved Mohammed, Chip Aquadro, Xiaoping Zhu and members of both the University of Massachusetts Medical School (UMMS) and École Polytechnique Fédérale de Lausanne (EPFL) communities for the suggestion of the maximum likelihood model of individual nucleotide reconstruction and other useful feedback. This work was supported by grants from the Swiss National Science Foundation, a European Research Council (ERC) Starting Grant to JDJ, and the National Institutes of Health grant R01NS066586 to FBG.

ADDITIONAL SUPPLEMENTARY INFORMATION

The following additional data are included with the online version of this work, and are reproduced here for convenience. Figures III-S1a and III-S1b show correlations between total miRNA binding sites and number of turnover events inferred (gains in Figure III-S1a and losses in Figure III-S1b), necessitating the measurement of turnover rate as a proportion of total miRNA binding sites. Table III-S1 is an approximation of the increased significance associated with an 'A' opposite nucleotide 1 of conserved miRNAs that do not begin with a 'U' relative to the reverse complement of the first nucleotide. Table III-S2 describes the derivation of parameters used for SFS_CODE. Supplemental Methods III describes the process used for 3' UTR calling, alignment, and filtering in primate genomes, using the program LASTZ, as well as the process used to generate the parameters for SFS_CODE simulations of neutral DNA. All scripts used to generate the data can be found online at https://github.com/alfredsimkin/AMTA.

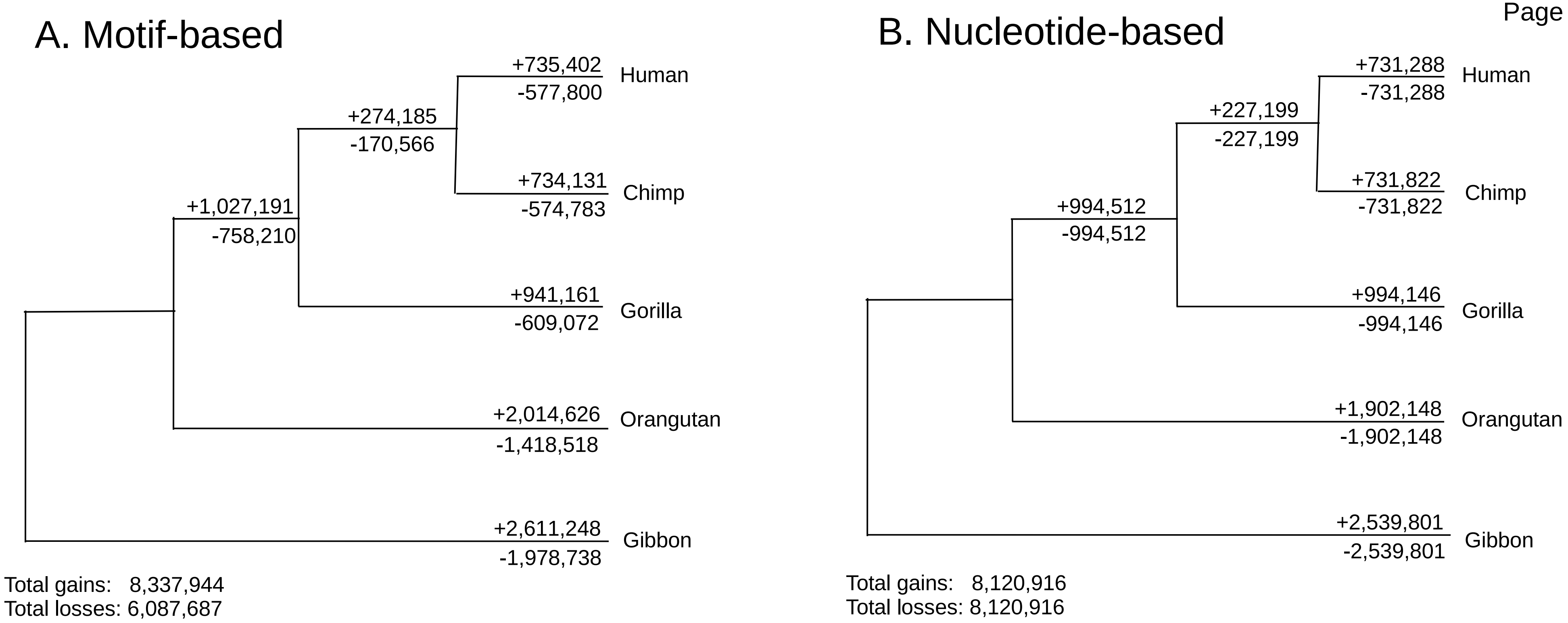


Figure III-1: Two methods for inferring turnover rates in syntenically defined primate 3' UTRs. In Figure III-1A, the ancestral states of 3' UTRs (containing or not containing a binding site) are inferred using presence/absence calls in extant species at the level of motifs, and gains and losses are in turn inferred from the inferred ancestors. Losses appear to occur much less frequently than gains. In Figure III-1B, a revised maximum likelihood approach using individual reconstructed ancestral nucleotides was applied to primate 3' UTRs. The skew toward gains is completely ameliorated.

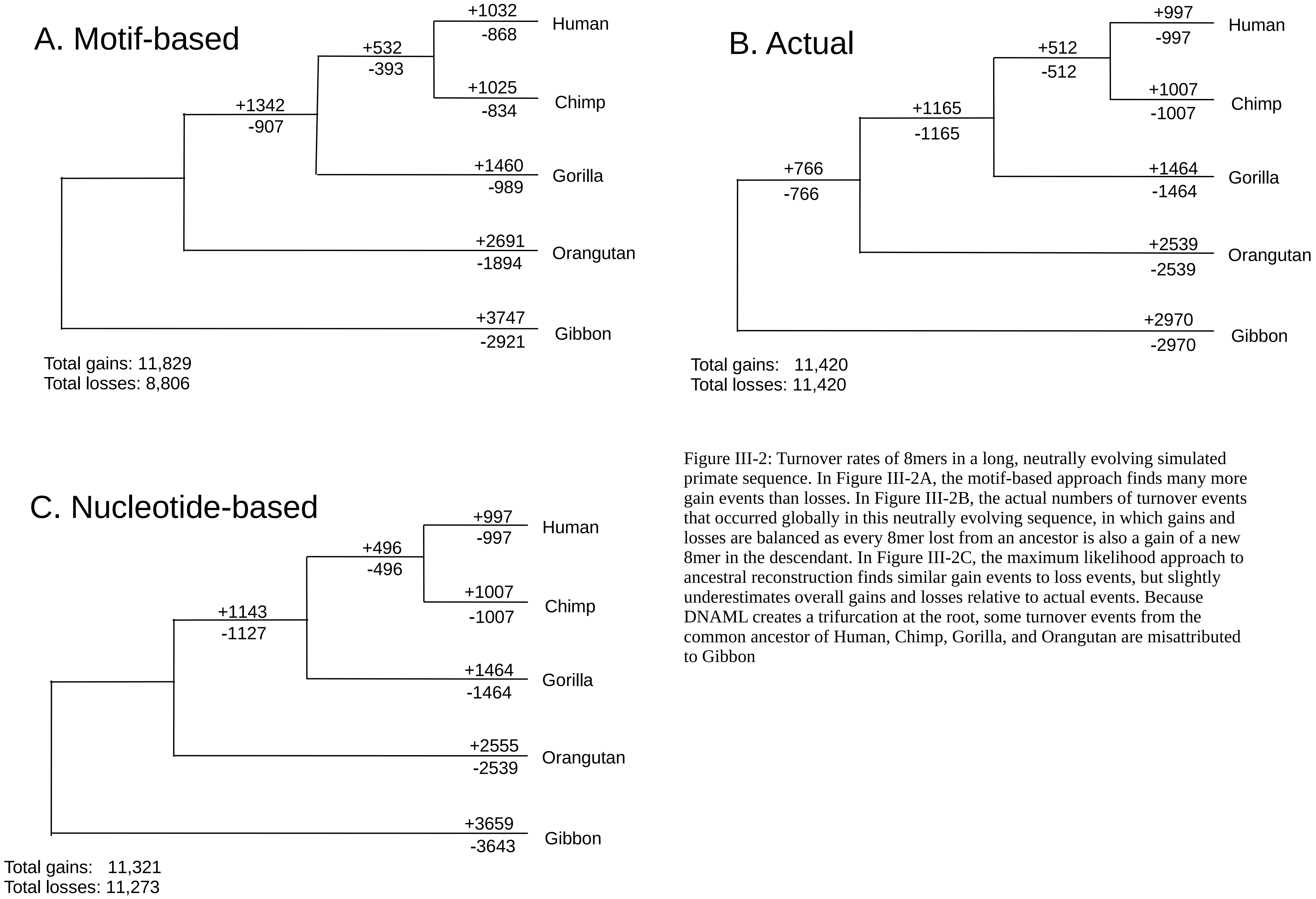


Figure III-2: Turnover rates of 8mers in a long, neutrally evolving simulated primate sequence. In Figure III-2A, the motif-based approach finds many more gain events than losses. In Figure III-2B, the actual numbers of turnover events that occurred globally in this neutrally evolving sequence, in which gains and losses are balanced as every 8mer lost from an ancestor is also a gain of a new 8mer in the descendant. In Figure III-2C, the maximum likelihood approach to ancestral reconstruction finds similar gain events to loss events, but slightly underestimates overall gains and losses relative to actual events. Because DNAML creates a trifurcation at the root, some turnover events from the common ancestor of Human, Chimp, Gorilla, and Orangutan are misattributed to Gibbon

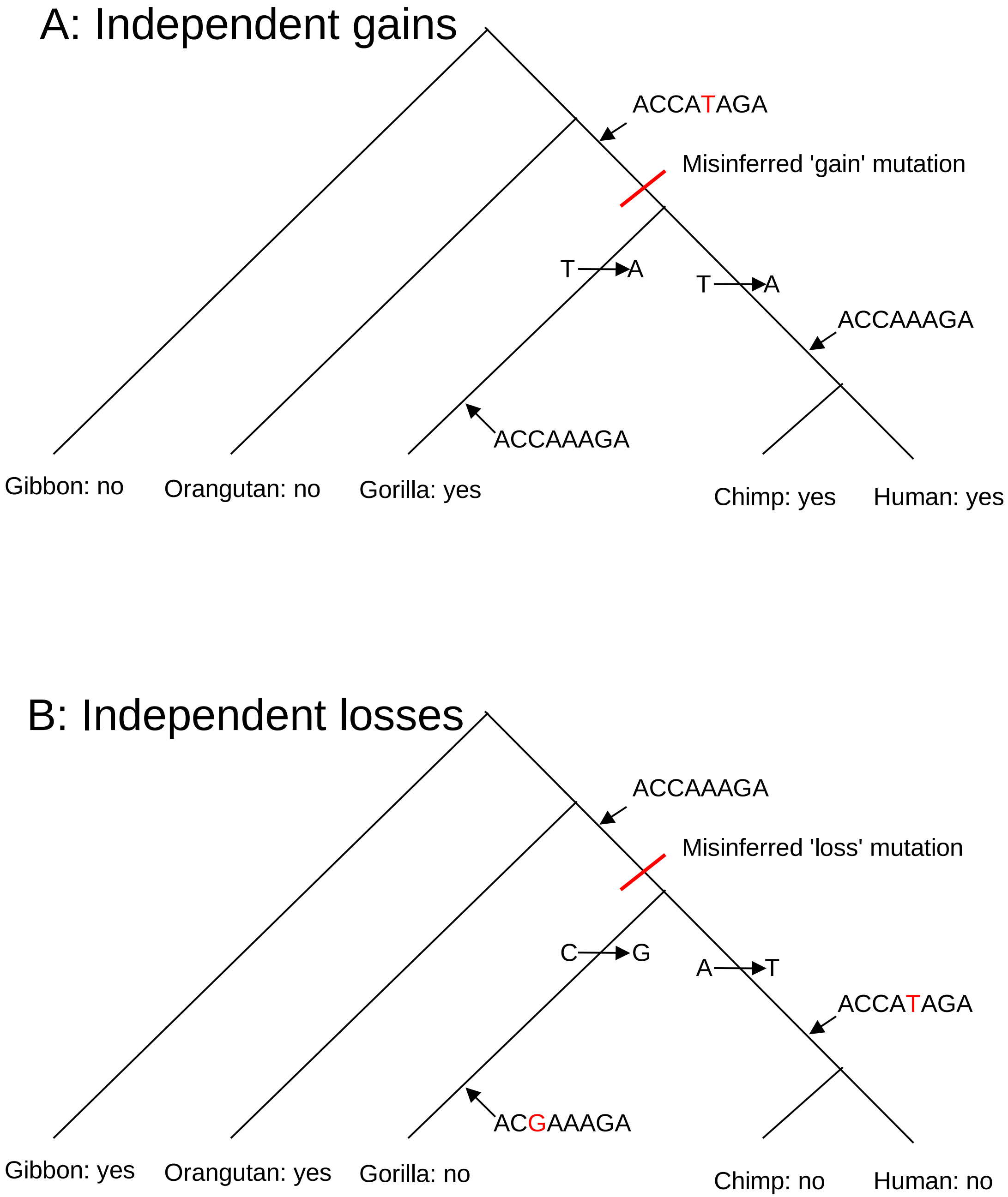


Figure III-3: Independent gains are less likely than independent losses. While parsimony approaches may misinfer independent gain substitutions in multiple descendants as single events (A) and independent loss substitutions as single events (B), independent gains in an ancestral sequence one step away from being a binding site can only occur by one type of substitution (represented by two independent T-->A events at position 5), while independent loss substitutions can occur by any event anywhere within an existing site (here represented by C-->G at position 3 and A-->T at position 5). Independent, misinferred loss substitutions are therefore much more likely to occur than independent, misinferred gains.

