RNA: Algorithms, Evolution and Design

by

Michael Schnall-Levin

A.B., Harvard University (2005)

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

Author .....
Department of Mathematics
April 29, 2011

Certified by .....
Bonnie Berger
Professor of Applied Mathematics
Thesis Supervisor

Accepted by .....
Michel X. Goemans
Chairperson, Applied Mathematics Committee

Accepted by .....
Bjorn Poonen
Chairman, Department Committee on Graduate Students

RNA: Algorithms, Evolution and Design

by

Michael Schnall-Levin

Submitted to the Department of Mathematics
on April 29, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Modern biology is being remade by a dizzying array of new technologies, a deluge of data, and an increasingly strong reliance on computation to guide and interpret experiments. In two areas of biology, computational methods have become central: predicting and designing the structure of biological molecules and inferring function from molecular evolution. In this thesis, I develop a number of algorithms for problems in these areas and combine them with experiment to provide biological insight.

First, I study the problem of designing RNA sequences that fold into specific structures. To do so I introduce a novel computational problem on Hidden Markov Models (HMMs) and Stochastic Context Free Grammars (SCFGs). I show that the problem is NP-hard, resolving an open question for RNA secondary structure design, and go on to develop a number of approximation approaches.

I then turn to the problem of inferring function from evolution. I develop an algorithm to identify regions in the genome that are serving two simultaneous functions: encoding a protein and encoding regulatory information. I first use this algorithm to find microRNA targets in both *Drosophila* and mammalian genes and show that conserved microRNA targeting in coding regions is widespread. Next, I identify a novel phenomenon where an accumulation of sequence repeats leads to surprisingly strong microRNA targeting, demonstrating a previously unknown role for such repeats.

Finally, I address the problem of detecting more general conserved regulatory elements in coding DNA. I show that such elements are widespread in *Drosophila* and can be identified with high confidence, a result with important implications for understanding both biological regulation and the evolution of protein coding sequences.

Thesis Supervisor: Bonnie Berger
Title: Professor of Applied Mathematics

Acknowledgements

I owe enormous gratitude to my advisor Bonnie Berger who accepted me into the Berger mafia and was a tremendous source of support and guidance over the last four and a half years. And I owe many thanks to the other members in Bonnie's group who were great teachers and collaborators and who made my time here a lot of fun. In particular, thank you to Patrice Macaluso, Leonid Chindelevitch, Patrick Schmid, Nathan Palmer, Ragu Hosur, Charlie O'Donnell, Jerome Waldispuhl, Po-Ru Loh, Jason Trigg and Michael Baym.

This thesis wouldn't have been possible without Norbert Perrimon, who has been an unofficial co-advisor and who opened his lab to me when I had no clue how to do biology experiments (in the interest of space, I won't describe the numerous mistakes I made during this process). It also wouldn't have been possible without the Perrimon lab manager, Richard Binari, without whom the lab would quickly grind to a halt, or the many members of the Perrimon lab who patiently offered advice and tolerated my many pranks. In particular, thank you to Yong Zhao, Rami Rahal, Shu Kondo, Michele Markstein, Richelle Sopko, Phil Karpowicz and Meghana Kulkarni. And thank you to Carlos Loya and Tudor Fulga who worked with me during my time in the Perrimon lab.

I also want to thank David Bartel for his guidance and help over the last year and a half as we have worked together. And I want to thank Olivia Rissland for being a great collaborator. Working together has been, and I hope will continue to be, a lot of fun.

I was generously supported during this time, both financially and intellectually, by the

Fannie & John Hertz Foundation. While I'm still convinced the decision to award me a fellowship was a clerical error, I have nevertheless reaped the rewards. Interacting with the other fellows has been a constant source of both humility and inspiration. In addition, the NDSEG program deserves my gratitude for their support during my first three years of graduate school.

Most importantly, I want to thank my family and friends for their endless support. In particular, I thank my family, who are my biggest fans (yes, Mom, the Nobel prize is on its way) and my greatest source of strength. And thanks to old friends who have kept me laughing and who have kept me sane over the years. Sara, you have done your share of both. And to new friends I have met in Boston over the last four and a half years who have helped me get away from the A's, C's, G's and T's once in a while. Finally thanks to Carmel: you have been my best friend over these past three years. Together we will conquer the world.

Dedicated to T., my muse.

Previous Publications of this Work

Results described in Part I were published in the Proceedings of the 25th International Conference on Machine Learning (ICML 2008) [88]. Those described in Part II were published in the Proceedings of the National Academy of Sciences [89]. Those described in Part III are under the final stages of review at Genome Research. Those described in Part IV are currently in preparation for publication.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

Biology is in the midst of a revolution. Data is being collected at a scale unimaginable only a few years ago, and experiments are being automated and performed massively in parallel. Many results that used to take years can now be achieved in weeks or even days. This has provided tremendous opportunities but also great challenges. How do we manage and analyze the deluge of data? How do we extract meaning in complex biological systems? And how do we co-opt technologies that nature has invented over billions of years of evolution for our own needs? Increasingly, mathematical modeling and computation are crucial to finding the answers.

This thesis takes a number of modest steps toward answering some of these challenges. A broad unifying theme underlying this work is the application of algorithms to the study of RNA. While seen classically as a passive molecule, shuttling information from DNA (the information storer of the cell) to proteins (the machinery of the cell), RNA has now grown to be seen as a functional molecule in its own right. Indeed in the last ten years, a class of RNA genes called microRNAs, a major subject of investigation of this thesis, has been shown to be one of the most important components of cellular regulation. The work presented here goes further in establishing the

scale of that role. Below, I give a high-level introduction to the problems addressed in the following sections of this thesis. More detailed and technical introductions for each subject are given in the chapters at the start of each section.

1.1 RNA Secondary Structure Design

In the first section of this thesis, I study the problem of RNA secondary structure design (Fig 1-1). This problem can be seen as an inverse to the RNA secondary structure prediction problem. In the prediction problem, one is given the nucleotide sequence of an RNA molecule and the goal is to find the base-pairing structure that sequence will form. In the design problem, one starts with a desired structure of base-pairing, and the goal is to find a nucleotide sequence that will take on this structure. A solution to this problem has been a long sought-after goal [48, 3, 11], as designing structure is a first step towards designing function.

I show how to abstract the RNA design problem to a problem on Stochastic Context Free Grammars (SCFGs) and in the process, define a novel problem on grammatical models. I use this abstraction to study the computational complexity of the problem. The major result of this section is a proof of the limitations of computers in solving this problem. Indeed, I show the design problem is NP-hard (a result that holds even on the simpler model of HMMs), proving that a polynomial time algorithm isn't possible unless P = NP. I then go on to offer a number of approximation approaches.

1.2 Regulatory Codes in Protein-Coding Regions

In the remaining three sections of this thesis, I study a problem of overlapping biological codes. I show that within many regions in DNA that encode the instructions on

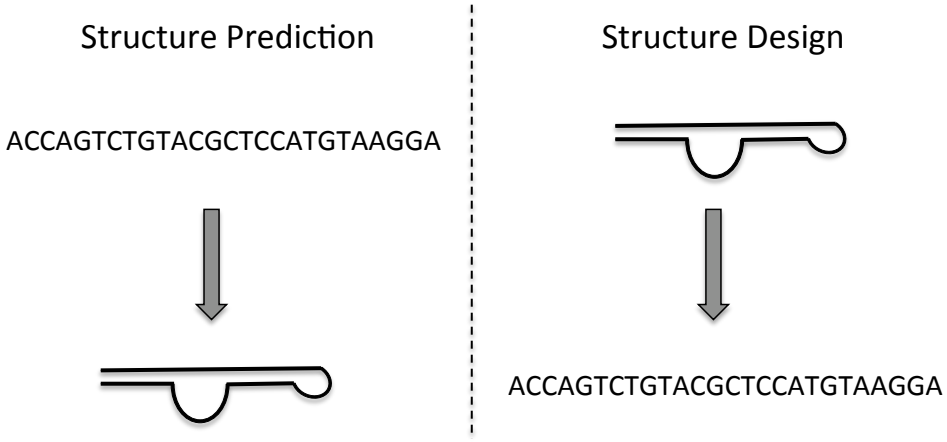

Figure 1-1: Section I of this thesis examines the RNA structure design problem, which is an inverse to the RNA structure prediction problem.

how to make a protein, there are simultaneous codes in the DNA specifying regulation of the gene (Fig 1-2). One of the difficulties in identifying such regions is that the signals for regulatory codes are masked by biases introduced by the code for making a protein. I develop algorithms that utilize the genomes from related species to infer such regions by their evolutionary signatures. In particular, I develop methods that explicitly control for the evolutionary effects of the the protein-coding aspect of such regions in order to reveal the effects of the regulatory codes alone. I use these tools to examine three situations in more detail.

MicroRNA Target Prediction in Coding Regions: In Section II, I perform a search for the genes regulated by microRNAs. microRNAs are very short RNA molecules that regulate cellular state by turning off specific genes before they can be made into proteins. Central to understanding the functional role played by microRNAs has been an effort to map out the genes regulated by them. Most effort in the

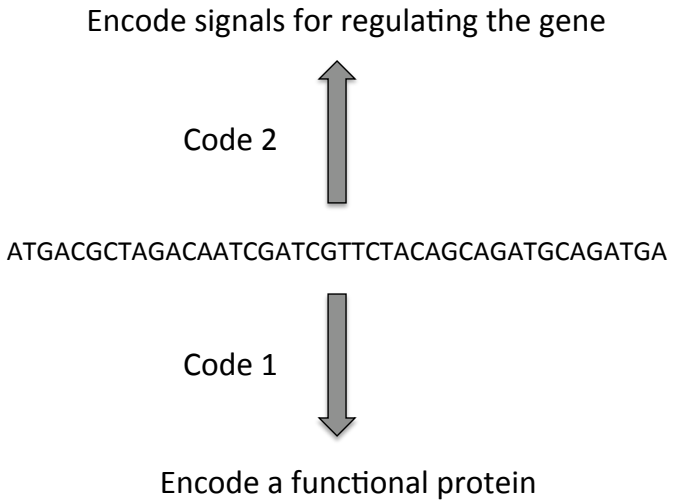

Figure 1-2: Sections II - IV of this thesis examine regions in the genome of dual functionality. Such regions both encode functional proteins and simultaneously encode regulatory signals.

search for such targets of a microRNA has focused on targeting that occurs within the non-coding regions of potential target genes. I show, however, that microRNA targeting is far more widespread in coding regions than has often been appreciated. Importantly, I also develop tools for predicting such targets and perform a number of experiments to verify my predictions.

Repeat-Mediated microRNA Targeting: In Section III, I explore a novel phenomenon involving repeated regions in the genome. Repeated sequences of many types make up a large portion of many genomes, including that of humans. The functional consequences of many of these repeats are still being explored. I show that the accumulation of some classes of repeats that occur in protein coding regions can lead to surprisingly strong targeting by microRNAs. Experiments confirm the targeting relationships predicted from genome sequence, and a phylogenetic analysis

shows how an evolutionary process leads to the accumulation of such repeats. These results demonstrate a novel mechanism through which weak regulatory signals can combine to create substantial regulation.

Widespread Non-Coding Regulation in *Drosophila* Coding Regions: In Section IV, I turn to the problem of finding more general non-coding regulatory signals in coding DNA. In addition to microRNA targeting, many other regulatory processes exist that can be specified by sequences within protein coding regions. I show evidence that conserved codes for regulating such process are surprisingly prevalent in *Drosophila* coding regions, making up as much as 10% of such regions. In particular, I examine a set of ultraconserved regions that show strikingly high levels of conservation and investigate the possible causes for this conservation.

Part I

The Structure Design Problem

Chapter 2

Grammatical Models and Structure Design

Probabilistic grammatical formalisms such as hidden Markov models (HMMs) and stochastic context-free grammars (SCFGs) have been applied towards a diverse set of problems in computational biology. Because of their intuitive representation, their power to capture some of the essential relationships present in data, and the existence of polynomial-time algorithms (e.g. the Viterbi algorithm) and practical training procedures (e.g. the Baum-Welch algorithm), these formalisms have enjoyed tremendous popularity in the past decades.

Previously, three natural problems for a grammatical model have been described: the decoding problem (given a model and a sequence, find the most likely derivation), the evaluation problem (given a model and a sequence, find the likelihood of the sequence being generated), and the learning problem (given a set of sequences, learn the parameters of the underlying model). In the first section of this thesis, I formulate another natural problem on HMMs and SCFGs, which is the inverse of the decoding problem: given a derivation and a model, find a sequence for which this derivation is

the most likely one. Because the decoding problem is solved by the Viterbi algorithm in HMMs and by the CKY algorithm in SCFGs, I refer to the problem on these two models as the Inverse-Viterbi and the Inverse-CKY problem, respectively.