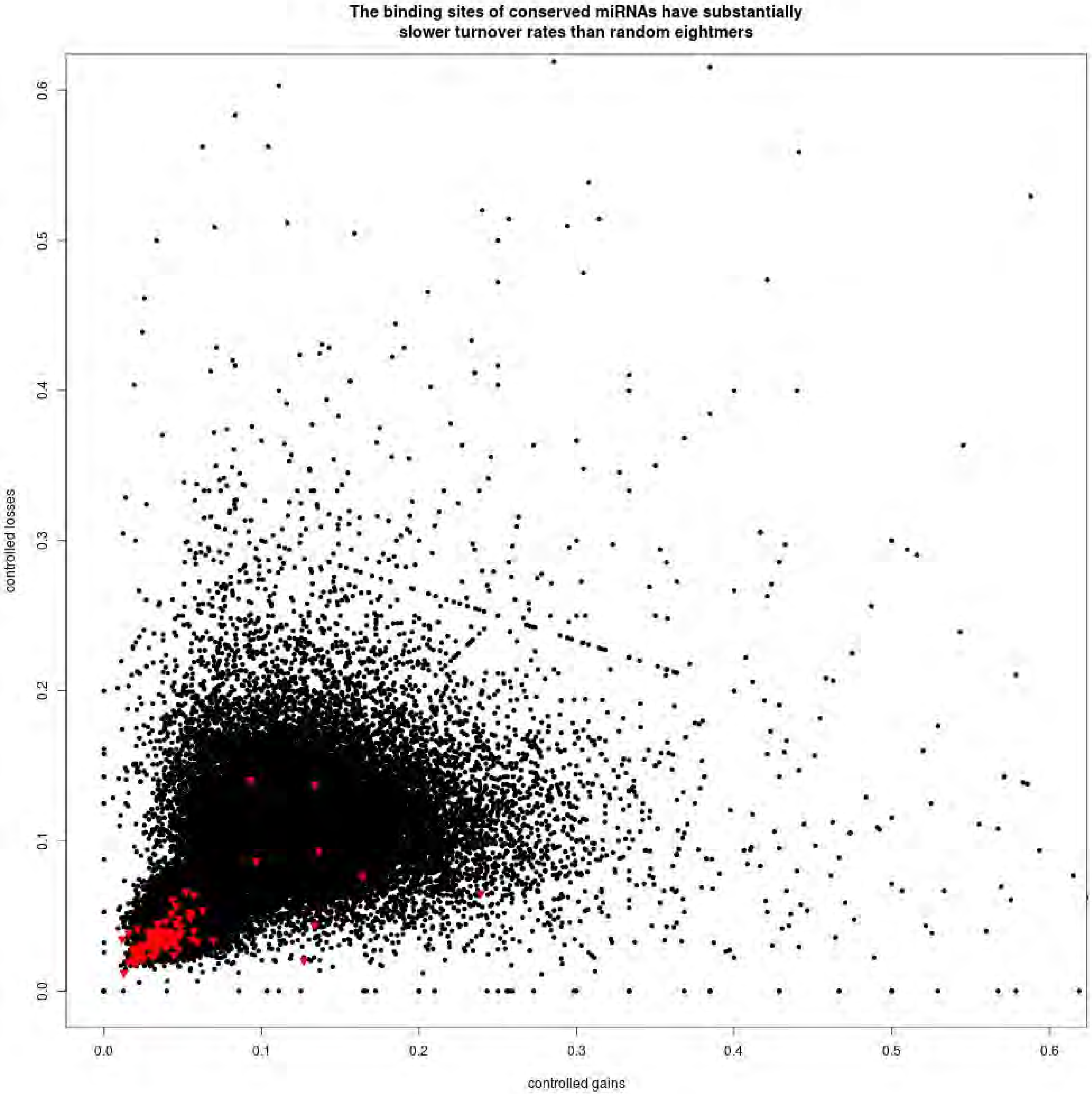


Figure III-4: The binding sites of conserved miRNAs have substatially slower turnover rates than random 8mers.

After normalizing gain and loss rates by total number of sites, it is apparent that the binding sites of conserved miRNAs (in red) have slower turnover rates than those of other 8mer sequences in 3' UTRs

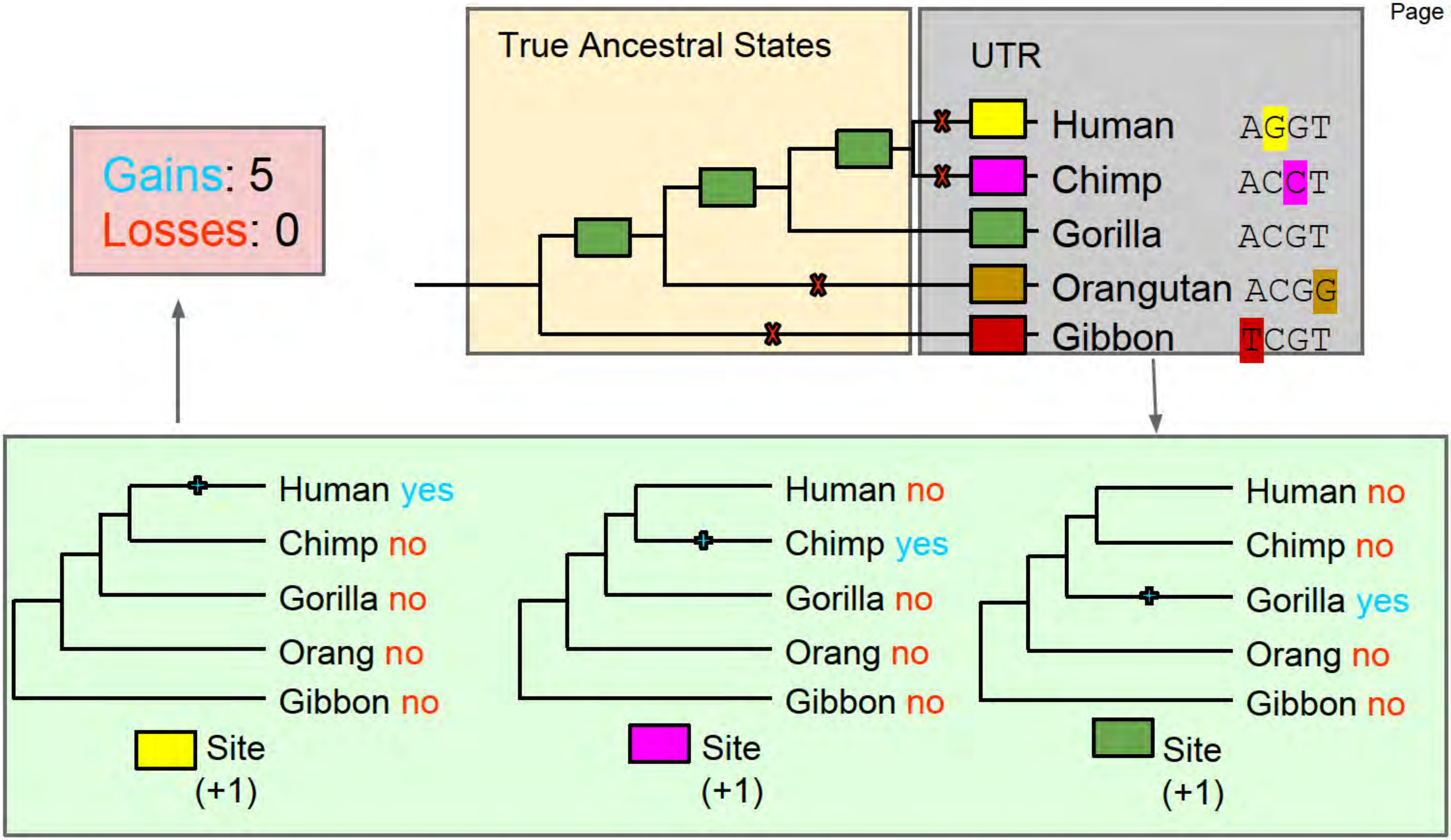


Figure III-5: In extreme circumstances, no ancestors are inferred. In this schematic, the motifs corresponding to five different 4mers, AGGT, ACCT, ACGT, ACGG, and TCGT, are examined within a 3' UTR (gray box). Each motif is summarized by a colored rectangle. When each of these motifs is examined independently, a species specific gain is inferred (green box, shown for three of the five species). When analyzed in aggregate, five species specific gains are inferred and no losses are inferred (pink box). In reality, four species specific losses of the green motif (ACGT) have occurred, as well as four species specific gains of each new motif, while one species (gorilla) maintains the ancestral state, but in the absence of nucleotide-resolution data, this ancestral state is never reconstructed (in fact no ancestral state is reconstructed), such that all existing traits are falsely inferred as newly arising. When analyzed on a global level, this inability to infer ancestral states leads to a loss of power to detect lineage-specific loss events, leading to a skew toward gains.