The motivation for the inverse problem comes from protein and RNA design. The design of biological molecules with a desired structure is a long sought-after goal in computational biology. While a number of achievements have been made in protein structure design, the problem remains difficult [13, 81, 83]. For RNA, there has been recent interest in secondary structure design [8], and a number of fairly successful heuristics have been developed to solve this problem [48, 3, 11]. Generally, structure design can be divided into two goals: the positive-design aspect of finding a sequence that has low energy in the desired structure, and the negative-design aspect of blocking the sequence from having low energy in other structures. While some work has explored the negative-design aspect in protein structure design [13], most work has focused solely on the positive-design aspect. In RNA secondary structure design, the positive-design aspect is largely trivial (desired paired positions in the secondary structure can simply be chosen to be complementary bases) and the negative-design aspect, which involves attempting to block erroneous base pairings in other structures, is central to solving the problem.

In Chapter 3, I show that the inverse problem is NP-hard for HMMs (and as a result for SCFGs). I then give approaches for making the problem tractable in some cases. In particular, for HMMs I give a branch-and-bound algorithm. This algorithm can be shown to have fixed-parameter tractable running time: if there are K states, the emission alphabet is Σ, the path length is n, and all of the log-probabilities in the model are greater than $-B$ (so that all the non-zero probabilities in the model are greater than e^{-B}) and are defined to a precision δ, then the branch-and-bound algorithm has worst-case running time $O((2B/\delta+1)^{K-2}nK^2|\Sigma|)$, which is exponential

in the number of states but linear in the path length. I also show how to cast the problem as a simple mixed integer linear program.

The hardness proof provides a negative solution to an open problem on the existence of a polynomial time algorithm for RNA secondary structure design. A polynomial-time algorithm that only depends on the energy model for RNA secondary structure being SCFG-like, as is the case for the Zuker energy model (the most successful model curently available for RNA secondary structure prediction [107]), without making additional assumptions on the particular form of the energy model, is not possible unless $P = NP$. It does, however, remain possible that a polynomial-time algorithm could exist for certain specific energy models. I discuss this point further in Section 3.4.

Chapter 3

The Design Problem

3.1 Problem Definition

3.1.1 Definition of the Models

An HMM consists of a set $\mathcal{N}$ of K states and an alphabet Σ. The symbols in Σ are emitted on transitions between the states. The probability of emitting the symbol a when transitioning from the state s_k to the state s_l is specified by the value of the parameter $p^a_{s_k,s_l}$. Without loss of generality, there is a unique initial state S.

The normalization condition requires that

$$\sum_{s_l \in \mathcal{N}} \sum_{a \in \Sigma} p^a_{s_k,s_l} = 1 \text{ for } k = 1, \ldots, K$$

Similarly, an SCFG consists of a set $\mathcal{N}$ of K non-terminal symbols, and a set Σ of terminal symbols. The non-terminals are rewritten according to a set $\mathcal{R}$ of rewriting rules. The probability of applying each rewriting rule α is specified by the value of the parameter p_α. These parameters determine the SCFG. Without loss of generality, there is a unique starting non-terminal symbol S.

Every rule α replaces a single non-terminal with a string γ of non-terminals and terminals:

$$\alpha = N_k \rightarrow \gamma$$

Here N_k (the terminal symbol being rewritten) is referred to as the left-hand side of the rule, abbreviated as $l(\alpha)$.

The normalization condition requires that

$$\sum_{\{\alpha \in \mathcal{R} | l(\alpha) = N_k\}} p_\alpha = 1, \text{ for } k = 1 \ldots, , K$$

I don't insist that the SCFG be in Chomsky Normal Form (CNF) because in some applications (such as RNA secondary structure design), the correspondence between the design and inverse problem defined in this paper may only be natural if the SCFG is not converted to CNF.

I'll use boldface letters to indicate sequences of symbols. Thus, a state-path of length n in the HMM is written as $\boldsymbol{\pi} = \pi_1 \ldots \pi_n$, where each π_i is a state in the HMM. Such a path emits a sequence of $n-1$ emission symbols, $\boldsymbol{\omega} = \omega_1 \ldots \omega_{n-1}$ where each ω_i is a symbol from Σ. The joint probability of a state-path $\boldsymbol{\pi}$ and an emission sequence $\boldsymbol{\omega}$ is given by $\Pr(\boldsymbol{\pi}, \boldsymbol{\omega}) = \prod_{i=1}^{n-1} p^{\omega_i}_{\pi_i, \pi_{i+1}}$. It is frequently more convenient to deal with sums rather than products, which can be achieved by working in log-space, taking $q^a_{s_1,s_2} := \log(p^a_{s_1,s_2})$ and therefore $\log(\Pr(\boldsymbol{\pi}, \boldsymbol{\omega})) = \sum_{i=1}^{n-1} q^{\omega_i}_{\pi_i, \pi_{i+1}}$.

A derivation of length n in the SCFG is the successive application of rewriting rules, beginning with the starting symbol S, which generates a yield $\boldsymbol{\omega} = \omega_1 \ldots \omega_n$ where each ω_i is a symbol from Σ. The derivation can be summarized in the form of a tree $\mathcal{T}$. The joint probability of a derivation tree $\mathcal{T}$ and a yield $\boldsymbol{\omega}$ is given by $\Pr(\mathcal{T}, \boldsymbol{\omega}) = \prod_{\alpha \in \mathcal{R}(\mathcal{T})} p_\alpha$, where $\mathcal{R}(\mathcal{T})$ denotes the multiset of rewriting rules used to derive $\mathcal{T}$. As with HMMs, it is convenient to work instead with the log-probabilities,

$q_\alpha := \log p_\alpha$, which gives $\log(\Pr(\mathcal{T}, \boldsymbol{\omega})) = \sum_{\alpha \in \mathcal{R}(\mathcal{T})} q_\alpha$.

3.1.2 Definition of the Direct Problem

In the original Viterbi problem, one is given an emission sequence $\boldsymbol{\omega_0}$ from an HMM and the goal is to find the most likely state-path to have generated $\boldsymbol{\omega_0}$: the $\boldsymbol{\pi}$ that maximizes the conditional probability given the emission $\Pr(\boldsymbol{\pi}|\boldsymbol{\omega_0})$. Since $\Pr(\boldsymbol{\pi}|\boldsymbol{\omega_0}) = \frac{\Pr(\boldsymbol{\pi}, \boldsymbol{\omega_0})}{\Pr(\boldsymbol{\omega_0})}$, and $\boldsymbol{\omega_0}$ is fixed, it is equivalent to simply maximize the joint probability $\Pr(\boldsymbol{\pi}, \boldsymbol{\omega_0})$. The Viterbi problem can therefore be expressed as: given $\boldsymbol{\omega_0}$, find an element of $\arg\max_{\boldsymbol{\pi}} \Pr(\boldsymbol{\pi}, \boldsymbol{\omega_0})$ (here arg max is the *set* of all arguments maximizing the function). For an HMM with K states and an emission of length n, the Viterbi algorithm finds the best state-path using dynamic programming in time $O(nK^2|\Sigma|)$ [103].

Similarly, the direct problem for an SCFG is formulated as follows: given a yield $\boldsymbol{\omega}$, find the derivation tree $\mathcal{T}$ which maximizes the joint probability $\Pr(\mathcal{T}, \boldsymbol{\omega})$. In other words, given $\boldsymbol{\omega}$, we find an element of $\arg\max_{\mathcal{T}} \Pr(\mathcal{T}, \boldsymbol{\omega})$. The optimal derivation is referred to as the Viterbi parse of $\boldsymbol{\omega}$. For a derivation of length n in an SCFG with rewriting rules $\mathcal{R}$ in Chomsky Normal Form, the CKY algorithm finds the Viterbi parse in time $O(n^3|\mathcal{R}|)$ [26]. Modified versions of the CKY algorithm can also handle SCFGs in similar forms, such as those used in RNA structure prediction, with the same time complexity (for example see [25]).

3.1.3 Definition of the Inverse Problem

In the Inverse-Viterbi problem, a desired output of the Viterbi algorithm is known and the goal is to design an input to the Viterbi algorithm that will return this output. In mathematical terms the problem is: given a state-path $\boldsymbol{\pi_0}$, find an $\boldsymbol{\omega}$ so that $\boldsymbol{\pi_0}$ is in $\arg\max_{\boldsymbol{\pi}} \Pr(\boldsymbol{\pi}, \boldsymbol{\omega})$, or determine that none exists.

In an HMM used for structure prediction, the above definition of the inverse problem captures what it means to do structure design: one knows the structure (state-path) and tries to find a sequence that has a higher score with that structure than with any other structure. It is important to emphasize that for many $\boldsymbol{\pi}$ there will be no such $\boldsymbol{\omega}$. In fact, it can be shown that only polynomially many paths are designable [30]. This captures the intution that many physical structures are not designable: there is no sequence that will lead to an RNA molecule folding into these structures.

Upon encountering the inverse problem, the first reaction of many is to suspect that it can be easily solved by taking the most likely emission string given the desired state-path. To illustrate why this is not the case, consider the 2-state HMM shown in Figure 3-1. Say that the desired state-path to design is $B^n = B \dots B$. The most likely emission given this state-path is $a^{n-1} = a \dots a$, but when run on such a path the Viterbi algorithm will not return B^n. In fact, the only sequence that the Viterbi algorithm will return B^n on is b^{n-1}. This simple case illustrates that to design a path of all B's it is important not just to pick emissions likely given this path, but to simultaneously block other possible paths (in this case those paths containing A's). Note further that the probability of b^{n-1} being emitted from B^n at random is $(0.2)^{n-1}$. Therefore, neither picking the most likely emission sequence nor randomly generating sequences from the state-path will in general solve the Inverse-Viterbi problem with probability greater than exponentially small in the length of the state-path.

I incorporate one generalization into the definition of the problem of inverting the Viterbi algorithm, because it seems natural to the design problem. I allow constraints on the emissions that can be chosen in any position (given as the Σ_i below). The algorithms developed in this paper handle this generalization without any added complexity.

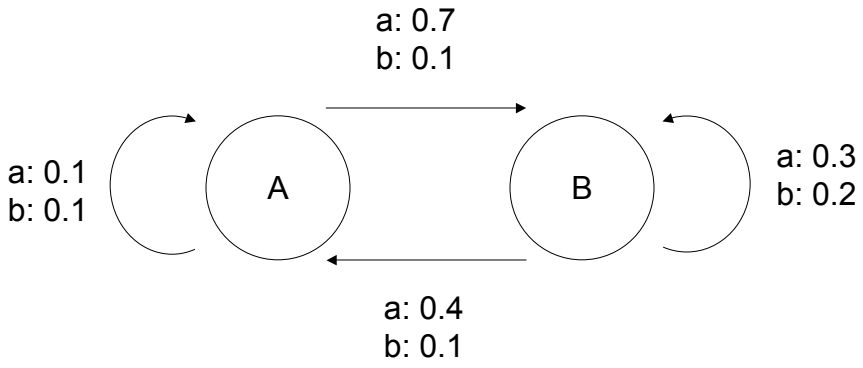

Figure 3-1: A 2-state HMM illustrating the distinction between the Inverse-Viterbi problem and the trivial problem of finding the most likely emission from a given state-path. The 2 states are A and B, while the 2 possible emissions are a and b. Each transition is marked with the possible emissions followed by their corresponding probabilities. In order to design B^n the only possible sequence is b^{n-1}, which is the least likely sequence to be produced by B^n.

INVERSE-VITERBI

Input: An HMM, a state-path $\boldsymbol{\pi_0}$ of length n and for every position i in $1, \ldots, n$ a set $\Sigma_i \subseteq \Sigma$ giving allowed emissions at position i.

Output: An $\boldsymbol{\omega}$ where each $\omega_i \in \Sigma_i$ so that $\boldsymbol{\pi_0}$ is in $\arg\max_{\boldsymbol{\pi}} \Pr(\boldsymbol{\pi}, \boldsymbol{\omega})$, or $\emptyset$ if no such $\boldsymbol{\omega}$ exists.

Similarly, the inverse problem for an SCFG requires one to find an input that corresponds to a given output. In other words, given a derivation $\mathcal{T}_0$, we would like to find an $\boldsymbol{\omega}$ such that $\mathcal{T}_0$ is in $\arg\max_{\mathcal{T}} \Pr(\mathcal{T}, \boldsymbol{\omega})$, or determine that none exists. Note that this problem only makes sense if the tree $\mathcal{T}_0$ has had all of its leaves removed (I'll call such a tree "naked"); in other words, the tree includes the specification of non-terminals but not the terminal symbols produced.

INVERSE-CKY

Input: An SCFG, a naked derivation tree $\mathcal{T}_0$ that corresponds to an emitted string of n terminals and for every position i in $1, \ldots, n$ a set $\Sigma_i \subseteq \Sigma$ giving the allowed emissions at position i.