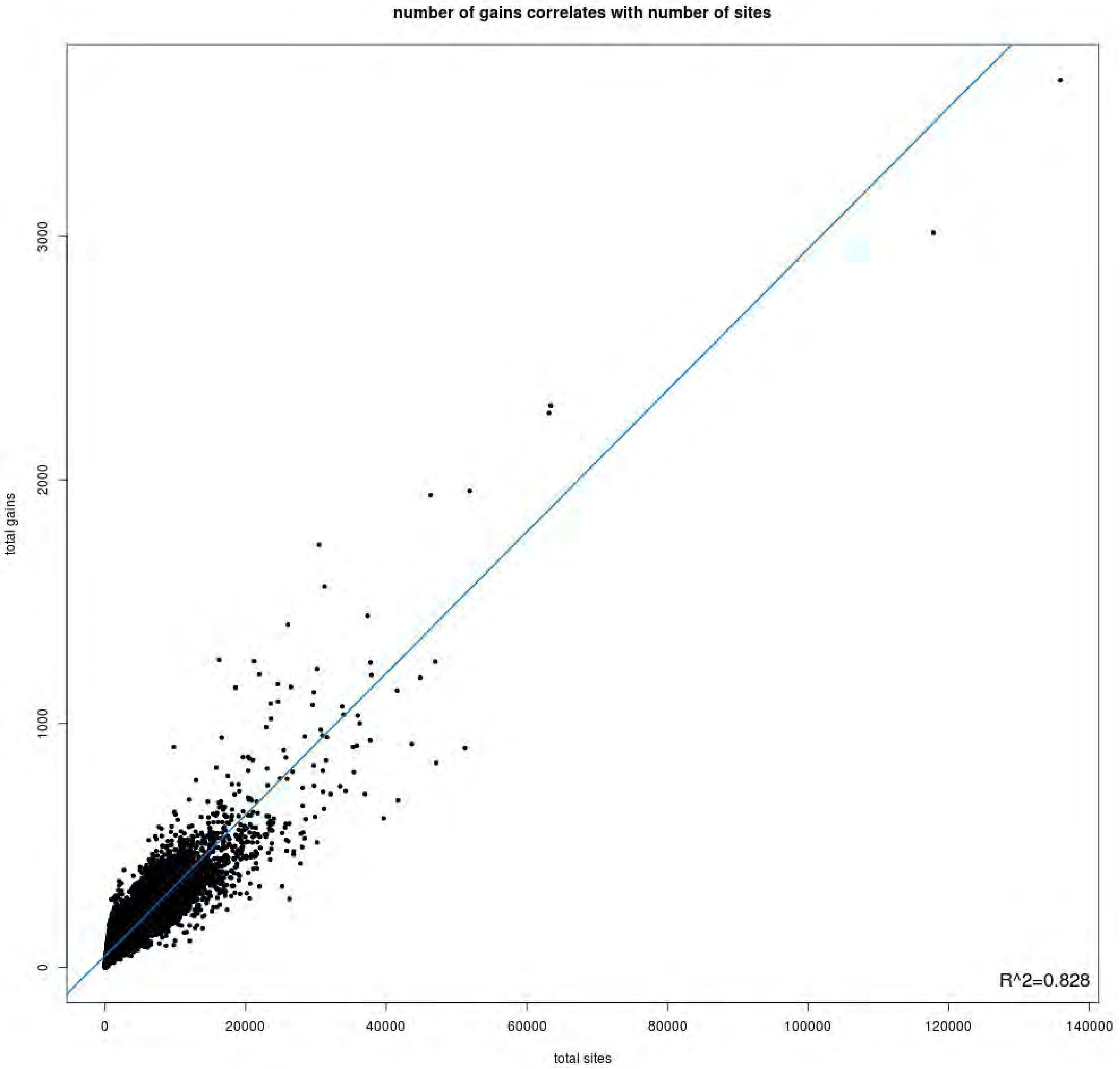
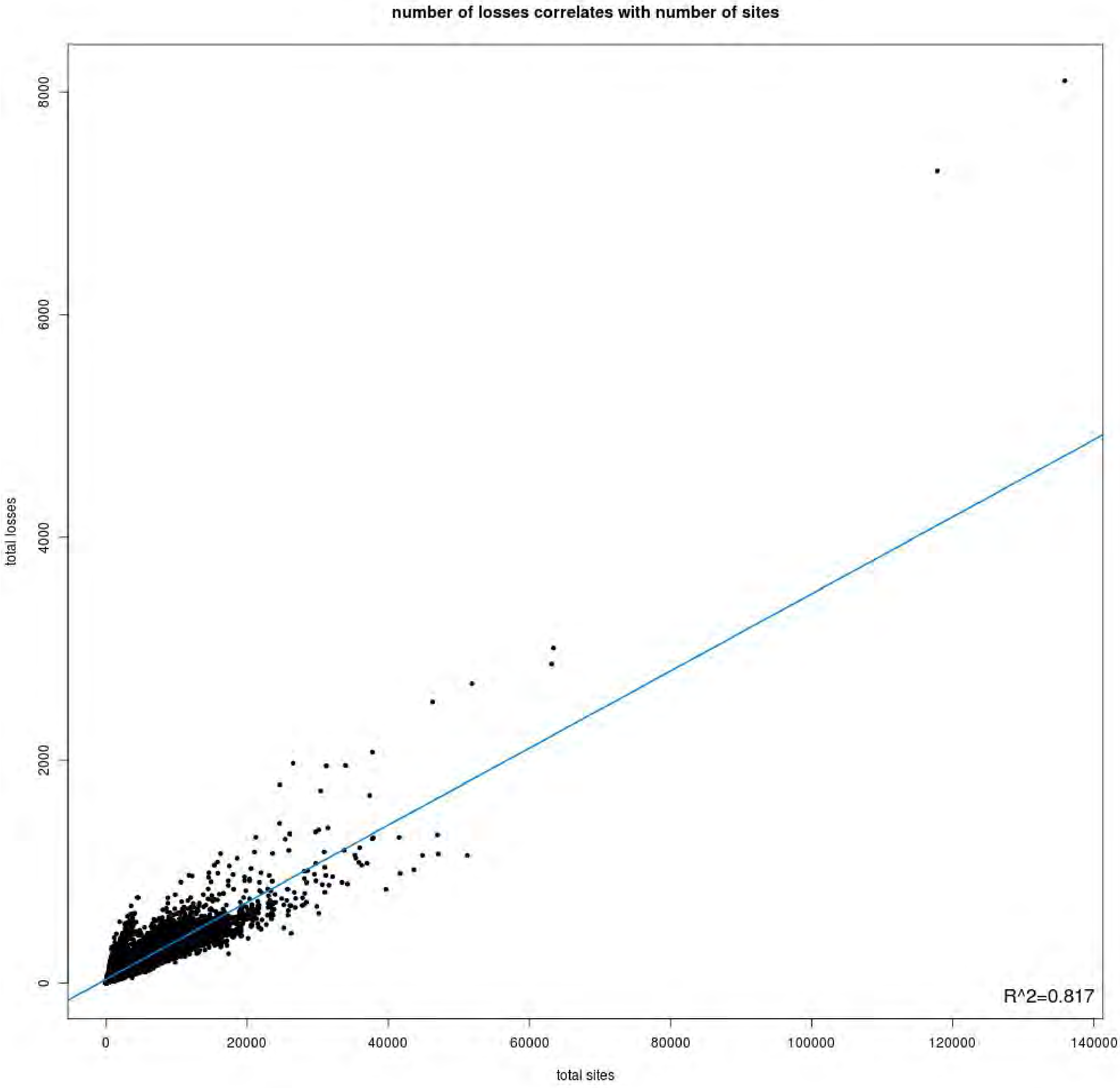


Figure III-S1A: Inferred gain rates correlate with total number of binding sites

8mers that occur frequently have a higher likelihood that one or more 8mers will have a gain event occur. It is therefore necessary to normalize turnover rates by dividing the number of inferred gains by number of total observed sites



Supplemental Figure III-S1B: Inferred loss rates correlate with total number of binding sites

8mers that occur frequently have a higher likelihood that one or more 8mers will have a loss event occur. It is therefore necessary to normalize turnover rates by dividing the number of inferred losses by number of total observed sites

Table III-1: Parsimony Approaches Scoring motifs as Present/Absent Undercount Losses

Dataset	Species surveyed	gains	losses
Meunier et al. 2013[1]	Human, Macaque, Mouse, Oppossum, Platypus, Chicken	719	140
Nozawa et al. 2010[1]	(*Drosophila*) *melanogaster, simulans, sechellia, yakuba, erecta, ananassae, pseudoobscura, persimilis, willistoni, mojavensis, virilis, grimshawi*	101	48
Nozawa et al. 2012[1]	*Arabidopsis*, Papaya, Poplar, *Medicago*, Soybean, Grape, Rice, *Sorghum*, Maize, Moss, Green Algae	743	77
Xiao et al. 2013[1]	*Oryza Sativa*, *Phoenix Dactylifera*, *Populus Trichocarpa*, *Malus domestica*, *Glycine max*, *Solanum lycopersicum*, *Citrus sinensis*, *Arabidopsis thaliana*	167	4
current study, experimental primate parsimony[2]	Human, Chimp, Gorilla, Orangutan, Gibbon	8337944	6087687
current study, simulated neutral primate parsimony[3]	Human, Chimp, Gorilla, Orangutan, Gibbon	11829	8806

[1] Earlier studies examining the turnover rates of miRNA genes find many more gain events than losses.

[2] When using these methods, more gains than losses are inferred for the turnover rates of miRNA binding sites in the 3' UTRs of primates

[3] Simulated neutral datasets (which have been simulated to have identical gain and loss rates in reality) also infer more gains than losses.

Table III-2: Enrichment Levels of miRNA Binding Sites Having a Given Turnover Rank

	7mer[4]	**8mer**[4]	**9mer**[4]
Total conserved miRNAs	94	93	92
Total short sequences	16384	65536	262144
gain rank slowest and loss rank slowest	4/94 vs. 24/16384 (P=1.26E-005*)	7/93 vs. 60/65536 (P=4.768E-012*)	0/92 vs. 388/262144 (P=1)
gain rank<median rank and loss rank<median rank	60/94 vs. 4854/16384 (P=7.47E-012*)	54/93 vs. 15736/65536 (P=2.459E-012*)	45/92 vs. 63850/262144 (P=2.98E-007*)
gain rank<median rank and loss rank>median rank	3/94 vs. 3291/16384 (P=0.9999997889)	5/93 vs. 13315/65536 (P=0.9999900675)	21/92 vs. 64558/262144 (P=0.6937253547)
gain rank>median rank and loss rank<median rank	15/94 vs. 3427/16384 (P=0.9083059651)	14/93 vs. 13396/65536 (P=0.9261940449)	18/92 vs. 64332/262144 (P=0.893040712)
gain rank>median rank and loss rank>median rank	12/94 vs. 4812/16384 (P=0.9999594283)	5/93 vs. 16118/65536 (P=0.9999998491)	8/92 vs. 69404/262144 (P=0.9999954775)
gain rank fastest and loss rank fastest	0/94 vs. 16/16384 (P=1)	0/93 vs. 35/65536 (P=1)	0/92 vs. 85/262144 (P=1)

[4] For every seed length, conserved miRNAs were defined as those whose seed exists as a miRNA in human, mouse, and zebrafish. This resulted in 94, 93, and 92 'real' miRNAs for 7mer, 8mer, and 9mer seeds, respectively. For each seed length, every seed's turnover rates were ranked relative to sequences one mutational step away, with a rank of 1 corresponding to the slowest turnover rates, and the highest rank corresponding to the fastest turnover rates.

[*] Significant P-values are marked with an asterisk. P-values were calculated using binomial distributions on the empirical data. Low probabilities near 0 indicate that having the same or more events occur by chance in a random draw from the dataset is extremely unlikely, while high probabilities near 1 indicate that observing this number of events or more in a random draw from the dataset is virtually guaranteed.

original 8mer	sites rank	gain rank	loss rank		modified 8mer	sites rank	gain rank	loss rank	rank sum before modification	probability of improved rank by chance	rank sum improved?
AACCACTG	8	3	10		AACCACTA	11	9	1	13	0.1056	yes
AACGGTTT	13	22	8		AACGGTTA	10	12	19	30	0.6304	no
AACTGGAC	14	7	6		AACTGGAA	5	6	1	13	0.1056	yes
AAGGGCTT	11	6	2		AAGGGCTA	14	2	8	8	0.0336	no
AATGTGAT	6	10	2		AATGTGAA	2	1	5	12	0.088	yes
ACACTGGG	8	6	18		ACACTGGA	5	3	5	24	0.4048	yes
ACGCACAG	10	10	20		ACGCACAA	12	7	9	30	0.6304	yes
ACTGAAAG	6	9	11		ACTGAAAA	6	19	3	20	0.2736	no
ACTGCAGT	10	21	3		ACTGCAGA	13	16	9	24	0.4048	no
ACTTTATG	9	5	9		ACTTTATA	8	3	18	14	0.1248	no
AGCACTTT	1	1	11		AGCACTTA	5	6	10	12	0.088	no
ATGCTGCT	3	3	6		ATGCTGCA	7	3	2	9	0.0448	yes
ATGTAGCT	14	15	9		ATGTAGCA	6	2	20	24	0.4048	yes
CACCAGCT	7	11	4		CACCAGCA	5	3	2	15	0.1456	yes
CACTGCCT	4	13	4		CACTGCCA	6	8	5	17	0.192	yes
CATTTCAC	8	11	5		CATTTCAA	8	6	12	16	0.168	no
CCTGCTGT	6	17	14		CCTGCTGA	7	4	1	31	0.664	yes
CGAATTTG	11	17	25		CGAATTTA	12	7	14	42	0.928	yes
GCAAAAAG	12	10	11		GCAAAAAA	9	7	6	21	0.304	yes
GCAAAACT	9	13	4		GCAAAACA	10	15	14	17	0.192	no
GCACTTTG	1	1	9		GCACTTTA	5	2	2	10	0.0576	yes
GGCAGCTT	6	1	13		GGCAGCTA	10	3	1	14	0.1248	yes
GGCCAGTT	14	13	16		GGCCAGTA	11	2	5	29	0.5952	yes
GTGCAATG	7	16	12		GTGCAATA	4	1	1	28	0.5584	yes
GTGCAATT	15	2	7		GTGCAATA	4	1	1	9	0.0448	yes
GTGTCATT	10	12	16		GTGTCATA	17	18	4	28	0.5584	yes
TAATAATG	9	16	5		TAATAATA	4	4	14	21	0.304	yes
TACGGGTT	22	14	1		TACGGGTA	17	7	8	15	0.1456	no
TAGGTCAT	14	17	11		TAGGTCAA	19	6	12	28	0.5584	yes
TCGATGGT	12	19	2		TCGATGGA	22	16	9	21	0.304	no
TGAATGTT	5	3	3		TGAATGTA	9	4	3	6	0.016	no
TTGCACTG	12	8	5		TTGCACTA	10	4	1	13	0.1056	yes
											21
median prob of reduced rank				0.192							
binomial approx. significance				1.2459E-008							
exact probability of 21 or more eightmers having reduced rank				6.16E-007							
Table III-S1: significance associated with replacing non-A binding sites with 'A' binding sites											

branch number	common name	Species 1	Species 2	Divergence time (mya)	thousands of generations	thousands of generations since ancestor (TS)	derivatives	final parameters
1	gibbon			20.4	1360	0	none	
2	human-orang anc	Human	Orangutan	15.7	1046.6666666667	313.3333333333	3 & 4	-TS 313.333 2 3 -TE 313.333 2 -TS 313.333 3 4
3	orangutan			15.7	1046.6666666667	313.3333333333	none	
4	human-gorilla anc	Human	Gorilla	8.8	586.6666666667	773.3333333333	5 & 6	-TS 773.333 4 5 -TE 773.333 4 -TS 773.333 5 6
5	gorilla			8.8	586.6666666667	773.3333333333	none	
6	human-chimp anc	Human	Chimp	6.3	420	940	7 & 8	-TS 940 6 7 -TE 940 6 -TS 940 7 8
7	human			6.3	420	940	none	
8	chimp			6.3	420	940	none	
Table III-S2: SFS parameters								

Supplemental Methods III

lastz

In order to generate 3' UTR alignments across primate genomes, the following procedure was used:

1. The full list of human protein coding genes (spliced mRNA complete with 5' UTR and 3' UTR) was downloaded from UCSC genome browser
2. This list was filtered to remove genes with no 3' UTR, genes with multi-exon 3' UTRs, and genes with the same identifier assigned to multiple locations in the genome.
3. 5 primate genomes (hg19, gorGor3, panTro3, nomLeu1, and ponAbe2) were downloaded from the UCSC genome browser
4. Lastz was run with each primate genome (including humans) as the query, the set of filtered human genes as the subject, a match score of 1, mismatch score of 3, and a step size of 200.
5. Orthologs across primate species were called using the following criteria:
    1. Lastz hits with less than 95% identity to any human gene were discarded.
    2. Lastz hits to a single human gene were sorted with respect to where they fell on the human gene and with respect to where they fell relative to each other, and those which fell in a linear sequence (with respect to both the human gene and the query genome) were grouped together as putative transcripts. At this stage, many individual human genes had multiple putative transcripts, which were sorted by percentage of the human gene covered by the transcript.
    3. Cases in which the best 'transcript' covered less than 80% of the human gene were discarded.
    4. Of the human genes with a best hit transcript covering more than 80% of the human gene, cases in which a second best hit transcript was also present covering more than 20% of the best hit transcript were also discarded as putatively duplicated genes.
    5. Of the remaining genes, cases in which more than 20% of the nucleotides covered the human gene redundantly were also discarded. to avoid genes which have undergone excessive internal duplications.
    6. Of the genes remaining at this stage, those in which any of the putative transcript covered an annotated human 3' UTR more than once were discarded, to ensure proper alignment of 3' UTRs.
6. After calling orthologs using the above procedure within individual primate genomes, the lists of orthologs were cross referenced, and genes with high confidence nonduplicated orthologs present in all primate species were kept.
7. Using the 11way primate multiz alignment maf file for the gorilla genome, maf alignments with overlap to 3' UTRs with non-redundant genome coverage in any species were truncated to match the boundaries of 3' UTRs and assembled into complete 3' UTR alignments (with an N inserted in cases involving 3' UTRs composed of multiple maf alignments).

sfs_code

The parameters used for sfs code were based on the following estimates and assumptions:

- 15 year generation times for all primates
- A scaled theta value of 0.00002 to reproduce observed levels of sequence divergence between human and chimpanzee
- Divergence times (and therefore split times) based on timetree
- A simulated UTR space of 30,000 nucleotides

These values were used to produce parameters for sfs_code using table III-S2. Using estimates from this table, the final sfs_code command was as follows:

```
sfs_code 9 1 -t 0.00002 -L 1 30000 -TS 0 0 1 -TE 0 0 -TS 0 1 2 -TS 313.333 2 3 -TE 313.333 2 -TS 313.333 3 4 -TS 773.333 4 5 -TE 773.333 4 -TS 773.333 5 6 -TS 940 6 7 -TE 940 6 -TS 940 7 8 -TE 1360
```

sfs output was parsed into fasta format, assuming that mutations present in greater than half of any population would have been sampled if a reference genome of that species had been produced.

probability calculations for non-canonical miRNAs:

While most miRNAs begin with a 'U' at nucleotide one of the mature miRNA, there are 32 well-conserved miRNAs from our study that do not follow this tendency. We used our dataset to evaluate whether in these cases an 'A' is preferred opposite the first nucleotide, or whether the reverse complement of this first nucleotide is preferred, and used turnover ranks as a readout for reverse complement vs. 'A' preference. We used the following steps to test this theory:

(1) We summed the gain rank and loss rank associated with each miRNA binding site to assign an overall 'turnover rank' to each non-canonical miRNA binding site.
(2) To assign a probability of a reduced turnover rank occurring by chance, we considered that an eightmer ending in 'A' could have one of 25 gain ranks and one of 25 loss ranks, for 25^2 possible rank combinations, and enumerated all such combinations that would lead to an improved overall turnover rank relative to the turnover rank of the unmodified eightmer. This gave each of our 32 miRNAs its own probability of a reduced turnover rank occurring by chance.
(3) We counted the number of miRNAs that had a reduced turnover rank after replacing the reverse complement of the first nucleotide with an 'A' and observed that 21 out of 32 miRNA binding sites exhibited reduced turnover ranks when this substitution was made, suggesting that the reverse complement of nucleotide '1' may not be as strongly constrained as an 'A' opposite this nucleotide.
(4) To assign significance, we modeled a reduced turnover rank as a success (1), and a non-reduced turnover rank as a failure (0). Each possible outcome could therefore be converted into a string. For example, successes at the 6th and 22nd input eightmers could be represented as follows:

00000100000000000000010000000000

We next matched the position of each '0' or '1' to its corresponding probability of success (or the additive inverse of the success probability for a failure) from the list of probabilities calculated in (2) and took the product of these probabilities for the cumulative probability of a particular combination of successes and failures.

By repeating this process over all strings with more than 21 1's, and summing the resulting probabilities, we obtained a probability of 21 or more 'successes' occurring by chance. This process can be viewed as a series of binomial events, where each binomial event has a different underlying probability. In practice, we used cumulative probabilities superimposed on taxicab geometry with a dynamic programming approach (whose conceptual framework was suggested by Xiaoping Zhu) to dramatically reduce the computation of these probabilities (see the script ‘multiple_binomial_cumulative_dist.py’ for our implementation)

In order to comprehensively examine scenarios under which trait-based approaches are misled into misinferring ancestral events or counting an excess of gains relative to losses, I here examine the primate phylogeny of human, chimpanzee, gorilla, orangutan, and gibbon as a sample case with the following topology taken from the UCSC genome browser:

```
      |---human
    |-|
    | |---chimp
  |-|
  | |-----gorilla
|-|
| |-------orangutan
|
|---------gibbon
```

Note that there are nine total branches that can have substitutions fall on them. These branches are human, chimp, gorilla, orangutan, gibbon, the human-chimp ancestor, the human-chimp-gorilla ancestor, and the human-chimp-gorilla-orangutan ancestor. I will abbreviate these branches, respectively, with the terms hum, chi, gor, ora, gib, hc, hcg, and hcgo. Within an ungapped alignment of these species, each species may be said to have some motif present, which may be identical or different from the motif present in the other species. For the purposes of summary, I represent these motif states at some aligned location in each of five species by five characters. Motif states that are identical to one another are represented with the same character, while motif states that are different from one another are represented by different characters. Thus 00112 would represent a motif with a common state in human and chimpanzee, a second shared state in gorilla and orangutan, and a third, unique state in gibbon. When represented in this way, there are seven basic types of patterns:

all species sharing the same motif (5)
four species having one motif with the fifth having a second (4 1)
three having one motif and two sharing a second (3 2)
3 having one motif, 1 having a second, and 1 having a third (3 1 1)
2 having one motif, 2 having a second, and 1 having a third (2 2 1)
2 with one, 1 with a second, 1 with a third, 1 with fourth (2 1 1 1)
all five species having different motifs (1 1 1 1 1)

Some of these basic patterns have multiple sub-patterns associated with them, while others do not. For example, the pattern (5) with the same motif present in all five species only has one representation (00000) while the pattern (4 1) has five different ways in which one species can differ from all the others. To avoid repeating the same pattern twice, the human state (position 1) is always represented with a 0, the second motif state encountered from left to right is given a 1, and so on until all motif states are exhausted. Thus the five ways one species can differ from the others are (00001, 00010, 00100, 01000, 01111). This analysis produces the following breakdown of 52 patterns:

(5)	(4 1)	(3 2)	(3 1 1)	(2 2 1)	(2 1 1 1)	(1 1 1 1 1)
00000	00001	00011	00012	00112	00123	01234
	00010	00101	00102	00121	01023	
	00100	00110	00120	00122	01123	
	01000	00111	01002	01012	01203	
	01111	01001	01020	01021	01213	
		01010	01112	01022	01223	
		01011	01121	01102	01230	
		01100	01200	01120	01231	
		01101	01211	01122	01232	
		01110	01222	01201	01233	
				01202		
				01210		
				01212		
				01220		
				01221		

Several of these patterns can come about through multiple mutational origins. 00001, for example, can come about through a mutation in gibbon (gib), or through a mutation in the the common ancestor of human, chimpanzee, gorilla, and orangutan (hcgo). Because our analysis is restricted to point mutations, and every point mutation both destroys the pre-existing motif and creates a new mutated motif, every mutation creates exactly as many gains as losses.

Below is a table of all patterns from above (1st column below), one potential minimum set of actual events that could produce the observed patterns (2nd column), the actual number of gains and losses that would have been involved (3rd column), a series of presence-absence values encoded by 1's and 0's, analyzed independently from the perspective of each constituent motif making up the overall species pattern (4th column) and the inferences that would be formed by trait-based parsimony based approaches that used the patterns from the 4th column. Colors are discussed below.

Supplemental File III-4

pattern	actual mutation branches	actual gainsl/losses	trait-centric patterns	inferred gains/losses
00000	no gains/losses	0G 0L	11111	0G 0L
00001	gib	1G 1L	11110 00001	**0G 2L max** 2G 0L min
00010	ora	1G 1L	11101 00010	1G 1L
00100	gor	1G 1L	11011 00100	1G 1L
01000	chi	1G 1L	10111 01000	1G 1L
01111	hum	1G 1L	10000 01111	1G 1L
00011	hcg	1G 1L	11100 00011	1G 1L
00111	hc	1G 1L	11000 00111	1G 1L
00101	gor, gib	2G 2L	11010 00101	**0G 4L max** 4G 0L min
00110	gor, ora	2G 2L	11001 00110	**0G 4L max** 4G 0L min
01001	chi, gib	2G 2L	10110 01001	2G 2L
01010	chi, ora	2G 2L	10101 01010	2G 2L
01011	hum, gor	2G 2L	10100 01011	**1G 3L max** 3G 1L min
01100	chi, gor	2G 2L	10011 01100	**1G 3L max** 3G 1L min
01101	hum, ora	2G 2L	10010 01101	2G 2L
01110	hum, gib	2G 2L	10001 01110	2G 2L
00012	ora, gib	2G 2L	11100 00010 00001	3G 0L
00102	gor, gib	2G 2L	11010 00100 00001	2G 2L max 4G 0L min
00120	gor, ora	2G 2L	11001 00100 00010	2G 2L max 4G 0L min
01002	chi, gib	2G 2L	10110 01000 00001	2G 2L
01020	chi, ora	2G 2L	10101 01000 00010	2G 2L
01112	hum, gib	2G 1L	10000 01110 00001	2G 2L
01121	hum, ora	2G 2L	10000 01101 00010	2G 2L
01200	chi, gor	2G 2L	10011 01000 00100	2G 2L max 3G 1L min
01211	hum, gor	2G 2L	10000 01011 00100	2G 2L max 3G 1L min
01222	hum, chi	2G 2L	10000 01000 00111	2G 1L

01222	hum, chi	2G 2L	10000 01000 00111	2G 1L
00112	gib, hc	2G 2L	11000 00110 00001	2G 2L max 4G 0L min
00121	hc, ora	2G 2L	11000 00101 00010	2G 2L max 4G 0L min
00122	hc, gor	2G 2L	11000 00100 00011	2G 1L
01022	hcg, chi	2G 2L	10100 01000 00011	2G 2L max 3G 1L min
01122	hcg, hum	2G 2L	10000 01100 00011	2G 2L max 3G 1L min
01201	chi, gor, gib	3G 3L	10010 01001 00100	5G 0L
01202	hum, chi, ora	3G 3L	10010 01000 00101	3G 2L max 5G 0L min
01210	hum, gor, gib	3G 3L	10001 01010 00100	5G 0L
01212	hum, chi, ora	3G 3L	10000 01010 00101	3G 2L max 5G 0L min
01220	hum, chi, gib	3G 3L	10001 01000 00110	3G 2L max 5G 0L min
01221	hum, chi, gib	3G 3L	10000 01001 00110	3G 2L max 5G 0L min
01012	chi, ora, gib	3G 3L	10100 01010 00001	4G 1L max 5G 0L min
01021	chi, ora, gib	3G 3L	10100 01001 00010	5G 0L
01102	hum, ora, gib	3G 3L	10010 01100 00001	4G 1L max 5G 0L min
01120	chi, ora, gor	3G 3L	10001 01100 00010	4G 1L max 5G 0L min
00123	gor, ora, gib	3G 3L	11000 00100 00010 00001	5G 0L
01023	chi, ora, gib	3G 3L	10100 01000 00010 00001	4G 1L max 5G 0L min
01123	hum, ora, gib	3G 3L	10000 01100 00010 00001	4G 1L max 5G 0L min
01203	chi, gor, gib	3G 3L	10010 01000 00100 00001	5G

01213	hum, gor, gib	3G 3L	10000 01010 00100 00001	5G
01223	hum, chi, gib	3G 3L	10000 01000 00110 00001	3G 2L max 5G 0L min
01230	chi, gib, ora	3G 3L	10001 01000 00100 00010	5G
01231	hum, gor, ora	3G 3L	10000 01001 00100 00010	5G
01232	hum, chi, ora	3G 3L	10000 01000 00101 00010	3G 2L max 5G 0L min
01233	hum, chi, gor	3G 3L	10000 01000 00100 00011	3G 1L
01234	hum, gor, ora, gib	4G 4L	10000 01000 00100 00010 00001	5G

This table is divided into three main sections. In the first portion (pale green background), patterns that can be generated with 0 or 1 substitutions are shown. Most of the patterns in this section either correctly infer identical numbers of gain and loss substitutions (bright green highlight) with one pattern having multiple interpretations, with the potential to equally parsimoniously infer either too many gain or too many loss substitutions (blue highlight).