Output: An $\boldsymbol{\omega}$ where each $\omega_i \in \Sigma_i$ so that $\mathcal{T}_0$ is in $\arg\max_{\mathcal{T}} \Pr(\mathcal{T}, \boldsymbol{\omega})$, or $\emptyset$ if no such $\boldsymbol{\omega}$ exists.

3.2 NP-hardness of the Inverse Problem

I now prove that the Inverse-Viterbi problem is NP-hard. To do so, I introduce the decision problem corresponding to Inverse-Viterbi:

DESIGNABLE

Input: An HMM and a state-path $\boldsymbol{\pi_0}$

Output: YES if there is an $\boldsymbol{\omega}$ so that $\boldsymbol{\pi_0}$ is in $\arg\max_{\boldsymbol{\pi}} \Pr(\boldsymbol{\pi}, \boldsymbol{\omega})$, otherwise NO.

An algorithm that solves Inverse-Viterbi would also solve Designable and so by proving Designable is NP-complete, I show that Inverse-Viterbi is NP-hard.

Theorem. *Designable is NP-Complete*

Proof. Clearly Designable is in NP so I just need to show Designable is NP-hard. I do so by presenting a polynomial-time reduction from 3-SAT to Designable.

In outline, the construction is achieved by creating an HMM with one component that can emit all possible non-satisfying assignments for the 3-SAT problem along with a special state outside of this component that can emit all binary strings, but that does so with smaller probability. Because this probability is small, the path consisting

of repeatedly being in the special state is only designable if a specific sequence of 0's and 1's could not possibly be emitted by the component corresponding to the 3-SAT formula. And such a sequence is, by the construction, a satisfying assignment of the 3-SAT formula.

In full detail, the construction is as follows (see Fig 3-2). Assume the 3-SAT formula consists of m variables and r clauses. The HMM consists of a begin state B, two special states S and T and $r(m+1)$ states labelled $X_{i,j}$ where $1 \leq i \leq r$ and $1 \leq j \leq m+1$. The emission alphabet consists of 0, 1, and the special symbol #. The state B transitions to either S or any of $X_{i,1}$ with equal probability, $\frac{1}{r+1}$, while emitting #. The state S transitions to itself while emitting 0 or 1, each with probability $\frac{1}{2}$. The state T transitions to itself with probability 1 while emitting #. The r sets of states $X_{i,1}, \ldots, X_{i,m+1}$ for $1 \leq i \leq r$ are arranged in independent chains, each corresponding to the ith clause, that emit all strings $\{0,1\}^m$ that do not satisfy the ith clause. Such a chain is constructed by the following: if the ith clause contains the jth variable un-negated then $X_{i,j}$ transitions to $X_{i,j+1}$ while emitting 0 with probabilty 1, if the ith clause contains the jth variable negated then $X_{i,j}$ transitions to $X_{i,j+1}$ while emitting 1 with probabilty 1, and if the ith clause doesn't contain the jth variable then $X_{i,j}$ transitions to $X_{i,j+1}$ while emitting 0 or 1 each with probability $\frac{1}{2}$. Finally, $X_{i,m+1}$ transitions to T while emitting # with probability 1.

The state-path to design is BS^{m+1}. Observe that the joint probability of this state-path and an emission sequence of the form $\#\{0,1\}^m$ is $(\frac{1}{r+1})(\frac{1}{2})^m$, and that only emissions of this form have non-zero probability for this state-path. Further observe that the only other state-path that could emit such a sequence must be of the form $BX_{i,1} \ldots X_{i,m+1}$, and the joint probability of such a sequence and such a state-path is $(\frac{1}{r+1})(\frac{1}{2})^{m-3}$ if the emission sequence contains a # followed by a non-satisfying assignment to the 3-SAT formula, but the joint probability is zero if the emission

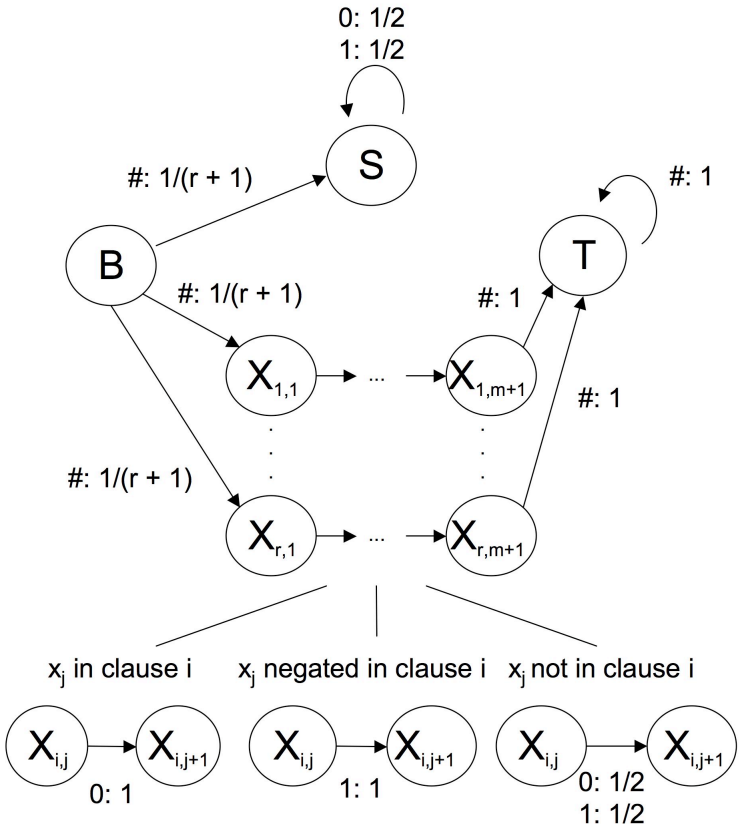

Figure 3-2: The reduction from 3-SAT to DESIGNABLE. Each transition is marked with all non-zero probability emissions followed by their corresponding probabilities.

sequence contains a # followed by a satisfying assignment. Since $(\frac{1}{r+1})(\frac{1}{2})^{m-3} > (\frac{1}{r+1})(\frac{1}{2})^{m}$, the only sequence that could design BS^{m+1} is a # followed by a satisfying assignment and therefore BS^{m+1} is designable if and only if there is a satisfying assignment to the 3-SAT formula.

The above construction is done in polynomial time, and therefore I have successfully given a polynomial reduction from 3-SAT to Designable. □

For further clarity, an example of the HMM constructed for 3-SAT instance

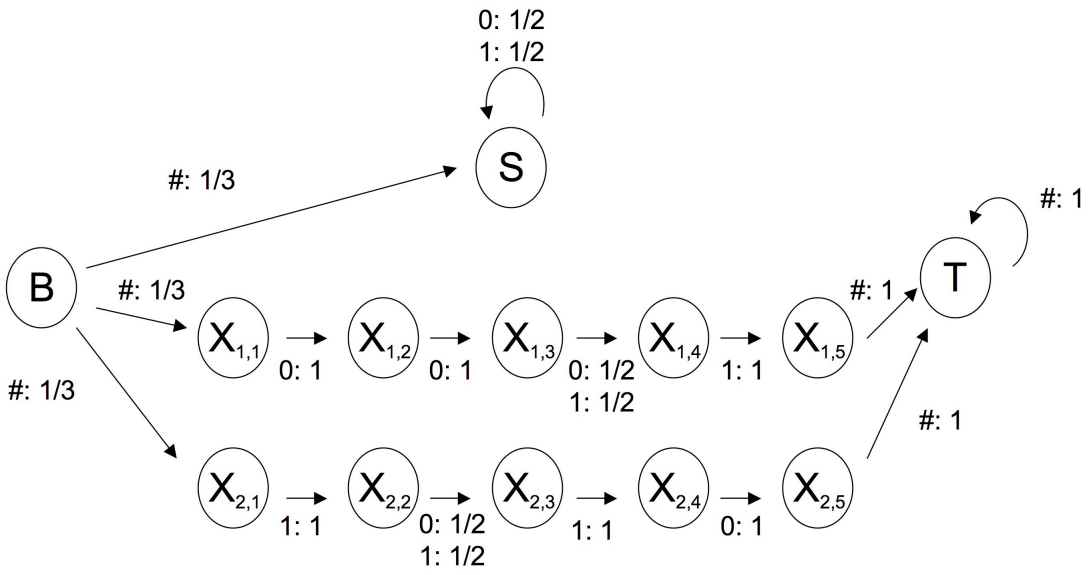

Figure 3-3: The reduction from 3-SAT to DESIGNABLE illustrated for the specific 3-SAT instance $(x_1 \vee x_2 \vee \hat{x}_4) \wedge (\hat{x}_1 \vee \hat{x}_3 ix \vee x_4)$.

$(x_1 \vee x_2 \vee \hat{x}_4) \wedge (\hat{x}_1 \vee \hat{x}_3 \vee x_4)$ is given in Figure 3-3.

Corollary 1. *Inverse-CKY is NP-hard.*

Proof. An HMM can be thought of as an SCFG with a non-terminal corresponding to each state and a terminal to each letter in the emission alphabet. Every branching rule rewrites a state as a letter and another state, so that all derivation trees are right-branching. Since the problem is hard on HMMs it is also hard on the extended class of SCFGs. □

3.3 Approximation Approaches

In this section, I give two approaches for finding a solution to the inverse problem, a branch-and-bound algorithm and a formulation of the problem as a mixed integer linear program. Both of these are derived from the same basic approach, based on a set of constraints I develop that are satisfied by an $\boldsymbol{\omega}$ if and only if it is a solution to the inverse problem. Below I first develop these constraints. Similar constraints and a mixed integer linear program can be developed for SCFGs.

3.3.1 Constraint Formulation

Conceptually, the set of inequalities for HMMs is derived by looking at how the Viterbi algorithm works and enforcing constraints on $\boldsymbol{\omega}$ so that the Viterbi algorithm is forced to return the desired state-path $\boldsymbol{\pi_0}$.

The Viterbi algorithm calculates an n by K table of values $M_{i,s}$ of the best log-probability scores for the state-path from positions 1 to i with final state s. Because of the special form of the HMM score, this table can be filled in iteratively:

(1) $M_{1,S} = 0$ and $M_{1,s} = -\infty$ for all $s \neq S$.

(2) $M_{i,s} = \max_{s'}(M_{i-1,s'} + q_{s',s}^{\omega_{i-1}})$ for $2 \leq i \leq n$ and all s.

The best state in the nth position is then read off as $\pi_n \in \arg\max_s(M_{n,s})$, and the earlier ones are read off by a traceback routine: the best state in position $n-1$ is an s' that maximized $(M_{n-1,s'} + q_{s',\pi_n}^{\omega_{n-1}})$, and so on.

From the above, one can directly read off the constraints on the emission symbol ω_i in position i for $1 \leq i \leq n-1$, that need to be satisfied in order to design a state-path with states π_i. For the Viterbi algorithm to return the desired path, we need for every state in this path to traceback to the previous state in the desired path and for the last state in this path to have the best log-probability score:

(3) $M_{i,\pi_i} + q^{\omega_i}_{\pi_i,\pi_{i+1}} \geq \max_{s \neq \pi_i}(M_{i,s} + q^{\omega_i}_{s,\pi_{i+1}})$ for $1 \leq i \leq n-1$

(4) $M_{n,\pi_n} \geq \max_{s \neq \pi_n}(M_{n,s})$

3.3.2 Branch-and-Bound Algorithm

What is particularly nice about inequalities **(1)** - **(4)** is that they allow for an inductive method for choosing possible ω_i in an emission sequence based only on the choices of ω_j for $1 \leq j \leq i-1$. This is because the inequality constraining the choice of ω_i (inequality **(3)** above) only depends on the values for $M_{i,s}$. And the values for $M_{i,s}$ only depend on the choices made for ω_1 through ω_{i-1}. This naturally leads to a branch-and-bound algorithm. Branch-and-bound algorithms are frequently useful in solving computationally hard problems. A branch-and-bound algorithm is complete (it always finds the correct answer) and frequently efficient on many problem instances.

The branch-and-bound algorithm steps through position i from 1 to $n-1$, at each step maintaining a list of emission sequences of length i that could be extended to possible length $n-1$ sequences the algorithm will ultimately return. At each step i, the algorithm forms emission sequences of length i from the emission sequences of length $i-1$ stored in the previous stage by appending possible emission symbols onto the sequences from the previous stage. In order to avoid performing an exhaustive search, at every stage the algorithm prunes the search space by applying two elimination rules. The first elimination rule ensures that for a given length $i-1$ sequence from the previous stage, an ω_i is only appended onto this sequence to form a length i sequence if the traceback constraint (constraint **(3)**) is satisfied by the choice ω_i. The second elimination rule examines pairs $\boldsymbol{\omega}$ and $\tilde{\boldsymbol{\omega}}$ of partial strings of length i that remain after the application of the first elimination rule. It eliminates $\boldsymbol{\omega}$ due to $\tilde{\boldsymbol{\omega}}$, if given that $\boldsymbol{\omega}$ can be extended to a solution to the design problem, then $\tilde{\boldsymbol{\omega}}$ must also

be able to be extended to a solution.

Specifically, the second elimination rule is based on the following observation. If for all states s, $M_{i+1,\pi_{i+1}} - M_{i+1,s}$ is at least as large under $\tilde{\boldsymbol{\omega}}$ as it is under $\boldsymbol{\omega}$ (i.e. if for all states s, the relative preference of $\tilde{\boldsymbol{\omega}}$ for π_i to state s is at least as large as that of $\boldsymbol{\omega}$), then the traceback constraints (inequality **(3)** above) on all positions j for $j > i$ and the ending constraints (inequality **(4)** above) can only be easier to satisfy when extending $\tilde{\boldsymbol{\omega}}$ than when extending $\boldsymbol{\omega}$.

It is important to note that for the case of a 2-state HMM the branch-and-bound algorithm is an exact polynomial-time algorithm. This is because there is only one $M_{i+1,\pi_{i+1}} - M_{i+1,s}$ value to compare the choices for ω_i on (there is only one state s other than π_i at every position since there are only 2 states to choose from), and so there is always a *best* choice for ω_i at every position based on the past choices.

The branch-and-bound algorithm is exact for all HMMs, but has no guaranteed worst-case running time. If one makes additional assumptions about the HMM, however, it can be shown that the algorithm also has fixed-parameter tractable running time. Specifically, I assume that all q values (the log-probabilities) satisfy $q \geq -B$ (except for the zero-probability transitions, which have value $-\infty$). Furthermore, I assume that these q values have been rounded off to precision δ.

Under these assumptions, any two values $M_{i,s}$ and $M_{i,s'}$ satisfy $|M_{i,s} - M_{i,s'}| \leq B$ (or else the difference is equal to $\pm\infty$). This follows from the definitions:
$M_{i,s} = \max_{s'}(M_{i-1,s'} + q_{s',s}^{\omega_{i-1}})$ and
$M_{i,s'} = \max_{s}(M_{i-1,s} + q_{s,s'}^{\omega_{i-1}})$.
Let the maximum in the expression for $M_{i,s}$ be attained with s_0. Then

$$M_{i,s'} \geq M_{i-1,s_0} + q_{s_0,s'}^{\omega_{i-1}}$$

$$= M_{i-1,s_0} + q_{s_0,s}^{\omega_{i-1}} + (q_{s_0,s'}^{\omega_{i-1}} - q_{s_0,s}^{\omega_{i-1}})$$

Algorithm 1 Branch-and-Bound Algorithm

Input: An HMM, a desired state-path $\boldsymbol{\pi_0}$ of length n, and for every position i in $1, \ldots, n$ a set $\Sigma_i \subseteq \Sigma$ giving the allowed emissions at position i

Output: A sequence $\boldsymbol{\omega}$ such that $\boldsymbol{\pi_0}$ is in $\arg\max_{\boldsymbol{\pi}} \Pr(\boldsymbol{\pi}, \boldsymbol{\omega})$ or $\emptyset$ if no such sequence exists.

Variables: A list L_i of all partial sequences of length i considered at the ith iteration each together with its corresponding K-vector of values $M_{i,s}$.

Initialize: $L_0 = \{(\epsilon, \mathbf{0})\}$

for $i = 1$ **to** $n - 1$ **do**
  Set $L_i = \emptyset$
  for all $(\boldsymbol{\omega^{i-1}}, \boldsymbol{v^{i-1}}) \in L_{i-1}$ and all $\omega_i \in \Sigma_i$ **do**
    Form $\boldsymbol{\omega^i} = \boldsymbol{\omega^{i-1}}\omega_i$ by concatenation
    Compute the K-vector $\boldsymbol{v^i}$ of values $M_{i+1,s}$
    Add ($\boldsymbol{\omega^i}$, $\boldsymbol{v^i}$) to L_i iff Elim Rule 1 doesn't apply
  end for
  for all $(\boldsymbol{\omega^i}, \boldsymbol{v^i}) \in L_i$ **do**
    From $\boldsymbol{v^i}$ compute and store the $(K-1)$-vector $\boldsymbol{u}$ of values $M_{i+1,\pi_{i+1}} - M_{i+1,s}$ for $s \neq \pi_{i+1}$
  end for
  Apply Elim Rule 2 to all pairs of entries of L_i
end for
for all $(\boldsymbol{\omega^{n-1}}, \boldsymbol{v^{n-1}}) \in L_{n-1}$ **do**
  if $M_{n,\pi_n} < \max_{s \neq \pi_n}(M_{n,s})$ **then**
    Remove $(\boldsymbol{\omega^{n-1}}, \boldsymbol{v^{n-1}})$ from L_{n-1}
  end if
end for
Return: An element of L_{n-1} or $\emptyset$ if L_{n-1} is empty.
Elim Rule 1: Eliminate $\boldsymbol{\omega^i}$ if $M_{i,\pi_i} + q^{\omega_i}_{\pi_i,\pi_{i+1}} < \max_{s \neq \pi_i}(M_{i,s} + q^{\omega_i}_{s,\pi_{i+1}})$
Elim Rule 2: Eliminate $\boldsymbol{\omega^i}$ due to $\tilde{\boldsymbol{\omega}}^{\boldsymbol{i}}$ if $\tilde{\boldsymbol{\omega}}^{\boldsymbol{i}} \in L_i$ has $(K-1)$-vector $\boldsymbol{u}$ componentwise $\geq$ that of $\boldsymbol{w^i}$

$$= M_{i,s} + (q^{\omega_{i-1}}_{s_0,s'} - q^{\omega_{i-1}}_{s_0,s}),$$

so that, upon rearranging,

$$M_{i,s} - M_{i,s'} \leq q^{\omega_{i-1}}_{s_0,s} - q^{\omega_{i-1}}_{s_0,s'} \leq 0 - (-B) = B,$$

and by symmetry, we also get $M_{i,s'} - M_{i,s} \leq B$, so finally, $|M_{i,s} - M_{i,s'}| \leq B$.

In particular, only $2B/\delta$ distinct non-infinite values are possible for each of the $(K-1)$ possible $M_{i,\pi_i} - M_{i,s}$ values, along with the value ∞ (the value $-\infty$ isn't possible for any acceptable $\boldsymbol{\omega}$, since this would have failed to satisfy constraint (3) in the previous step) . In the branch-and-bound algorithm, it is only impossible to remove either $\boldsymbol{\omega}$ or $\tilde{\boldsymbol{\omega}}$ (both of length i) due to the other if they are incomparable: the values one gives for $M_{i,\pi_i} - M_{i,s}$ are larger for some s and smaller for some other s. But there are only $(2B/\delta + 1)^{K-2}$ incomparable values: for two sequences that share the first $(K-2)$ $M_{i,\pi_i} - M_{i,s}$ values, any values for the last $M_{i,\pi_i} - M_{i,s}$ will make them comparable.

Therefore, in the branch-and-bound algorithm there are at most $(2B/\delta + 1)^{K-2}$ sequence possibilities that must be retained at any stage, and so with a careful implementation the running time of the algorithm is $O((2B/\delta + 1)^{K-2} n K^2 |\Sigma|)$. This bound is exponential in the number of states, but linear in the length of the structure to be designed. (This bound is independent of the base used to get the q values (log-probabilities), because changing the base introduces a factor into both B and δ that cancels.)

For SCFGs in CNF, a similar idea allows one to obtain an exact algorithm that runs in polynomial time if there are only 2 non-terminal symbols. However, the idea used above for candidate string elimination does not immediately generalize to SCFGs because of their non-linear nature; an HMM outputs one symbol per state, but a non-terminal in an SCFG can generally end up producing any substring of the output string.

3.3.3 Casting as a Mixed Integer Linear Program

One can also start with the inequalities that must be satisfied for $\boldsymbol{\omega}$ and cast the inverse problem as the problem of finding a feasible solution to a mixed integer linear program. I provide this simple formulation because it allows both practical and theoretical tools developed for integer programming to be applied directly to our problem.

The formulation as a mixed integer linear program is done by defining 0-1 variables $\epsilon_{i,j}$, where $\epsilon_{i,j} = 1$ indicates that the jth emission symbol is chosen for ω_i. Enforcing that there is only one emission choice made at every position is equivalent to requiring $\sum_j \epsilon_{i,j} = 1$ for $i = 1$ to $n-1$. Each maximum in the constraints is replaced by $\geq$, while the traceback constraints are enforced by additional equalities.

Integer Linear Program For HMMs

Objective:Feasible Solution

Variables:

$\epsilon_{i,j}$, 0-1 valued, for $1 \leq i \leq n-1$ and $1 \leq j \leq |\Sigma|$

$M_{i,s}$, for $1 \leq i \leq n$ and $1 \leq s \leq K$

Constraints:

$\sum_j \epsilon_{i,j} = 1$ for all $1 \leq i \leq n-1$

$\epsilon_{i,j} = 0$ if $j \notin \Sigma_i$ for all $1 \leq i \leq n-1$

$M_{1,S} = 0$ and $M_{1,s} = -\infty$ for all $s \neq S$

$M_{i,s} \geq \sum_j \epsilon_{i-1,j}(M_{i-1,s'} + q^j_{s',s})$ for all s, s' and all $i \geq 2$

$M_{i,\pi_i} = \sum_j \epsilon_{i-1,j}(M_{i-1,\pi_{i-1}} + q^j_{\pi_{i-1},\pi_i})$ for all $i \geq 2$

$M_{n,s} \leq M_{n,\pi_n}$ for all $s \neq \pi_n$

3.3.4 Simulations

In order to demonstrate that in practice the branch-and-bound algorithm can provide for significant time savings under some settings, I present some very simple simulations from an implementation of the branch-and-bound algorithm. The simulations were performed in the following manner. First, HMMs were randomly generated by drawing each-transition-emission pair probability from the uniform distribution and then normalizing the values, rounding off to precision $\delta = 0.01$. Separately, both arbitrary state-paths and designable state-paths were generated at random from this HMM (the latter by randomly sampling emission sequences and running the Viterbi algorithm on these sequences) and the branch-and-bound algorithm was timed on both types of instances. The algorithm ran significantly faster on arbitrary paths, the majority of which are not designable, than on arbitrary designable paths (taking milliseconds rather than seconds per run).

Figure 3-4 shows running times of simulations on random designable state-paths for different numbers of states K and path lengths n, with fixed emission alphabet of size $|\Sigma| = 20$. For each pair of K and n values, 10 HMMs were generated at random and for each of these HMMs, 10 designable paths were generated at random, as described above. The branch-and-bound algorithm was then run and the average time to design a sequence over these 100 runs was recorded. On these problem instances, the running time of the algorithm scales roughly linearly with path length n. Interestingly, while the running times initially increased with increasing K values, the running times were lower for $K = 50$ and $K = 100$ than for $K = 20$, an observation that was repeated for multiple experiments. The longest run of the algorithm took 80 seconds. A solution by exhaustive search would require examining $|\Sigma|^n$ possible sequences, which for that run would have been 20^{400} sequences. All code was implemented in Matlab and run on a 3.06 GHz Intel Xeon PC.

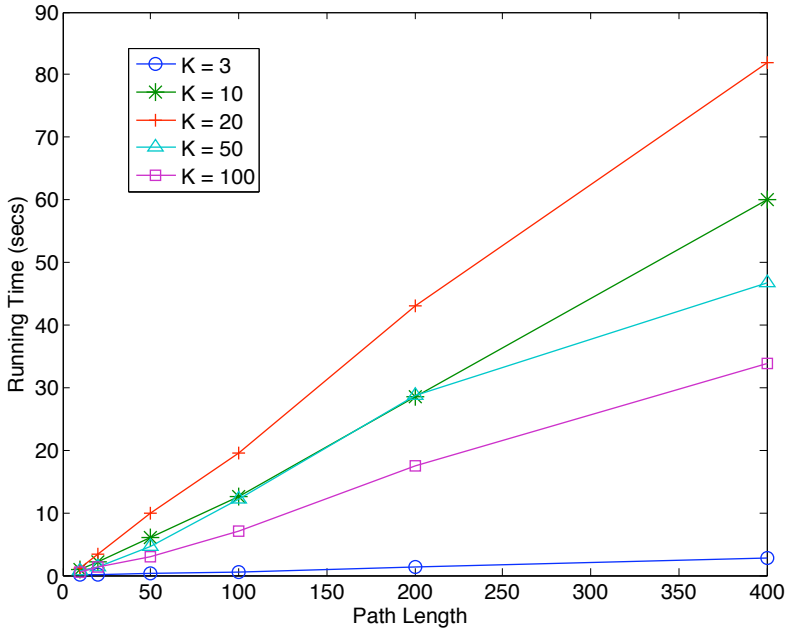

Figure 3-4: Running times of the branch-and-bound algorithm on designable paths. Simulations shown for number of states $K = 3$, 10, 20, 50, 100, path lengths $n = 10$, 20, 50, 100, 200, 400 and emission alphabet size $|\Sigma| = 20$.