In the second portion (pale orange background), patterns that can be generated with two substitutions are shown. Within this portion, 8 patterns lead to the correct inference of gains and losses, 4 can equally parsimoniously be misinferred as too many losses or too many gains, 8 can be misinferred as an imbalance in favor of gains or correctly inferred as equal gains and losses (bright orange highlight), and 3 will always be misinferred as an excess of gains relative to losses (bright red highlight).

In the third portion (pink background), patterns that can only be generated by assuming three or more substitutions are shown. Of these 21 patterns, all lead to a misinference of an excess of gains relative to losses.

This table has been generated to minimize its assumptions. While the species mutated in column 2 to produce the pattern in column 1 are arbitrary, the patterns produced in column 4 are not arbitrary, and come as a direct result of the patterns from column 1 regardless of which species have mutated.

Similarly, it can be said that by ignoring the assumption of parsimony, one could choose a more complicated series of mutations that reproduced the patterns of column 4, and that by making these mutations, it might be possible to balance gains and losses. This may be true, but in these cases, the choice of which additional mutations to make will be arbitrary if only the input patterns of column 4 are known, and inherently less likely than the smaller number of mutations assumed by maximum parsimony approaches.

Assuming maximum parsimony, all possible most parsimonious solutions are examined here, with 'tied' equally parsimonious patterns reporting a maximum and minimum number of allowable losses, along with the gains needed to produce these bounds (column 5).

Within this parameter space, it is apparent that when the number of substitutions is extremely low, most patterns will have only one or two substitutions within them, and will therefore fall in the green portion, and correctly infer equal turnover rates. Depending on the implementation used, very rarely the pattern '00001' might lead to a false inference of an excess of losses or an excess of gains. Some minority of aligned sequences may have been subject to two or more mutations, producing motifs that appear to have undergone multiple substitutions, causing these patterns to fall in the orange or even pink class of sites, and these sites are much more likely to be misinferred.

An ideal trait-based algorithm would correctly infer equivalent numbers of gains and losses in blue highlighted patterns, and would likewise give greater probabilities to recurrent losses than recurrent gains such that it would infer equivalent numbers of gains and losses in the case of orange highlighted sites. However, because red highlighted patterns are guaranteed to produce an excess of gains relative to losses, and there is no additional information in phyletic patterns to choose the correct sequence of mutations ('01234' tells us nothing about which state, if any, is ancestral), even an ideal trait-based algorithm not based on parsimony would still misinfer more gains than losses, or misinfer which species had these gains and losses, or both, far more often than it misinferred in favor of losses or correctly inferred actual events by chance. This bias worsens with worse algorithms that allow orange sites to be misinferred in favor of gains, and also worsens with increased substitution rates, which push more aligned sequence motifs into the higher substitution rate categories that have a greater likelihood of being misinferred in favor of gains.

CHAPTER IV

Conclusion

This work investigates several aspects of the evolution of genetic regulation by the piRNA and miRNA pathways, yet there are still many unanswered questions and fruitful directions for continued work.

Unanswered piRNA questions

Although it is well-established that piRNA pathway proteins appear to be rapidly evolving, and the rapid evolution of piRNA pathway proteins relative to the rest of the genome and relative to miRNA pathway proteins suggest an arms race dynamic with TEs, it is unclear exactly how such a dynamic would play out. As mentioned in the conclusion of Chapter II above, any variant of the piRNA pathway proteins which promoted the establishment of epigenetic or genetic sources of piRNAs in response to some novel TE would only be required in one individual, at which point it would be the newly established piRNA source that would sweep to fixation rather than the mutated protein. In order to explain rapid evolution in the piRNA pathway, it therefore seems much more likely that some TE-encoded inhibitors of the piRNA pathway must exist.

By operating on individual proteins in the piRNA pathway, an inhibitor of the piRNA pathway would impose strong and persistent selective pressure, favoring the emergence of piRNA pathway proteins that continue to function in the piRNA pathway while being unaffected by these inhibitors. Previous research into viruses has revealed over 50 diverse examples of viral encoded inhibitors of the evolutionarily-homologous RNAi pathway, which operates in much the same way as the piRNA pathway (reviewed in Obbard et al. 2009a). These viral inhibitors of the RNAi pathway operate through a variety of mechanisms, flooding the system with false RNAi substrates or allosterically blocking RNAi proteins, and in some cases the inhibitors are formed from structured RNA molecules rather than protein. Many TE families appear to derive from viruses that have lost the ability to invade new cells, and several encode many of the key proteins required for viral particle formation. Given that some components of RNAi machinery (such as Argonaute2) have homologs known to be essential to the piRNA pathway (Piwi, Argonaute3, and Aubergine), it seems plausible and even likely that some TEs have found ways to inhibit the piRNA pathway, either through repurposing of existing inhibitors in the RNAi pathway, or the denovo evolution of piRNA inhibitors.

Given the rapid evolution of the piRNA pathway and the strong possibility of TE encoded inhibitors of the piRNA pathway, a logical next step in understanding the evolutionary dynamics of this biological system would be the isolation of these inhibitors. The most comprehensive database of annotated TEs is known as RepBase. This database examines TEs across a broad swath of species, and is freely available upon written request from a principal investigator. The database includes several Drosophila species, and annotates the predicted open reading frames and consensus sequences of hundreds of TEs. Oddly, this database contains hundreds of TEs annotated as belonging to *D. melanogaster*, yet only a scattering (<10) of TEs are annotated for any other *Drosophila* species. It is unclear whether this discrepancy is due to biological differences in TE composition, sampling bias in the discovery of species-specific TEs from model vs. non-model organisms, or errors in the genome assemblies of non-model organisms. I conducted a blat search of the *D. melanogaster* genome, using *D. melanogaster* consensus elements annotated in Repbase. Most of the elements had at least one high confidence hit in the genome, but several elements had no high confidence copies. Many well-studied, relatively abundant elements, such as Het-A, do not appear to be present in the official melanogaster genome release, suggesting that in some cases unfinished assemblies may explain missing copies. Of the copies that are found in the D. melanogaster genome, a large fraction have polymorphisms associated with them, which could reflect coevolutionary arms race dynamics or the neutral accumulation of deleterious alleles in quiescent elements.

Using the polymorphisms to map TE reads to individual copies within the genome, I have examined RNAseq libraries generated by the Theurkauf lab in various piRNA pathway mutants and found that in many cases one or a few TEs appear to be responsible for a large amount of the observed TE expression. If the hypothesis that TE expression correlates with TE mobilization within the genome is correct, then transcriptionally active copies of TEs might represent functional elements with intact open reading frames, and these open reading frames may encode inhibitors of the piRNA pathway.

Several alternative possibilities may undermine a correlation between TE expression level and TE function; TEs that duplicate efficiently within genomes may not be those that are highest expressed, and highly expressed TEs may be expressed due to fortuitous placement in regions adjacent to highly expressed transcripts deriving from protein-coding genes and other highly expressed sources.

Even if transcription can be used as a proxy for TE activity, the nonspecific nature of piRNA pathway inhibitors pose interesting functional and evolutionary challenges. From a functional perspective, because a piRNA pathway inhibitor allows all TEs to become activated, the most actively transposing elements may not necessarily be those that encode inhibitors of the piRNA pathway. Indeed, there are several classes of TE that transpose quite effectively without any endogenous source of transposase, 'piggybacking' off of the transposases of other elements. There is no reason why the same dynamic might not play out with regard to piRNA inhibition. From an evolutionary perspective, TEs compete not only against the piRNA pathway defenses of the host genome but also against other elements. Those elements that encode inhibitors of the piRNA pathway should allow all functional elements in a genome to become active in a nonspecific way. To some extent, this property undermines the assumption in evolutionary biology that beneficial alleles sweep through populations in a largely deterministic manner. Elements that are highly active in the absence of the piRNA pathway may outcompete those that encode a TE inhibitor but transpose at low levels, thus creating the appearance of a selective sweep for a non-mutated TE and allowing for the loss of the beneficial allele that encodes a piRNA pathway inhibitor (which may itself occur before or after modifications in piRNA pathway proteins targeted by this mystery inhibitor).

This trait of piRNA pathway inhibition, in combination with ascertainment biases in the discovery of novel TEs from non-model organisms, may largely explain the results of Castillo et al. 2011, who found that rapid evolution in the piRNA pathway does not appear to correlate with TE family numbers or the copy numbers of individual TEs. Rather than searching for a TE that has expanded dramatically in correlation with a given set of piRNA pathway proteins, it may be productive, in looking for a TE encoded inhibitor of the piRNA pathway, to search for individual TE-encoded proteins whose amino acid substitutions correlate with amino acid substitutions in piRNA pathway proteins that have been examined in this thesis. On identifying candidate TEs that encode a piRNA pathway inhibitor, it would be a relatively simple matter to create a reporter of trans activation of TE elements on overexpression of the protein in question, in a cell culture system possessing a piRNA pathway naïve to this protein, as proposed in a personal conversation with Dr. Travis Thomson during his time in the Theurkauf lab. In all cases, such an analysis will depend on a thorough and accurate knowledge of the sequences of the TEs expressed across a large number of closely related *Drosophila* species. Although there are prominent examples in the literature of individual TE families that have been annotated across many *Drosophila* species, from my search and prior personal communications with Dr. Casey Bergman, I am not convinced that such a comprehensive, unbiased dataset exists in *Drosophila* species beyond *D. melanogaster* as of yet.

Recurrent selection within piRNA clusters

As mentioned above, any response to novel TEs that invade naïve populations by vertical inheritance, or that invade naïve species by horizontal transfer through some infectious particle, would strongly select for epigenetic changes in the pool of mature piRNAs, or genetic changes in piRNA clusters that encode piRNAs.

Even preexisting TEs appear to undergo periods of inactivity, interspersed with 'bursts' of transposition. This process may result from extended periods of transcriptional silencing (perhaps through methylation of TE-encoding regions of DNA) followed by fortuitous reactivation, or may come about through extended periods of efficient piRNA silencing, interspersed with dramatic failures of the piRNA pathway due to the theoretically encoded piRNA pathway inhibitors of other TEs.

Regardless of the source, when transpositional activity fluctuates, the selective pressure for a robust genetic response to TEs must also fluctuate, and this should be reflected in the dynamics of piRNA clusters. The Theurkauf lab has sequenced hybrid dysgenic lines of flies, and found hundreds of kilobases of new insertion material within piRNA clusters in the space of a single generation. This work also has found weak biases toward transposition into existing piRNA clusters (Khurana et al. 2011). Because piRNA clusters are composed of large numbers of TEs that have inserted within one another (Brennecke et al. 2007) it may be the case that the piRNA system has found a way to harness the act of transposition as a means of generating a strengthened genetic response to the most active elements. Taken together with the large amount of inserted genomic material over the course of a single generation, and the observed stable size of the *D. melanogaster* genome over many generations, these results imply that piRNA clusters must undergo continual turnover in their TE composition.

This functional evidence of rapid turnover within piRNA clusters is in agreement with evolutionary theory, which would predict that selection will be strongest for the suppression of the most deleterious elements, and considerably weaker for less active elements, with strong pressure to maintain sequence homology to active TEs, and weaker or nonexistent selection on the maintenance of homology to quiescent elements. By using cluster composition, it may therefore be possible to obtain a secondary perspective on the elements that pose the greatest threat to existing *Drosophila* populations, and to correlate the degeneracy and age of an inferred TE sequence with the times at which these TEs may have been active. If this information can be generated, it may even be possible to infer relationships between active TEs and changes in piRNA pathway proteins.

An initial hypothesis from this theoretical framework is that functional portions of piRNA clusters will experience lower substitution rates than nonfunctional portions. Because piRNA clusters have not been identified in other *Drosophila* species, and experimental evidence suggests that piRNA clusters may change rapidly within a single generation, I turned to polymorphism data from 139 sequenced D. melanogaster individuals, which were publicly available as part of the second phase of the Drosophila Genomics Reference Panel project (DGRP2). I analyzed a statistic of the fraction of pairwise nucleotide comparisons that are non-identical within an alignment (traditionally denoted in population genetics with the greek symbol π) and found that piRNA clusters do not have distinctive signatures of π relative to the rest of the genome. However, on closer examination, the DPGP2 dataset masked insertions and deletions relative to the reference strain with 'N' characters, and the piRNA clusters of these additional strains may have suffered from significant assembly errors due to the repetitive nature of the TE sequences contained within them. Additionally, although there are 142 piRNA clusters identified in Julius Brennecke's work (Brennecke 2007), only flamenco has any published experimentally validated piRNA pathway phenotype. It may be possible that only a small fraction of piRNA clusters are functional. Future evolutionary and molecular studies will be needed to identify clusters that are biologically relevant.

While I examined substitution rates within piRNA clusters, most of the turnover events observed in TEs are likely to be in the form of novel insertions into clusters, or deletions of unused material from clusters that have grown too large. Jiali Zhuang has developed a program that can parse through deep sequence data and find insertion and deletion events for TEs, both within and outside of piRNA clusters. This program has been successfully used to detect generation-specific insertion/deletions (indels) relative to the reference genome, and it seems that the approach could be used productively in the investigation of the individuals that have already been sequenced by the DPGP2 project.

I have used a slightly different approach in an attempt to reconstruct the evolutionary history of piRNA clusters. In this approach, I found short 19nt stretches of sequence having a perfect match to exactly two locations within the genome, one of which was constrained to be within a piRNA cluster. In several cases, I found that the same cluster had multiple 19nt matches to the same region of the genome, with identical spacing between the 19mers in the cluster and the corresponding

19mers at the other location in the genome. These observations strongly suggest a common origin between these regions of the genome. If piRNA clusters have some ability to direct transposition, and serve as sources of productive TE silencing, I would predict that these regions of identity would correspond to TEs or have TEs flanking them, and that these regions of perfect identity would occur much more often within clusters than within other regions of the genome. This technique might productively be used in future studies to infer the evolutionary history and most frequently preserved mechanism of transposition events into piRNA clusters, and would complement direct studies of currently occurring transposition events by adding the dimension of natural selection on these events across time.

Unanswered miRNA questions

My current work has found strong signatures of selective constraint, both in terms of the disruption of existing miRNA binding sites, and in terms of the creation of new miRNA binding sites, relative to other 3' UTR motifs. However, these conclusions only scratch the surface of the processes that may be occurring in this system. Although 7 out of 93 highly conserved miRNAs have gain and loss ranks that are both 1, and this number is extremely enriched relative to the 60 8mers having these ranks in the full dataset of 65,536, there are still 86 miRNAs that are highly conserved, yet have turnover rates that not in this slowest turning over class. I have re-examined the data for the purposes of this thesis, using a newer version of mirbase (mirbase 21 instead of mirbase 18) and found that in this new dataset, there are now 117 8mers that are found as nucleotides 1-8 of a miRNA in humans, mice, and fish. Of these 117, there are still 7 that have gain ranks of 1 and loss ranks of 1, which yields a new significance level of 2.4*10^-11. In order to repeat this process and investigate the probability distribution of the ranks of all 117 highly conserved miRNAs, I generalized this probability process by summing the ranks. The ranks 1 and 1 sum to 2, and can only be achieved with a gain rank of 1 and loss rank of 1. I next asked how many conserved miRNAs had ranks that sum to 3 (with either a gain rank of 2 and loss rank of 1, or vice-versa). I found 3/117 (0.026) conserved miRNAs whose ranks sum to 3, as compared to 109/65,536 (0.002)

8mers in the full dataset. The probability of finding 3 or more 8mers with ranks that sum to 3 out of a random set of 117 8mers is ~0.001, indicating that conserved miRNAs are also significantly enriched for gain and loss ranks that sum to 3 relative to the full dataset. These probabilities (0.026 vs. 0.002) were graphed, and the process was repeated for every sum between 2 (gain rank and loss rank of 1) and 50 (gain rank and loss rank of 25). Within the full dataset of all 8mers, the probabilities associated with these sums appear to approximate a normal distribution, with the highest probability at a gain/loss sum of 27, and a corresponding probability of .042 (2,784 8mers out of 65,536). Conserved miRNAs in general had significantly higher probabilities of being low ranking than random 8mers (in the gain/loss sums 2-12, 9 out of 11 sums have significantly higher probabilities of occurrence in the conserved miRNA dataset than random 8mers, 1 out of 11 had an insignificantly lower probability, and one had an insignificantly higher probability). In general, conserved miRNAs also had insignificantly lower probabilities of being high ranking than random 8mers (in the rank sums 39-49, all 11 ranks had insignificantly lower probabilities of being represented in the conserved miRNA dataset relative to the full dataset). Conserved miRNAs in general also had insignificantly lower probabilities of occupying the middle ranks (for the ranks 13-38, 4/26 had significantly lower probabilities of being found in these ranks, 2/26 had insignificantly higher probabilities of occupying these ranks, and 20/26 had insignificantly lower probabilities of occupying these ranks than the full dataset of 8mers). All of this data is graphed in Figure IV-1.

To assess the effect of conservation level of miRNAs on turnover rank of corresponding 8mers, I examined the turnover rates of miRNAs that were conserved in human, mouse, fish, and nematode (*C. elegans*), as well as miRNAs conserved in human, mouse, fish, and fly. I also examined the turnover rates of miRNAs conserved in human, mouse, fish, fly, and nematode. In all cases, I found significantly higher probabilities of low turnover rates for the highly conserved miRNAs, and insignificantly lower probabilities of high turnover rates, relative to the full dataset. I repeated the analysis using gain ranks alone and loss ranks alone, and found results consistent with those for summed ranks (namely, low gain ranks alone or loss ranks alone were consistently significantly over-represented in conserved miRNAs, while high ranks were consistently insignificantly under-represented in conserved miRNAs). None of the results obtained for more broadly conserved miRNAs were as significant as those obtained for miRNAs conserved in humans, mice and fish.

Taken together, these results suggest that conserved miRNAs are dramatically more likely to act in highly conserved contexts than random 8mers, while being consistently slightly less likely to be found in neutrally and rapidly evolving contexts than random 8mers. In addition, miRNAs that are not conserved beyond vertebrates tend to play roles in primates that are as conserved as the roles played in primates by miRNAs that are more deeply conserved in nematodes and/or insects.

While the results above are informative, they represent only the beginning of a more thorough analysis, which would ideally examine the binding sites of less well-conserved miRNAs, turnover rates within individual and novel species, and the turnover rates of different functional categories of genes. I would hypothesize that miRNAs that are only conserved within primates have roles that are more likely to be changing rapidly than more deeply conserved miRNAs, as these primate-specific miRNAs would have existed for too short of a time to acquire deeply conserved roles, and that for similar reasons vertebrate species groups that have few substitutions relative to the common ancestor of vertebrates are likely to show higher binding site turnover rates than miRNAs conserved out to nematode and insect.

It is striking that highly conserved miRNAs are so much more likely to be found in conserved contexts than random 8mers, especially when one considers that many 8mers matching the reverse-complement of nucleotides 1-8 of conserved miRNAs are likely to be without function. By examining the data in further detail, it may be possible to determine whether this signal of conservation derives from a few highly conserved interactions or a large number of moderately conserved processes, which may in turn suggest the number of functional targets that miRNAs tend to regulate over deep evolutionary time. By cross-referencing this dataset with large-scale unbiased targeting assays using PAR-CLIP and HITS-CLIP technology, it may be similarly possible to learn whether functional sites tend to be conserved or rapidly evolving, and what fraction of sites that are potentially functional according to evolutionary analysis are recovered by HITS-CLIP and PAR-CLIP techniques.

This work investigates turnover rates among binding sites rather than turnover rates in gene targeting. In this analysis, it is possible that binding sites turn over much more frequently than the gain and loss of targeting by a particular miRNA, and that because selection would be predicted to act more strongly on gene targeting than on binding site turnover that doesn't alter gene targeting, the observed turnover events are not being measured with the most relevant metric. I therefore examined all 117 miRNAs conserved in human, mouse, and fish, and found that of the 3' UTRs whose target sites for these miRNAs turned over (34,298 when examined across all possible ancestor/descendant pairs, Table IV-1), roughly 45 percent of site turnover events observed were associated with an ancestral single site that was lost in the descendant. An additional 46 percent of site turnover events were associated with an ancestral 3' UTR lacking a site that gained a site in the descendant. An additional 4 percent of site turnover events were associated with 3' UTRs that went from having 2 ancestral sites to 1 descendant site, and 3 percent of site turnover events were associated with 3' UTRs that went from having 1 ancestral site to 2 descendant sites. More complicated site turnover events (including site turnover events that do not change the total number of binding sites in a 3' UTR) make up less than 2 percent of all observed turnover events. It can therefore be concluded that roughly 91 percent of site turnover events reported in this study reflect gains or losses in gene targeting, and that within this recently diverged primate dataset, gains or losses in gene targeting can be approximated by gains and losses in binding sites. In future studies, especially for datasets having higher turnover rates in 3' UTRs, it will be useful to distinguish between binding site turnover and turnover in genes targeted or not targeted by miRNAs, and to investigate the extent to which natural selection operates in both scenarios.

While the approaches described so far can be used to infer the enrichment or depletion of miRNAs whose binding sites are turning over at a given rate, these methods do not have power to infer the significance of binding site turnover rates for any one miRNA. If the full dataset of all 8mers has a probability of 0.042 of having gain and loss ranks whose sum is 27, while the probability of ranks that sum to 27 in conserved miRNAs is 2/117, or 0.017, I can say that conserved miRNAs have a slightly lower probability of having this sum than the full dataset, but I can't say anything about whether the two real miRNAs having this sum are participating in processes that are rapidly turning over, are mostly without function in primates, or are participating in a small number of highly conserved processes that are overwhelmed by a large number of neutrally evolving artifactual targets with no biological significance. With further work, I hope to be able to address these types of questions for individual miRNAs. It seems possible and even likely that some examples can be found of miRNAs that are rapidly acquiring novel, biologically significant functions. I hypothesize that organisms that have recently colonized new environments, that have a large number of specialized adaptations to these environments, will be more likely to display rapidly evolved novel functions for conserved miRNAs than species occupying stable evolutionary niches. Future work will be directed to a more detailed investigation of the landscape of miRNA binding site targeting.

Conclusion

In my work with collaborators, I have examined both the piRNA and miRNA pathways, and used evolutionary approaches to make useful inferences about several aspects of these processes. Within the piRNA pathway, I have shown, with a range of evolutionary approaches, that several piRNA pathway proteins appear to be evolving quite rapidly relative to the rest of the genome, and that different piRNA proteins evolve rapidly at different evolutionary timescales. These rapid changes in the piRNA pathway could be driven by novel transposable element threats to the genome, which may occur with differing mechanisms and periodicities. My collaborators and I have posited at least two distinct mechanisms by which TEs might threaten genomes at different timescales, requiring either the epigenetic formation of new piRNAs to 'seed' the ping-pong amplification cycle when the novel TE invades vertically through the male germline, or the genetic creation of entirely new clusters to respond to novel TE threats that invade horizontally between species. These circumstances, although rare, could theoretically represent major challenges to genome integrity, and in these circumstances the re-establishment of silencing would be crucial to survival. This model does not address how a mutation that allowed the restoration of piRNA efficacy would sweep to fixation in the absence of continued selective pressure, but regardless of the nature of the novel threats that may be posed by TEs, it seems likely from our research that a small subset of the piRNA machinery responds in each case, and our work discovers essential roles that these proteins may play in rare circumstances that usually cannot be directly observed experimentally.