3.4 Implications and Future Directions

The NP-hardness result proves that there is no algorithm with running time polynomial in both the path size and the model parameters unless $P = NP$. It does, however, leave open the possibility that there are algorithms polynomial in the path size, but exponential in the model parameters. Indeed, the Branch and Bound algorithm is a PTAS (Polynomial Time Approximation Scheme) [102] for the Inverse Viterbi problem, according to the following argument. Instead of requiring an emission string to be an exact solution to the Inverse Viterbi problem, say we require only an ϵ approximate solution. That is we require an emission string $\boldsymbol{\omega}$ for which the desired path $\boldsymbol{\pi}$ is at worst a multiplicative factor of $1 + \epsilon$ less likely than some other path given $\boldsymbol{\omega}$. Then, working in log-space, we allow a discrepancy of at most

ϵ (here we use that $\log(1+\epsilon) \leq \epsilon$). This implies that for a path of length n, we can choose $\delta = \epsilon/(n-1)$, since there are $n-1$ emissions in a path of length n. And therefore, the Branch and Bound algorithm will have a worst-case running time of $O((2B(n-1)/\epsilon+1)^{K-2}nK^2|\Sigma|)$. Given a specific model (which fixes B, K and $|\Sigma|$), this is polynomial in the path length, n.

An open question is whether the Branch and Bound algorithm can be extended to SCFGs. It can be shown that the algorithm extends to SCFGs when the branching structure of the derivation (i.e. the shape of the derivation tree, but not the states themselves) can be assumed to be the same as the desired derivation tree. Indeed, this holds for SCFGs in specific forms (such as Chomsky Normal Form), where the branching structure is essentially fixed. However, while all SCFGs can be converted to Chomsky Normal Form, in the conversion process the interpretation of a particular derivation (i.e. the secondary structure in the RNA secondary structure design problem) from the original SCFG is lost. So far, I have not been able to find an equivalent algorithm that works on more general classes of SCFGs.

From a practical perspective, however, the most useful result section in this section is probably the NP-hardness result. Heuristic algorithms for designing RNA secondary structures exist and the latest have reasonable run-times for modest sized input structures [11]. Until the hardness result, it had remained an open question whether heuristics were really necessary or if an exact polynomial-time algorithm could be found. It's possible that some of the algorithms laid out above could be used to improve on the running time of some of these algorithms. However, given that the heuristic algorithms are already pretty good, the endeavor more likely to have an impact on the problem of structure design is the improvement of the energy models themselves.

Part II

MicroRNA Target Prediction in Coding Regions

Chapter 4

MicroRNAs and Comparative Genomics

4.1 MicroRNA Biogenesis and Targeting

In the model of molecular biology that emerged during the last century, RNA was relegated to a relatively passive role, shuttling information from DNA (the information storer) to proteins (the functional molecules of the cell). It wasn't until the discovery of the ubiquity of microRNAs that the extent of the revision needed to this simplified picture was revealed. Indeed, when discovered in 1993, the first known microRNA, lin-4, was thought to be a rare and nematode-specific phenomenon [72]. The discovery in 2000 of another C. elegans microRNA, let-7, conserved across both mammals and *Drosophila* quickly led to the realization that these early examples were part of a more widespread phenomenon [84]. Now, 11 years later, it is known that microRNAs are ubiquitous across both animals and plants, with many species containing hundreds or more of these genes [45]. Together microRNAs form a rich layer of post-transcriptional regulation, and control a wide variety of biological processes [12, 45] with important

implications for a number of human diseases [66, 1].

In the following paragraphs, I briefly outline the basics of microRNA biology, highlighting their biogenesis and function. As with most descriptions of biological phenomena, the 'rules' presented here are broad brush strokes. A small number of exceptions to these general rules have already been discovered (such as a microRNA that bypasses Dicer processing [21]) and still more are likely to be discovered after this writing.

MicroRNAs are formed through one of two pathways (Fig 4-1): (i) transcription of a long (up to many kilobases) primary transcript containing extensive self-complimentarity or (ii) transcription within an intron of a protein-coding gene [85]. In the first case, the primary transcript is processed by Drosha, which cuts the transcript, leaving an $\sim$70 nucleotide self-complimentary hairpin structure. In the later case, the $\sim$70 nucleotide hairpin is formed directly from a spliced out intron. In either case, this hairpin (called the pre-miRNA) is then exported from the nucleus by the protein Exportin5. Another protein Dicer cuts the pre-miRNA hairpin and retains one of the two $\sim$23 nucleotide strands from the stem of the hairpin as the mature miRNA which is then incorporated into a complex with an Argonaute protein.

Once loaded into the complex with Argonaute, the primary function of a microRNA is to direct the post-transcriptional silencing of protein-coding genes. While early studies suggested that this occurred primarily through translational repression [2], more recent work has suggested that the largest effect is due to destabilization of the mRNA itself (possibly through multiple mechanisms [40]). The silencing of a particular gene is largely mediated by Watson-Crick base-pairing between the 5' end of the miRNA–the so-called seed region–and a target message (Fig 4-2). Target sites can be grouped by the extent to which they match the region of the miRNA seed (nucleotides 2-7). The weakest site, with base-pairing to these 6 nucleotides (6mer

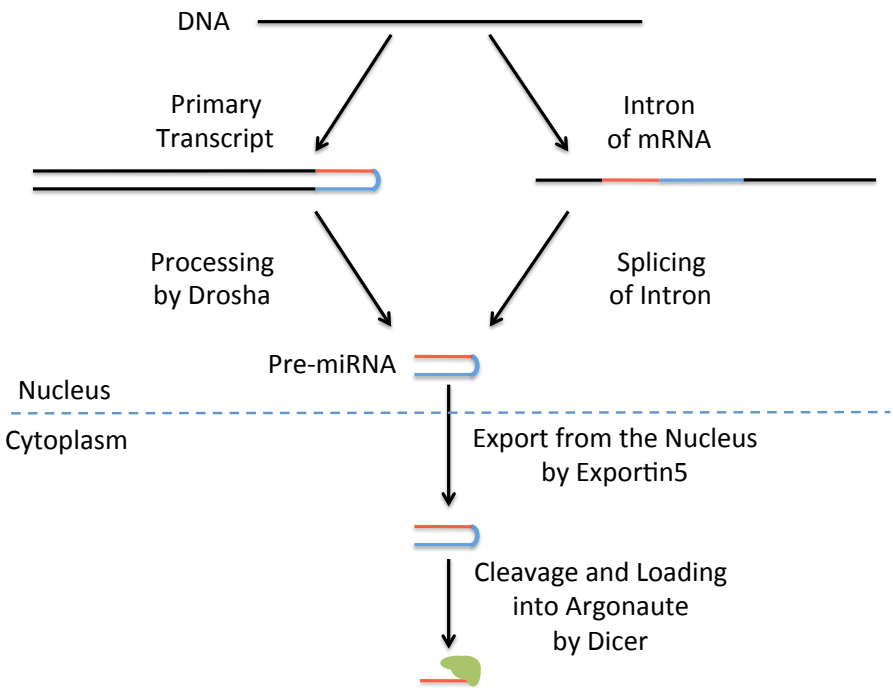

Figure 4-1: MicroRNAs are produced in one of two parallel pathways. In the first, they are processed from long primary transcripts with extensive self-complementarity, while in the second they are formed directly from the spliced introns of protein coding genes. In either case, the microRNA is subsequently exported from the nucleus, cleaved, and loaded into an Argonaute protein.

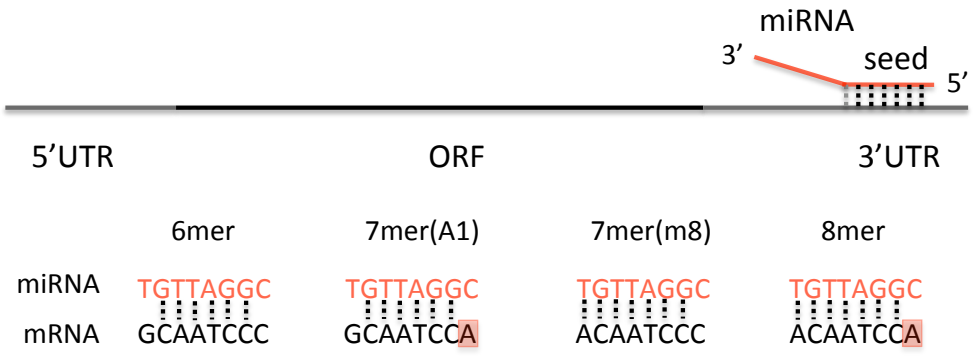

Figure 4-2: Canonical microRNA targeting occurs through base-pairing of the 5' end of a microRNA (the seed) to the 3'UTR of an mRNA. Seed matches can be grouped into different categories according to the extent of base-pairing.

site), usually confers only mild repression and is frequently augmented in functional sites. Those with an adenosine opposite nucleotide 1 (7mer-A1 site) generally confer more repression, followed by those with base-pairing to nucleotide 8 of the miRNA (7mer-m8 site), followed by those with both (8mer site) [73, 38]. Other context factors, such as local AU-content, also influence the repression mediated by an individual site [73, 38].

A focus of research in microRNAs has been, and continues to be, the identification of the genes that each microRNA targets. So far, the majority of characterized miRNA target sites have been in the 3' untranslated regions (UTRs) of mRNAs. In order to identify such 3'UTR targets, a number of target prediction tools have been designed. In addition to the presence of seed sites, these tools incorporate additional information, such as conservation, local AU-content, and the structural accessibility of the target mRNA to form a set of most likely targets [34, 67, 9, 64]. Such tools have proven to be an invaluable resource for miRNA researchers.

An open question has been the extent of targeting that does not fit this canonical

pattern, and in particular the extent of biologically relevant targeting outside of 3'UTRs. In this section of the thesis, this question is examined in detail. Through a combination of computational and experimental results, I provide evidence that such targeting is far more extensive than previously appreciated

4.2 Comparative Genomics

The last ten years have seen the mapping of genome sequences of an impressive number of species (at the time of this writing, genomes are available for 46 vertebrates, 15 insects, and 6 nematodes among numerous others). Availability of these sequences has enabled an extremely fruitful avenue of research: comparison of sequence data across species to infer the molecular evolutionary history. Knowledge of this history can provide, among other things, a map of the selective pressures exerted on different genomic features and can be an incredibly powerful tool for investigating biological function from sequence data alone. A foundation for this work is a simple model that I describe below: that of the evolution of an allele in a population under the simultaneous influence of selection and drift. Below, I give a brief derivation of the formula under this model for the fixation of a newly acquired allele under selection. While in most work in comparative genomics (including much of the work in this section) the application of this formula frequently gets reduced to something like 'Conserved equals functional', there are some subtleties in the formula that are important to remember for interpreting results. This will be particularly important for interpretation of the results in Part IV of this thesis.

The model taken is a simple one: a locus with 2 possible states A and B, in a population of N individuals of a hermaphroditic diploid organism (the hermaphroditic assumption is merely one of convenience–using a 2 sex model would only require a small correction to the population size used [36]). The individuals are assumed to

be randomly mating, so that all of the genotype frequencies can be assumed to be in Hardy-Weinberg equilibrium [36]. The genotypes are assumed to have the following relative fitnesses:

Genotype	Relative Fitness
AA	1
AB	1 - hs
BB	1 - s

The parameter s is called the selection coefficient and h the heterozygous effect. The heterozygous effect h can be used to capture all possible dominance relations (including heterozygous advantage, as found in the sickle-cell locus in some African groups [36]). Here, we take the simplifying assumption that $h = 1/2$, i.e. that the fitness effects are completely additive. Reproduction is considered to proceed in discrete generations and can be described as a resampling of the $2N$ alleles (with replacement) weighted according to the relative fitness of the genotypes.

Allele A is assumed to begin at a given frequency in the population, p. Starting from this point, different aspects of the dynamics of the frequency of A can be interrogated. Despite its simplicity, the model can capture a number of phenomena that are informative of the evolutionary history. For example, one can interrogate the distribution of allele frequencies within a single population of an allele under a given selection coefficient. From this, the frequency spectrum of given alleles can be used to identify selection [36]. One can also ask what the effect of selection on alleles has on other alleles physically linked to that allele, which is a question I will return to in more detail in Section IV. But in this section, I examine a simple question: what is the probability that allele A comes to take over the entire population? The derivation is given in outline and loosely follows [36].