Although this work provides a deeper understanding of the protein components that interact with TEs, it is far from complete. There are at least two additional, essential components of adaptation to novel transposons that might be productively examined in greater detail. First, while the proteins that regulate TEs are rapidly evolving, little is known about the TEs that have presumably coevolved as drivers of this process. Most TE copies cannot be easily mapped within the genome, and so it is difficult to estimate how fast these TE drivers are evolving. In addition, the biology of TEs is often unknown, with a nearly infinite and poorly studied plethora of replication methods, encoded proteins, and nested interdependencies within and across families, any one of which could present novel and significant threats to genomes. Further knowledge of TE biology and replication dynamics will be essential to an understanding of the exact evolutionary pressures faced by piRNA proteins.

As a second component of a complete picture of the evolutionary dynamics of TE silencing, a thorough understanding of the evolutionary dynamics of piRNA loci will be invaluable. While piRNA pathway proteins are the effectors of TE silencing, in current models it is piRNAs that are thought to serve as the adaptive component to novel TE threats. As such, one would expect piRNA loci to be adapting to TEs even more rapidly than piRNA pathway proteins. Moreover, because piRNAs maintain sequence homology to the TEs they silence, any selectively driven piRNA loci should reproduce a record of the active TEs that present a continual threat to genome integrity.

Within the miRNA pathway, I have demonstrated that the binding sites of conserved miRNAs are significantly enriched for slowly turning over motifs. I have also found that while eight nucleotide binding sites corresponding to the reverse complement of nucleotides 1-8 of the mature miRNA are more enriched for slowly turning over sequences than seven nucleotide motifs corresponding to nucleotides 2-8, constraining the first nucleotide to be an 'A' opposite the first nucleotide (rather than the reverse complement of the first nucleotide) causes even stronger enrichment for slowly turning over sequences. As a class, I've found that 7/93 highly conserved miRNAs have binding sites that fall at the highest conserved end of my scoring system. This implies that the functions of these several conserved miRNAs tend to be highly preserved over evolutionary time. Although the sample size is too small for meaningful formal statistics, it is interesting to note that fully three out of the seven well-conserved miRNAs in the slowest turning over class are highly expressed in neural tissue in human (miR-124, miR-9, and miR-92a/92b), suggesting that the functions of neural miRNAs may be highly constrained within primates. Equally interesting is the observation that several other famously highly conserved miRNAs, such as lin-4 and let-7, have target sites that do not fall within this most highly conserved group, while still other well-conserved miRNAs such as miR-10 appear to have targets that turnover at much higher rates than the median taken across all eightmers in the dataset. While it is important to stress that the turnover rates of individual miRNAs have no statistical significance associated, the highly significant number of miRNAs obtained within the 'slowest turnover' class of binding sites illustrate the power of this method to identify biological trends, and can be used to form intuitions as to which miRNAs are likely to have undergone dramatic shifts in function, and which seem to play conserved roles in developmental processes. This work begins to ask the fundamental question of how often miRNAs change their functions, which is tantamount to asking what biological roles miRNAs have evolved to play, and which can only be answered in a global way by evolutionary approaches similar to the one taken here.

This thesis set out to use evolutionary techniques in order to determine the rules that tend to predominate in small RNA pathways. Using these techniques, this study has been able to improve on existing knowledge concerning the historical relationship between several piRNA pathway proteins and the TEs they silence, and has determined that a subclass of miRNAs exist within primates whose overall targeting patterns are highly conserved over evolutionary time. This work has additionally found evidence in support of the notion that the eighth nucleotide of miRNA binding sites (opposite nucleotide 1 of mature miRNAs) contributes strongly to target recognition, but through mechanisms other than basepairing. Finally, this work has demonstrated a major flaw in existing methodologies to phylogenetic inference, as well as a means of successfully addressing this flaw. Although much remains to be done, and biology continues to present a seemingly infinite regress of new questions with every answer, with these results, we are one step closer to understanding how and why the miRNA and piRNA pathways came to be what they are today.

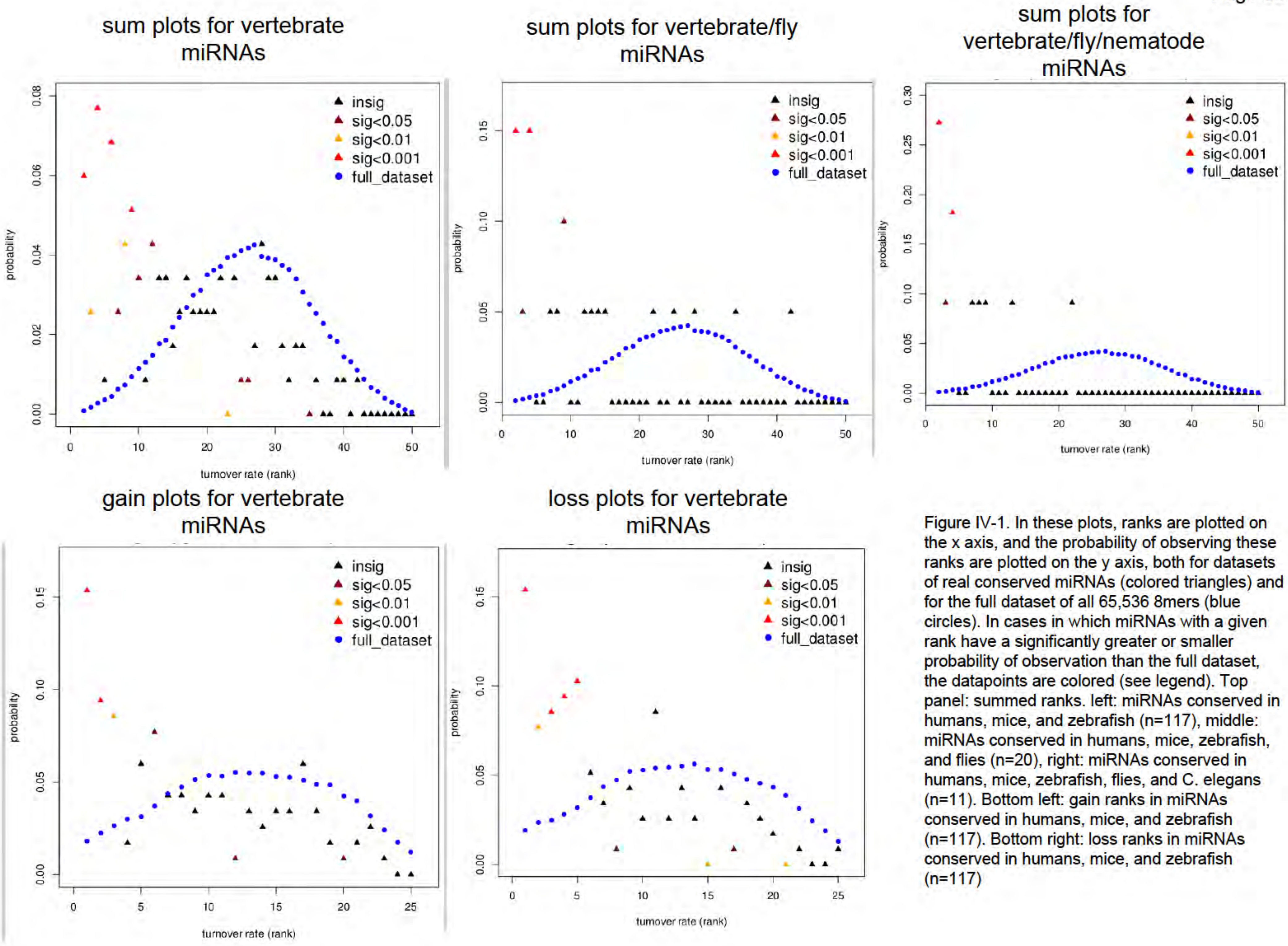


Figure IV-1. In these plots, ranks are plotted on the x axis, and the probability of observing these ranks are plotted on the y axis, both for datasets of real conserved miRNAs (colored triangles) and for the full dataset of all 65,536 8mers (blue circles). In cases in which miRNAs with a given rank have a significantly greater or smaller probability of observation than the full dataset, the datapoints are colored (see legend). Top panel: summed ranks. left: miRNAs conserved in humans, mice, and zebrafish (n=117), middle: miRNAs conserved in humans, mice, zebrafish, and flies (n=20), right: miRNAs conserved in humans, mice, zebrafish, flies, and C. elegans (n=11). Bottom left: gain ranks in miRNAs conserved in humans, mice, and zebrafish (n=117). Bottom right: loss ranks in miRNAs conserved in humans, mice, and zebrafish (n=117)

Table IV-1

ancestor	**descendan**	**count**	**proportion**
6	5	2	6E-005
2	4	2	6E-005
4	5	3	9E-005
4	2	3	9E-005
1	3	4	0.00012
2	2	11	0.00032
3	4	12	0.00035
3	1	20	0.00058
4	3	20	0.00058
0	2	42	0.00122
0	2	42	0.00122
2	0	61	0.00178
1	1	81	0.00236
2	3	124	0.00362
3	2	180	0.00525
1	2	1107	0.03228
2	1	1302	0.03797
1	0	15386	0.44866
0	1	15891	0.46339
		34293	total turnover instances

Table IV-1 turnover events from ancestral UTRs (first column) to descendant UTRs (second column) and the counts associated

REFERENCES

Aalto AP, Pasquinelli AE. 2012. Small non-coding RNAs mount a silent revolution in gene expression. Curr Opin Cell Biol 24: 333–40.

Anxolabéhère D, Kidwell MG, Periquet G. 1988. Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of Drosophila melanogaster by mobile P elements. Mol. Biol. Evol. 5: 252-269.

Bai Y, Dai X, Harrison AP, Chen M. 2014. RNA regulatory networks in animals and plants: a long noncoding RNA perspective. Brief Funct Genomics.

Bazzini AA, Lee MT, Giraldez AJ. 2012. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science 336: 233–7.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128: 1089-103.

Brennecke J, Stark A, Russell RB, Cohen SM. 2005. Principles of microRNA-target Recognition. PLoS Biol. 3: e85.

Brower-Toland B, Findley SD, Jiang L, Liu L, Yin H, Dus M, Zhou P, Elgin SCR, Lin H. 2007. Drosophila PIWI associates with chromatin and interacts directly with HP1a. Genes & development 21: 2300-11.

Cassidy JJ, Jha AR, Posadas DM, Giri R, Venken KJ, Ji J, Jiang H, Bellen HJ, White KP, Carthew RW. 2013. miR-9a minimizes the phenotypic impact of genomic diversity by buffering a transcription factor. Cell 155: 1556–1567.

Castillo DM, Mell JC, Box KS, Blumenstiel JP. 2011. Molecular evolution under increasing transposable element burden in Drosophila: A speed limit on the evolutionary arms race. BMC evolutionary biology 11:258.

Cech TR. 2000. Structural biology. The ribosome is a ribozyme. Science 289: 878–9.

Chen K, Rajewsky N. 2006a. Deep Conservation of microRNA-target Relationships and 3'UTR Motifs in Vertebrates, Flies, and Nematodes. Cold Spring Harb Sym. 71: 149–56.

Chen K, Rajewsky N. 2006b. Natural Selection on Human microRNA Binding Sites Inferred from SNP Data. Nat Genet. 38: 1452–1456.

Chen K, Rajewsky N. 2007. The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet. 8: 93–103.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203-18.

Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG, Chovnick A. 1990. Evidence for horizontal transmission of the P transposable element between Drosophila species. Genetics 124: 339-55.

Díaz-González J, Domínguez A, Albornoz J. 2010. Genomic Distribution of Retrotransposons 297, 1731, Copia, Mdg1 and Roo in the Drosophila Melanogaster Species Subgroup. Genetica 138: 579–86.

ENCODE project consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

Faehnle CR, Elkayam E, Haase AD, Hannon GJ, Joshua-Tor L. 2013. The Making of a Slicer: Activation of Human Argonaute-1. Cell Rep. 3: 1901–1909.

Farh KK, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP. 2005. The Widespread Impact of Mammalian MicroRNAs on mRNA Repression and Evolution. Science. 310: 1817–1821.

Fay JC, Wu CI. 2000. Hitchhiking Under Positive Darwinian Selection. Genetics 155: 1405-13.

Felsenstein J. PHYLIP [updated 2013 May 30]. Available from: http://evolution.genetics.washington.edu/phylip.html

Gao FB. 2010. Context-dependent functions of specific microRNAs in neuronal development. Neural Dev. 5:25.

Griffiths-Jones S, Hui JHL, Marco A, Ronshaugen M. 2011. MicroRNA Evolution by Arm Switching. EMBO Rep. 12: 172–7.

Guo H, Ingolia NT, Weissman JS, Bartel DP. 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 466: 835–840.