The evolution of frequency of the A allele can be modeled as a random walk. If

the population size, N, is sufficiently large, then it is possible to model the process continuously in what is called the diffusion approximation [36]. The probability of fixation (that allele A ultimately takes over the population) is written as $F(p)$. Then the probability of fixation from frequency p is given by the probability of going to a new frequency p' and reaching fixation from that point:

$$F(p) = \int_{p'} T(p, p')F(p')dp' \tag{4.1}$$

where $T(p, p')$ is the probability of going from p to p' in one reproduction step. For sufficiently large N (and under the additional assumption that selection is not so large, so that $s^2 << \frac{1}{N}$), $T(p, p')$ will be concentrated around p. In this case, it is possible to approximate $F(p')$ with a Taylor expansion:

$$F(p') = F(p) + (p' - p)\frac{dF(p)}{dp} + \frac{(p - p')^2}{2}\frac{d^2F}{dp^2} \tag{4.2}$$

Putting this back into into the 4.1 gives:

$$F(p) \approx \int_{p'} T(p, p')(F(p) + (p' - p)\frac{dF(p)}{dp} + \frac{(p - p')^2}{2}\frac{d^2F}{dp^2})dp' \tag{4.3}$$

To derive a formula for $T(p, p')$, it is necessary simply to obtain the probability of drawing a single copy of A from a Bernoulli process. The probability of drawing an A is the probability of drawing the genotype AA (which is present at fraction p^2 and has relative fitness 1) plus one-half the probability of drawing the genotype AB (which is present at fraction $2p(1 - p)$ and has relative fitness $1 - s/2$.) This probability is given by:

$$\frac{p - \frac{s}{2}(p(1 - p))}{1 - s(1 - p)} = p + \frac{s(p)(1 - p)}{2} + O(s^2) \tag{4.4}$$

Therefore, keeping only leading terms in s, the mean change in frequency of A in a

single step is given by $\frac{p(1-p)s}{2}$ and the variance is given by $\frac{p(1-p)}{2N}$. If N is large, the Gaussian is a good approximation for the binomial distribution. Therefore one can take $T(p,p') \approx \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-((p-p')-\mu)^2}{2\sigma^2}}$, where $\mu = \frac{p(1-p)s}{2}$ and $\sigma^2 = \frac{p(1-p)}{2N}$. Putting this into 4.1 gives:

$$\begin{aligned} F(p) = F(p) \int_{p'} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-((p-p')-\mu)^2}{2\sigma^2}} dp' + \quad & \frac{dF(p)}{dp} \int_{p'} (p'-p) \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-((p-p')-\mu)^2}{2\sigma^2}} dp' \\ & + \frac{d^2F}{dp^2} \int_{p'} \frac{(p-p')^2}{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-((p-p')-\mu)^2}{2\sigma^2}} dp' \\ = F(p) + \mu \frac{dF(p)}{dp} + \frac{(\sigma^2+\mu^2)}{2} \frac{d^2F}{dp^2} & \end{aligned} \tag{4.5}$$

Again the assumption is that $s^2 << \frac{1}{N}$ and so $\mu^2 << \sigma^2$. Using this and canceling common terms, the differential equation becomes:

$$0 = s\frac{dF(p)}{dp} + \frac{1}{2N}\frac{d^2F}{dp^2} \tag{4.6}$$

Solving the above differential equation on $F(p)$ with boundary conditions $F(0) = 0$ and $F(1) = 1$ gives the equation:

$$F(p) = \frac{1 - e^{-Ns/N}}{1 - e^{-2Ns}} \tag{4.7}$$

Of particular interest is the case of a new mutant A allele arising, which begins with frequency $\frac{1}{2N}$. This new allele then has the probability of fixation given by:

$$F(\frac{1}{2N}) = \frac{1 - e^{-s}}{1 - e^{-2Ns}} \tag{4.8}$$

This is one of the most important equations in comparative genomics. A number of observations on this equation can be made. (1) if $s < 0$ (i.e. the A allele causes lower fitness) then the probability of fixation decays exponentially in s. Therefore,

for significantly deleterious mutations (those with $|s| >> \frac{1}{2N}$), such a mutation will never get fixed. (2) Conversely, if $s > 0$ (i.e. the A allele is beneficial) then the probability of fixation grows with s. (3) If $|s| << \frac{1}{2N}$ then the probability of fixation is $\approx \frac{1}{2N}$ as in the neutral case (when $s = 0$). Regions under purifying selection (those where biological function is being conserved across related species) can then be found because mutations are accumulating slower than they would under neutral evolution. Conversely, regions under Darwinian selection (those where a new biological function is acquired by one species) evolve more quickly than expected under a neutral model. Both of these effects are thresholded, with the threshold set by the inverse population size $\frac{1}{2N}$. When population dynamics, such as bottlenecks, are considered the population size needs to be replaced with an effective population size N_e that may differ from the current size of the population of species, but the formulas above can continue to be applied in most cases.

The work described in Section II of my thesis focuses on finding instances of short motifs (microRNA seed sites) subject to purifying selection. In this case, the use of equation 4.8 is largely reduced to 'Conserved equals functional', (with the additional caveat that the level of function required to keep a feature from acquiring mutations is defined by the inverse effective population size). Therefore, the problem reduces to finding sequences that can be confidently scored as having mutated slower than they would have by chance. While in principle one might like to have an explicit model for neutral evolution and score features according to this model, in practice it is usually far better to take a more empirical approach. When scoring whether a class of features (such as instances of a sequence motif) is more conserved than expected under a neutral model, the empirical approach is to form a background set of features with similar sequence properties to the class under study (such as permuted versions of the sequence motif) and compare the conservation of the feature to the

conservation of the background set. The main advantage of this approach is that it doesn't make assumptions that (even when subtly wrong) could cause gross mis-estimates of the expected conservation of a feature under a model of neutral evolution. In particular, both subtle mutational biases as well as overlapping weak selection from other functional features make it difficult to very accurately model the expected rate of neutral evolution of a class of features from first principles. The work in Section II takes such an empirical approach. For the work described in Section IV, there are more subtle effects of equation 4.8 that need to be considered (such as the interaction with nearby loci under selection). This will be further examined in that section.

Chapter 5

Target Prediction in ORFs

5.1 Motivation

In order to analyze miRNA targeting in ORFs, I sought to use a conservation-based approach. The goals for this approach were two-fold: (i) to compare the extent of conserved targeting in ORFs to that in 3'UTRs, and (ii) to provide a tool to guide researchers in identifying the most likely ORF targets. Some preferential conservation of miRNA seed sites has previously been observed in ORFs in both vertebrates [73] and *Drosophila* [95], but it has been difficult to fully analyze the extent of conserved ORF targeting and to provide confident predictions of individual ORF targets. The main difficulty encountered in such an effort is that traditional techniques based on conservation are not designed for application to coding DNA.

To assess the evidence for selection on putative miRNA binding sites within coding regions, I developed an algorithm (called MinoTar) to score sequences within coding DNA for evidence of preferential conservation. Searching for preferentially conserved regulatory sequences within coding regions presents a unique challenge: coding DNA is already highly conserved due to selective pressures at the protein level, and such

selection may in general be far stronger than any additional selection at the nucleotide level. Furthermore, selection at the protein level causes highly biased conservation patterns, influenced both by the form of the genetic code and by codon bias.

Any method for locating preferentially conserved nucleotide sequences in coding regions must therefore remove the level of baseline conservation due to selective pressures at the amino acid level. In this work, a conservative approach (for lack of a better term) was taken. An underlying assumption was made that selection at the amino acid level dominates the conservation signal. According to this assumption, the conservation of nucleotide sequences was scored in a model where the amino acid sequence choices at all species in the alignment were fixed according to their observed values. To score nucleotide conservation, a model for codon evolution was learned for all branches in the phylogeny, and then random samples of codon evolution were generated using this model while holding the amino acid evolution fixed as observed.

Such an approach not only accounts for non-uniform levels of amino acid conservation, but also for the average effects of codon bias because the rates of codon evolution are measured empirically rather than assumed to follow a neutral model. Alternative choices for the model were possible. One alternative was to attempt to use knowledge of protein structure to directly model selection at the amino acid level. There are three reasons why I believe such an approach is inferior (at least for the clades considered in this work). (1) In *Drosophila* and (to a somewhat lesser extent) in vertebrates, conservation of amino acids beyond that expected under neutral evolution is extensive. (2) The current knowledge of most proteins is low, and so any model that attempted to capture the exact selective pressures on all amino acids would surely be wrong in a number of places. (3) Perhaps most importantly, the conservative approach taken doesn't result in the exclusion of most conserved sequences from scoring highly. There is enough codon choice for most amino acids (only 2 out

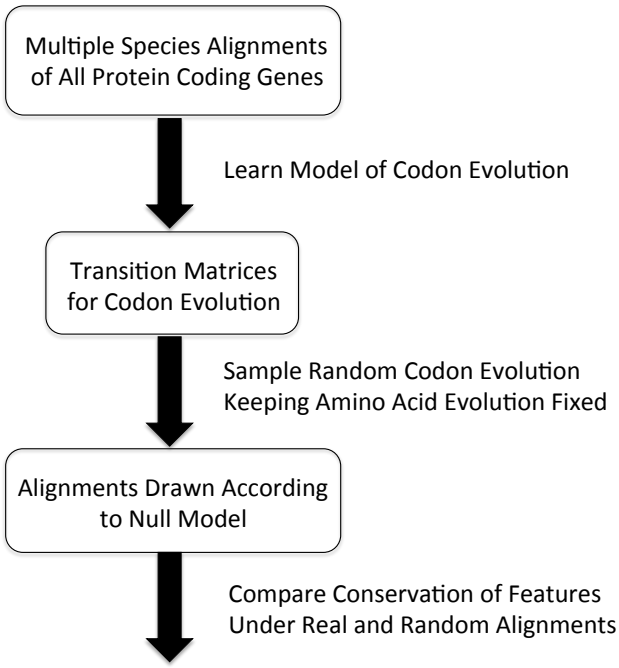

Figure 5-1: The stages of the algorithm for assessing nucleotide conservation in protein-coding regions.

of 64 have no codon choice), and there are enough species under consideration, that high confidence prediction of conserved nucleotide sequences is still possible under such a model.

5.2 Algorithm for Scoring Conservation

The central component of the method is an algorithm for forming randomized alignments that preserve amino acid evolution but contain randomly sampled codons according to an evolutionary model. These randomized alignments then form the basis for the microRNA target predictions made in this section, and also for some of the broader patterns of codon conservation studied in Section IV. A number of details of the methods used to evaluate and score targets are given in Section A.1.

The algorithm to generate these alignments proceeds in three main stages (Fig 5-1). In the first stage, a model for codon evolution is learned (Fig 5-2). The algorithm begins with a multiple species alignment of coding genes from which are removed all overlapping non-coding features that could confound analysis (including overlapping 3'UTRs and 5'UTRs from any transcript). First, at avery amino acid position in all alignments, the ancestral codons and amino acids are inferred using maximum parsimony (in the case of ties, this can result in multiple maximum parsimony trees). Then on every edge in the phylogeny, a 64 by 64 matrix is formed with all counts of codon transitions from parent species to child species (in the case of multiple maximum parsimony trees, counts from each possible tree are all given equal fractional weights– i.e. with 2 trees, each transition has weight 1/2). These counts are then used to form transition probabilities for codon evolution given fixed amino acid evolution. These probabilities are formed in both directions (forward evolution from parent to child and reverse evolution from child to parent) along all edges of the phylogeny. If we take C_1 and C_2 as the codons for origin and ending species respectively, AA_1 and AA_2 as their amino acids, and $N(C_1, C_2)$ as the transition counts obtained, then the probability is given by $Pr(C_1 \rightarrow C_2) = \frac{N(C_1,C_2)}{\sum_{AA(C)=AA_2} N(C_1,C)}$. In order to avoid artifacts from small counts in very rare amino acid transitions, counts are supplemented with small numbers of pseudocounts.

In the next stage, the algorithm draws randomized codon alignments according to the transition probabilities calculated in the first stage (Fig 5-3). The procedure is simple. First, the maximum parsimony amino acid sequences inferred on the ancestral species are retained (in the case of multiple maximum parsimony trees, one of these best trees is chosen at random for each random alignment). Then the codon in the reference species is fixed, and codons are drawn at random according to the previously calculated transition probabilities. The tree structure allows for these draws to be

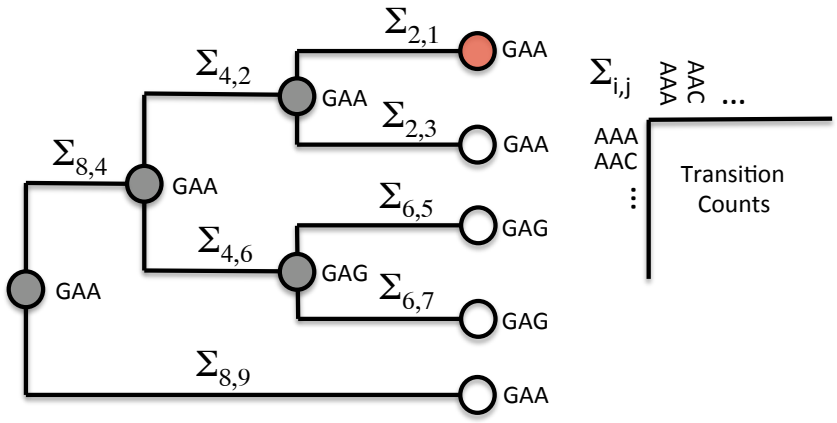

Figure 5-2: In the first stage of the algorithm, a model is learned for codon evolution. First amino acids and codons for ancestral species are inferred according to maximum parsimony. Then counts for codon transitions given amino acid transitions are used to define transition probabilities.

made independently (Fig 5-3). 1000 such random alignments are made for each sequence.

In the final stage of the algorithm, the conservation of any arbitrary feature (such as a sequence motif) can be scored (Fig 5-4). This is done by the following. First an instance of a feature is scored by the number of species it is conserved to in the real alignment N_{real}. Then a score is calculated for all of the random alignments, given again by the number of species the feature is conserved to in those alignments. Say these are labeled N_i, where i is the index of the random alignment. These scores can then be used to derive an empirical p-value: $p = \frac{\sum_i (N_{real} \leq N_i)}{S}$, where S is the number of sampled alignments. In addition to the p-value, a second score – the minimum possible p-value – is calculated. This is calculated by first finding N_{max}, the greatest number of species the feature could have possibly been conserved to while retaining the amino acid evolution, and then calculating the p-value as above with N_{max} instead of N_{real}. This second score gives a measure of the limit of information that can be inferred from conservation at that location, and allows for the exclusion of features

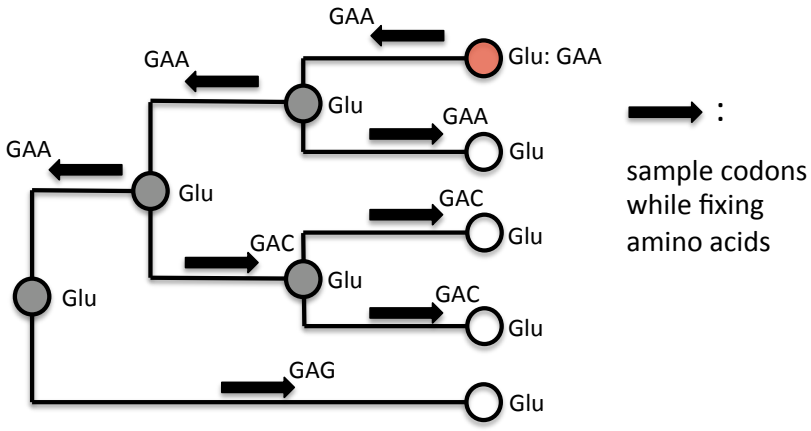

Figure 5-3: In the second stage of the algorithm, codons are sampled according to the transition probabilities learned in the first stage. The inferred amino acids are fixed for all species, along with the codon in the reference species. Codons are then evolved by sampling. Because of the tree structure, sampling can be performed efficiently, by proceeding in the direction shown by the arrows in the figure.

where nothing can be inferred: for example when there is no freedom of codon choice and nucleotide conservation is completely accounted for by amino acid conservation. In such a case, a feature can be said to be neither conserved nor non-conserved. No information can be extracted from its level of conservation.

The algorithm for sampling random alignments has a number of beneficial features:

1. It scales efficiently with the number of species. The sampling only requires $O(N)$ operations for N species (since all operations are proportional to the number of edges on the phylogenetic tree and there are $2N - 1$ edges).

2. Because the calculations are all made empirically according to the sample alignments, scoring of the conservation of any arbitrary motifs, including non-exact motifs and structural motifs, is easy to perform. This would be both difficult and computationally inefficient to calculate exactly.

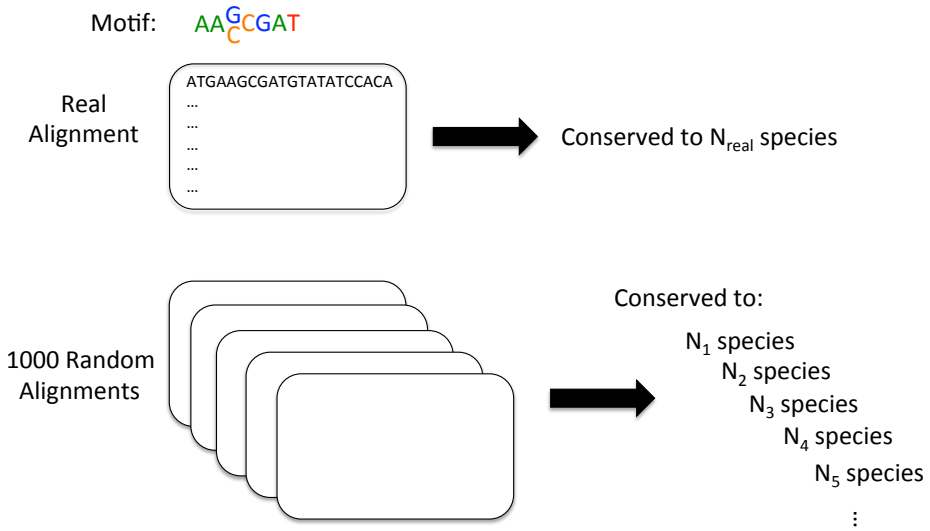

Figure 5-4: In the final stage of the algorithm, the conservation of a desired feature is scored. The number of species a feature is conserved to in the real alignment is compared to the number of species it is conserved to in the random alignments. A p value can then be calculated, indicating the level of surprise that should be attributed to the conservation of the feature.

3. It is easily extendible to more complicated models of codon evolution. While, as currently implemented, the model of evolution treats codon positions separately, the procedure could easily be generalized to allow for codon evolution to depend on the flanking sequences, or other features, as well.

All of the code has been written in Python and is freely available.

Chapter 6

Target Prediction Results

6.1 Computational Results

6.1.1 miRNA Seed Sites in ORFs Are Highly Conserved

To begin, I ran the MinoTar algorithm on every instance of all 8mers within *Drosophila* protein coding genes, and found that miRNA seeds accounted for the majority of the most highly conserved 8mers. To evaluate this, I formed a conservation score for each 8mer, given by the fraction of instances with p-value below 0.05. Most 8mers had scores close to 0.05 expected by chance, but a small group showed significantly more conservation (Fig 6-1). Grouping the top 26 most conserved 8mers into 10 motifs, it was observed that 8 of these 10 motifs correspond to miRNA seeds and 2 to other unknown motifs (Table 6.1).

Next, I verified that the increased conservation observed for miRNA seeds could not be explained by other sequence characteristics of these seeds. To do this, five sets of 8mers were formed: seeds for conserved miRNAs (those miRNAs largely present across all 12 *Drosophila* species), and four control sets, (i) reverse complements of these seeds, (ii) 8mers with identical dinucleotide content as these seeds, (iii) seeds

Motif	Annotation
AAGACTGA AGACTGAA	miR-14
CTGTGATA TGTGATAC ACTGTGAT TCTGTGAT TGTGATAT TGTGATAA TTGTGATA	K-box (miR-2a/6/11/13/308)
ACATTCCA CATTCCAA AACATTCC	miR-1
ATGAACAA ATGGACAA ATGTACAA	unknown
TCTAGTCA CTCTAGTC TCTAGTCT	miR-279/286/996
ACATATCA	miR-190
ACCAAAGA	miR-9
TGCATTTA GCATTTAG	miR-277
GTCAATTA	unknown
CAGTATTA AGTATTAA AAGTATTA	miR-8

Table 6.1: The top 26 most conserved 8mers form 10 motifs, 8 of which correspond to miRNA seeds.

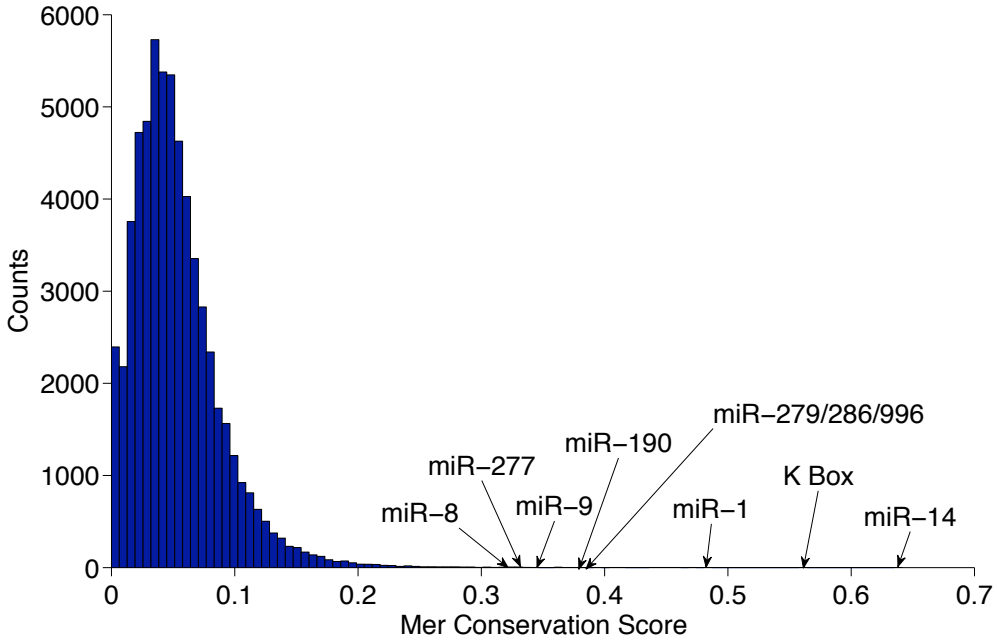

Figure 6-1: Histogram of conservation scores for all 65,536 8mers. Nearly all of the top conserved 8mers correspond to miRNA seed sites.

for non-conserved miRNAs (those not found beyond the melanogaster subgroup) and (iv) seeds for human miRNAs. The cumulative distributions for these 5 sets, as well as the set of all 8mers are plotted in Figure 6-2. While seeds for conserved miRNAs showed a very significant bias to be highly conserved, the four control sets all behaved similarly to the set of all 8mers, providing further evidence that the increased conservation seen was indeed evidence of selection on miRNA target sites.

In addition, I found that the MinoTar algorithm could produce very high-confidence target predictions. To test this, I investigated the effect of an increasingly stringent cutoff on the confidence of predicted sites at that cutoff. I pooled seeds for all conserved miRNAs together and calculated the fraction of instances of these seeds with p-values below each cutoff, repeating the same for the 4 control sets of 8mers defined above. For each cutoff, the signal-to-background ratio was calculated by dividing the fraction of conserved instances in a given set by the background fraction of instances

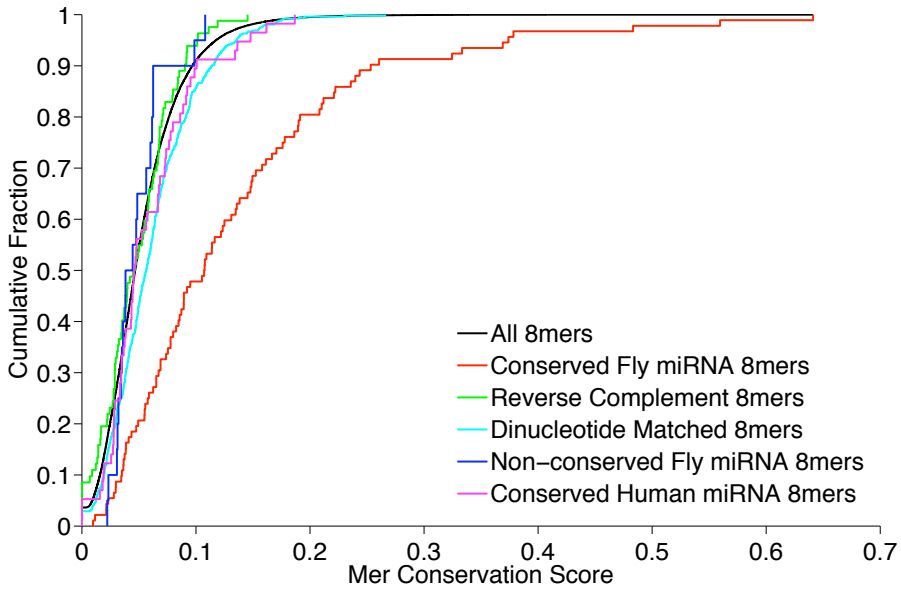

Figure 6-2: Cumulative plot of scores for different sets of 8mers. Only seeds for conserved microRNAs show preferential conservation while all the control sets show conservation similar to the set of all 8mers