Harris RS. 2007. Improved pairwise alignment of genomic DNA. PhD thesis. The Pennsylvania State University.

Hedges S, Blair JD, Kumar S. 2006. TimeTree: a Public Knowledge-base of Divergence Times Among Organisms. Bioinformatics. 22: 2971–2972.

Hernandez RD. 2008. A Flexible Forward Simulator for Populations Subject to Selection and Demography. Bioinformatics. 24: 2786–2787.

Hiard S, Charlier C, Coppieters W, Georges M, Baurain D. 2010. Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. Nucleic acids research. 38:D640–D651.

Fiston-Lavier AS, Singh ND, Lipatov M, Petrov DA. 2010. Drosophila melanogaster recombination rate calculator. Gene 463: 18-20.

Fletcher W, Yang Z. 2010. The Effect of Insertions, Deletions and Alignment Errors on the Branch-Site Test of Positive Selection. Molecular biology and evolution 27: 2257-2267.

Gao F-B. 2010. Context-dependent functions of specific microRNAs in neuronal development. Neural Dev 5: 25.

Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, Xu J, Kittler ELW, Zapp ML, Weng Z, et al. 2008. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. Science 320: 1077–81.

Griffiths-Jones S, Hui JHL, Marco A, Ronshaugen M. 2011. MicroRNA evolution by arm switching. EMBO Rep 12: 172–7.

Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC. 2007. A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in

Drosophila. Science 315: 1587-90.

Guo H, Ingolia NT, Weissman JS, Bartel DP. 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 466: 835–840.

Huisinga KL, Elgin SCR. 2009. Small RNA-directed heterochromatin formation in the context of development: what flies might learn from fission yeast. Biochimica et biophysica acta 1789: 3-16.

Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170: 1401-10.

Kazazian HH. 2004. Mobile elements: drivers of genome evolution. Science 303: 1626-32.

Khurana JS, Wang J, Xu J, Koppetsch BS, Thomson TC, Nowosielska A, Li C, Zamore PD, Weng Z, Theurkauf WE. 2011. Adaptation to P element transposon invasion in Drosophila melanogaster. Cell 147: 1551-63.

Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765-77.

Klattenhoff C, Xi H, Li C, Lee S, Xu J et al. 2009. The Drosophila HP1 homolog Rhino is required for transposon silencing and piRNA production by dual-strand clusters. Cell 138: 1137-49.

Kolaczkowski B, Hupalo DN, Kern AD. 2011. Recurrent adaptation in RNA-interference genes across the Drosophila phylogeny. Molecular biology and evolution 28: 1033-1042.

Kucherenko MM, Barth J, Fiala A, Shcherbata HR. 2012. Steroid-induced microRNA let-7 acts as a spatio-temporal code for neuronal cell fate in the developing Drosophila brain. EMBO J. 31: 4511–4523.

La Torre A, Georgi S, Reh TA. 2013. Conserved microRNA pathway regulates developmental timing of retinal neurogenesis. Proc Natl Acad Sci. 110: E2362–E2370.

Le Rouzic A, Capy P. 2005. The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. Genetics 169: 1033-1043.

Lecellier C-H, Dunoyer P, Arar K, Lehmann-Che J, Eyquem S, Himber C, Saïb A, Voinnet O. 2005.

A cellular microRNA mediates antiviral defense in human cells. Science 308: 557–60.

Lewis BP, Burge CB, Bartel DP. 2005. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates That Thousands of Human Genes Are microRNA Targets. Cell. 120: 15–20.

Lewis BP, I-hung S, Jones-Rhoades MW, Bartel DP, Burge CB. 2003. Prediction of Mammalian microRNA Targets. Cell. 115: 787–798.

Li C, Vagin VV, Lee S, Xu J, Ma S et al. 2009. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. Cell 137: 509-521.

Li X, Cassidy JJ, Reinke CA, Fischboeck S, Carthew RW. 2009. A microRNA Imparts Robustness Against Environmental Fluctuation During Development. Cell. 137: 273–282.

Li XZ, Roy CK, Dong X, Bolcun-Filas E, Wang J, Han BW, Xu J, Moore MJ, Schimenti JC, Weng Z, et al. 2013. An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. Mol Cell 50: 67–81.

Li Y, Wang F, Lee JA, Gao FB. 2006. MicroRNA-9a ensures the precise specification of sensory organ precursors in *Drosophila*. Genes Dev. 20: 2793–2805.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451-1452.

Lim AK, Kai T. 2007. Unique germ-line organelle, nuage, functions to repress selfish genetic elements in Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America 104: 6714-6719.

Liu G, Mattick JS, Taft RJ. 2013. A meta-analysis of the genomic and transcriptomic composition of complex life. Cell Cycle 12: 2061–72.

Liu N, Okamura K, Tyler DM, Phillips MD, Chung W-J, Lai EC. 2008. The evolution and functional diversification of animal microRNA genes. Cell Res 18: 985–96.

Loh YE, Yi SV, Streelman JT. 2010. Evolution of MicroRNAs and the Diversification of Species. Genome Biol Evol. 3: 55–65.

Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National Academy of Sciences of the United States of America 102: 10557-10562.

Lynch M. 2007. The Evolution of Genetic Networks by Non-adaptive Processes. Nat Rev Genet. 8: 803–813.

Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, Hannon GJ. 2009. Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. Cell 137: 522-535.

Mattick JS, Gagen MJ. 2001. Review Article The Evolution of Controlled Multitasked Gene Networks : The Role of Introns and Other Noncoding RNAs in the Development of Complex Organisms. 1611–1630.

McDonald JH Kreitman M. 1991. Adaptive Protein Evolution at the Adh Locus in Drosophila. Nature 351: 652-4.

Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, Hu H, Khaitovich P, Kaessmann H. 2013. Birth and expression evolution of mammalian microRNA genes. Genome Res. 23: 34–45.

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confi- dencemicroRNAs using deep sequencing data. Nucleic Acids Res. 42: D68–D69.

Miska EA, Alvarez-Saavedra E, Abbott AL, Lau NC, Hellman AB, McGonagle SM, Bartel DP, Ambros VR, Horvitz HR. 2007. Most Caenorhabditis Elegans microRNAs Are Individually Not Essential for Development or Viability. Plos Genet. 3: e215.

Morris KV. 2009. RNA-directed transcriptional gene silencing and activation in human cells. Oligonucleotides 19: 299–306.

Moshkovich N, Lei EP. 2010. HP1 recruitment in the absence of argonaute proteins in Drosophila. PLoS genetics 6: e1000880.

Moss EG, Tang L. 2003. Conservation of the Heterochronic Regulator Lin-28, Its Developmental Expression and microRNA Complementary Sites. Dev Biol. 258: 432–442.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular biology and evolution 3: 418-426.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. Genome research 15: 1566-1575.

Nozawa M, Miura S, Nei M. 2010. Origins and evolution of microRNA genes in Drosophila species. Genome Biol Evol. 2: 180–189.

Nozawa M, Miura S, Nei M. 2012. Origins and evolution of microRNA genes in plant species. Genome Biol Evol. 4: 230–239.

Obbard DJ, Gordon KHJ, Buck AH, Jiggins FM. 2009a. The evolution of RNAi as a defence against viruses and transposable elements. Philosophical transactions of the Royal Society of London 364: 99-115.

Obbard, DJ, Welch JJ, Kim K-W, Jiggins FM. 2009b. Quantifying adaptive evolution in the Drosophila immune system. PLoS Genetics 5: e1000698.

Obbard DJ, Jiggins FM, Bradshaw NJ, Little TJ. 2011. Recent and Recurrent Selective Sweeps of the Antiviral RNAi Gene Argonaute-2 in Three Species of Drosophila. Molecular Biology and Evolution 28: 1043-1056.

Obbard DJ, Jiggins FM, Halligan DL, Little TJ. 2006. Natural selection drives extremely rapid evolution in antiviral RNAi genes. Current biology 16: 580-585.

Pal-Bhadra M, Bhadra U, Birchler JA. 2002. RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in Drosophila. Molecular cell 9: 315-327.

Pal-Bhadra M, Leibovitch BA, Gandhi SG, Rao M, Bhadra U, Birchler JA, Elgin SCR. 2004. Heterochromatic silencing and HP1 localization in Drosophila are dependent on the RNAi machinery. Science (New York, and N.Y.) 303: 669-672.

Pane A, Wehr K, Schüpbach T. 2007. Zucchini and squash encode two putative nucleases required for rasiRNA production in the Drosophila germline. Developmental cell 12: 851-862.

Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI et al. 2000. Conservation of the Sequence and Temporal Expression of Let-7 Heterochronic Regulatory RNA. Nature. 408: 86–89.

Przeworski M. 2002. The signature of positive selection at randomly chosen loci. Genetics 160: 1179-1189.

Ramesh SV, Ratnaparkhe MB, Kumawat G, Gupta GK, Husain SM. 2014. Plant miRNAome and antiviral resistance: a retrospective view and prospective challenges. Virus Genes 48: 1–14.

Rogers K, Chen X. 2013. Biogenesis, turnover, and mode of action of plant microRNAs. Plant Cell 25: 2383–99.

Rozhkov NV, Aravin AA, Zelentsova ES, Schostak NG, Sachidanandam R, McCombie WR, Hannon GJ, Evgen'ev MB. 2010. Small RNA-based Silencing Strategies for Transposons in the Process of Invading Drosophila Species. RNA 16: 1634–45.

Rozhkov NV, Zelentsova ES, Shostak NG, Evgen'ev MB. 2011. Expression of Drosophila Virilis Retroelements and Role of Small RNAs in Their Intrastrain Transposition. PloS One 6:e21883.

Sampson TR, Weiss DS. 2014. Exploiting CRISPR/Cas systems for biotechnology. Bioessays 36: 34–8.

Seth M, Shirayama M, Gu W, Ishidate T, Conte D, Mello CC. 2013. The C. elegans CSR-1 argonaute pathway counteracts epigenetic silencing to promote germline gene expression. Dev Cell 27: 656–63.

Sienski G, Dönertas D, Brennecke J. 2012. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell 151: 964–80.

Siomi MC, Mannen T, Siomi H. 2010. How does the royal family of Tudor rule the PIWI-interacting RNA pathway? Genes & development 24: 636-646.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L et al. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450:219-232.

Swanson WJ, Wong A, Wolfner MF, Aquadro CF. 2004. Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection. Genetics 168: 1457-65.

Tang T, Kumar S, Shen Y, Lu J, Wu M-L, Shi S, Li W-H, Wu C-I. 2010. Adverse interactions between micro-RNAs and target genes from different species. Proc Natl Acad Sci 1–6.

Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P et al. 2009. FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res. 37:D555-D559.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler AD. 2002. The Human Genome Browser at UCSC. Genome Res. 12:996–1006.

Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD. 2006. A distinct small RNA pathway silences selfish genetic elements in the germline. Science 313: 320-324.

Vasudevan S, Tong Y, Steitz JA. 2007. Switching from repression to activation: microRNAs can up-regulate translation. Science 318: 1931–4.

Vermaak D, Henikoff S, Malik HS. 2005. Positive selection drives the evolution of rhino, a member of the heterochromatin protein 1 family in Drosophila. PLoS genetics 1: 96-108.

Wedeles CJ, Wu MZ, Claycomb JM. 2013. A multitasking Argonaute: exploring the many facets of C. elegans CSR-1. Chromosome Res 21: 573–86.

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168: 1041-1051.

Wu CI, Yang S, Tang T. 2009. Evolution Under Canalization and the Dual Roles of microRNAs: a Hypothesis. Genome Res. 19: 734–743.

Xiao Y, Xia W, Yang Y, Mason AS, Lei X, Ma Z. 2013. Characterization and Evolution of Conserved MicroRNA through Duplication Events in Date Palm (Phoenix Dactylifera). PloS One 8: e71435.

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3’ UTRs by comparison of several mammals. Nature. 434:338–345.

Xu J, Zhang R, Shen Y, Liu G, Lu X, Wu C-I. 2013. The evolution of evolvability in microRNA target sites in vertebrates. Genome Res. 23: 1810–1816.

Yang HP, Hung TL, You TL, Yang TH. 2006. Genomewide Comparative Analysis of the Highly Abundant Transposable Element DINE-1 Suggests a Recent Transpositional Burst in Drosophila Yakuba. Genetics 173: 189–96.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution 24: 1586-1591.

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. Journal of molecular evolution 46: 409-18.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Molecular biology and evolution 22: 2472-2479.

Zhang Z, Koppetsch BS, Wang J, Tipping C, Weng Z, Theurkauf WE, Zamore PD. 2014. Antisense piRNA amplification, but not piRNA production or nuage assembly, requires the Tudor-domain protein Qin. EMBO J 33: 536–9.