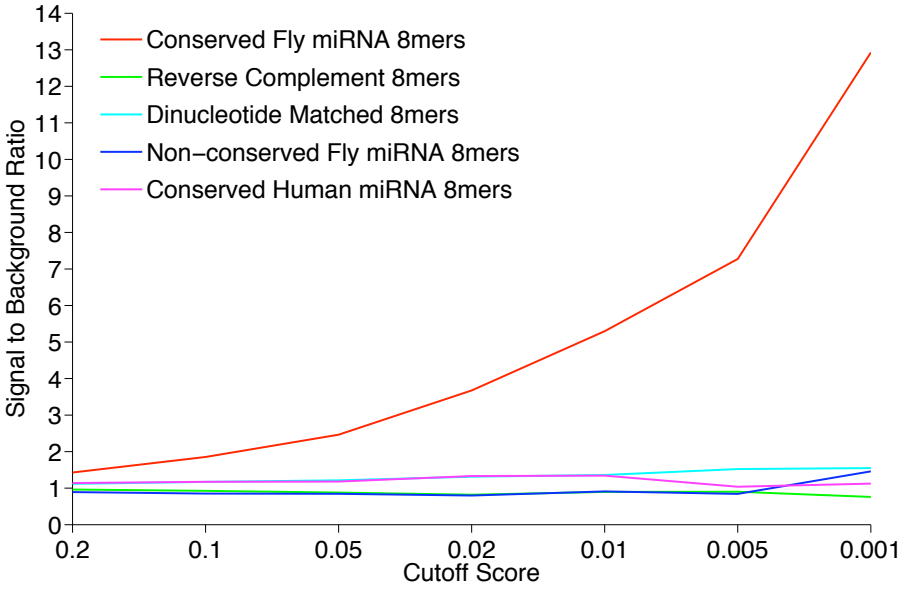

Figure 6-3: Imposing increasingly stringent conservation cutoff results in higher signal-to-background ratios for the set of *Drosophila* conserved miRNA seeds, while control sets behave as background at all cutoffs

reaching that cutoff (Fig 6-3). At the most stringent cutoffs, a signal-to-background ratio of over 10 (confidence $> 90\%$) could be achieved, producing a set of hundreds of very high confidence targets. The 4 control sets showed signal-to-background ratios near 1 at all cutoffs.

Finally, I evaluated the evidence for increased functionality of miRNA seed sites, when such sites were accompanied by additional 3' pairing to the miRNA. Following [38], I grouped seed sites according to the extent of additional 3' pairing starting at different positions within the miRNA and calculated the fraction of such sites with p-value below a cutoff of 0.05. Signal to background was computed by comparing this fraction to the expected fraction of seed sites reaching such a cutoff as explained in the previous section. I found that while the majority of conserved seed sites are not accompanied by extensive additional 3' pairing, those sites with such pairing showed some evidence of increased conservation (Fig 6-4). In particular, as has been observed in 3'UTRs [38], those seed sites with contiguous 4mer or 5mer base-pairing beginning at positions 1214 of the miRNA showed a statistically significant increase in conservation (number conserved seed sites: 2094 total; 4mers: 178 vs. 145 expected, $p < 0.002$; 5mers: 51 vs 38 expected, $p < 0.01$; Chi-Square test).

6.1.2 Extent of Targeting in ORFs, 3'UTRs and 5'UTRs

Next, I compared the extent of evidence for miRNA targeting in ORFs, 3'UTRs and 5'UTRs. For 3'UTRs and 5'UTRs, I performed an analysis similar to that done previously in 3'UTRs [65]: conservation of a miRNA seed site was judged by the number of species within the alignment the site was conserved to, and background conservation rates were estimated by nucleotide-matched background sets. The 3'UTR sets of conserved sites were nearly identical to those in [65], but repeating the analysis allowed for an evaluation of the strength of selection separately for 8mers and 7mers

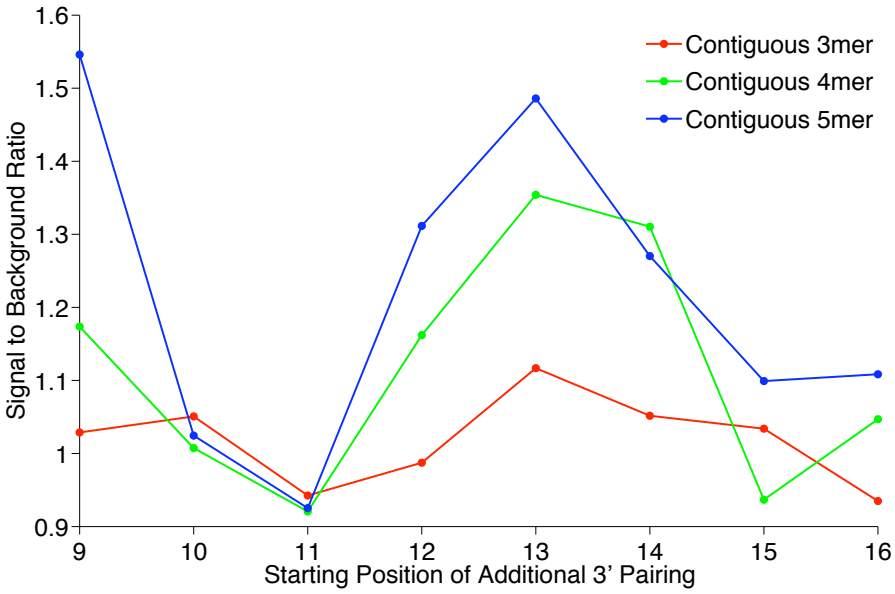

Figure 6-4: Signal to background, giving conservation of seed sites accompanied by 3' base-pairing to the miRNA, starting at different positions within the miRNA, as compared to the background conservation of seed sites.

within each region (7mers excluded those contained within 8mers).

I characterized the level of conservation by the fraction of potentially conserved sites that showed conservation above background level. The fraction of ORF sites preferentially conserved was about 60% that in 3'UTRs, while the fraction of sites in 5'UTRs was about 60% that in ORFs. Though miRNA seed sites are denser in AT-rich 3'UTRs than in ORFs, because of the significantly larger size of ORFs and smaller size of 5'UTRs (roughly 2×10^7 total bases in ORFs, 7×10^6 total bases in 3'UTRs and 2.5×10^6 total bases in 5'UTRs), the number of preferentially conserved sites in ORFs and 3'UTRs was very similar ($\sim$ 7000 sites within 3'UTRs versus $\sim$ 6500 in ORFs), while the number in 5'UTRs was significantly smaller ($\sim$ 700 sites) (Fig 6-5).

The extent of conservation of miRNA seed sites in ORFs varied considerably between different miRNAs. Those with the most conserved sites tended to be well known

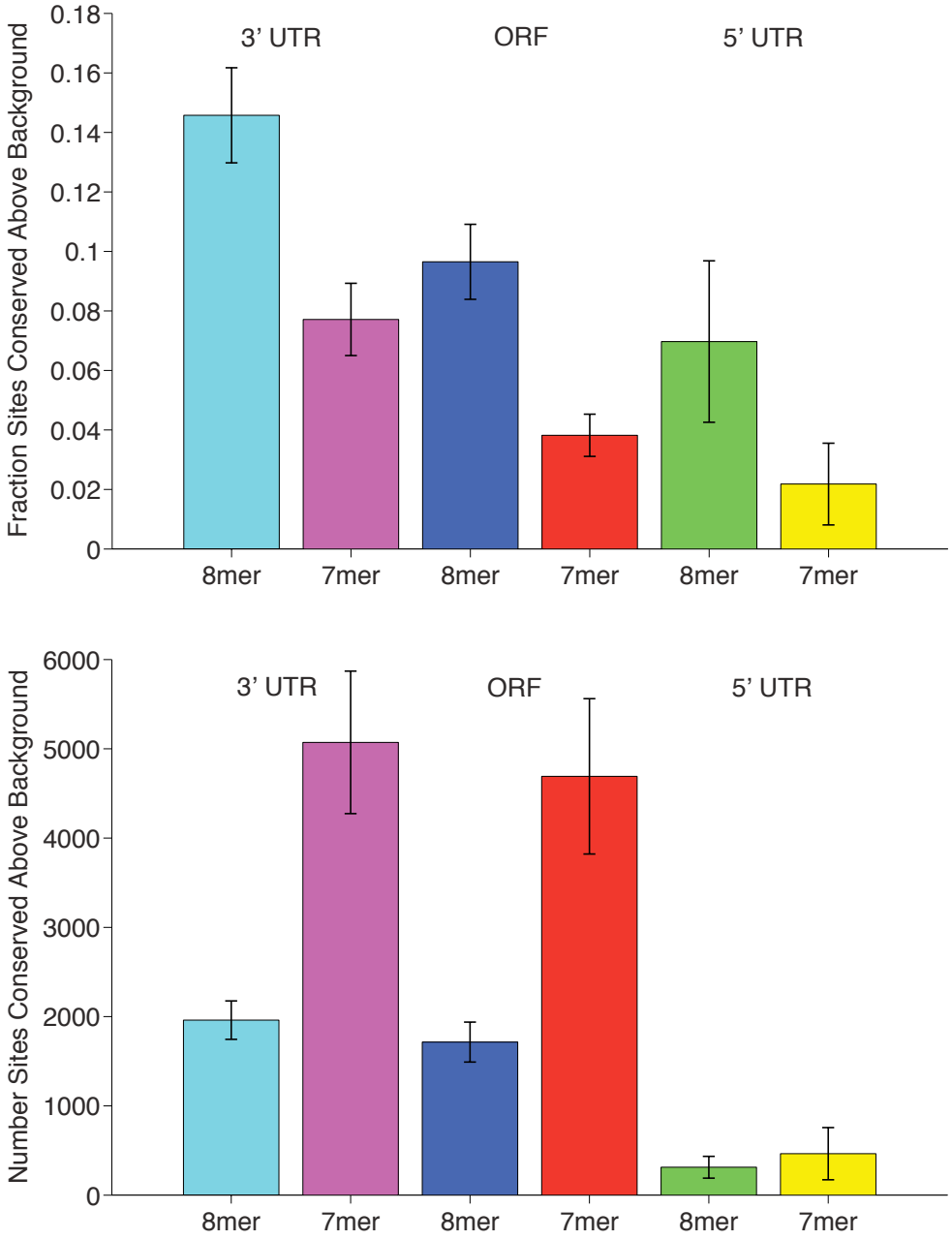

Figure 6-5: The scale of conserved miRNA targeting in 3'UTRs, ORFs and 5'UTRs. Top panel: Fraction of sites conserved above background for both 8mers and 7mers in 3'UTRs, ORFs and 5'UTRs. Bottom panel: Number of predicted sites above background for 8mers and 7mers in 3'UTRs, ORFs and 5'UTRs. Error bars show standard deviation in the estimates obtained from sampling of background sets.

and highly expressed miRNAs, while those with few conserved sites tended to be more recently discovered and expressed at lower levels. This suggested that the levels of conservation correlate with the number of targets each miRNA has acquired, with more widely and highly expressed miRNAs tending to have acquired more targets. To provide support for this, I compared the level of conservation of individual miRNA seeds within ORFs, 3'UTRs and 5'UTRs. If conservation levels reflect the number of acquired targets, then I surmised that the miRNAs with the most conserved targets in each region should largely agree. I looked at 8mer seeds and chose a cutoff for ORFs and 3'UTRs that gave predictions at 60% confidence (p = 0.05 for ORFs, conservation to 8 out of 12 species for 3'UTRs). For 5'UTRs I used the same cutoff as for 3'UTRs, as 60% confidence was not possible to achieve. For each seed, I compared the fraction of instances with conservation above background in the different regions (Fig 6-6). The level of conservation in ORFs and 5'UTRs were both highly correlated to those in 3'UTRs: mean conservation above background in ORFs and 5'UTRs was significantly higher for the top 50% most conserved miRNA seeds in the 3'UTR than for the bottom 50% (ORFs: 0.14 versus 0.02, $p < 10^{-7}$; 5'UTRs: 0.10 versus 0.01, $p < 3 \times 10^{-4}$; Mann-Whitney test). To ensure that I was not observing biases in conservation independent of selection due to miRNA targeting or conservation due to overlap with un-annotated 3'UTRs, I repeated the same procedure with promoter sequences (500 nucleotides upstream of the transcription start site). Within promoters, miRNA seed sites overall showed no preferential conservation and the top 50% most highly conserved seeds in the 3'UTR showed no tendency to be more conserved than the bottom 50% (-0.01 versus -0.014, $p = 0.32$; Mann-Whitney test).

To further analyze the set of predicted ORF targets, I formed a set of genes for each miRNA that had either a 7mer or 8mer conserved to the 60% confidence level. I compared these target lists to the set of predicted 3'UTR targets from the

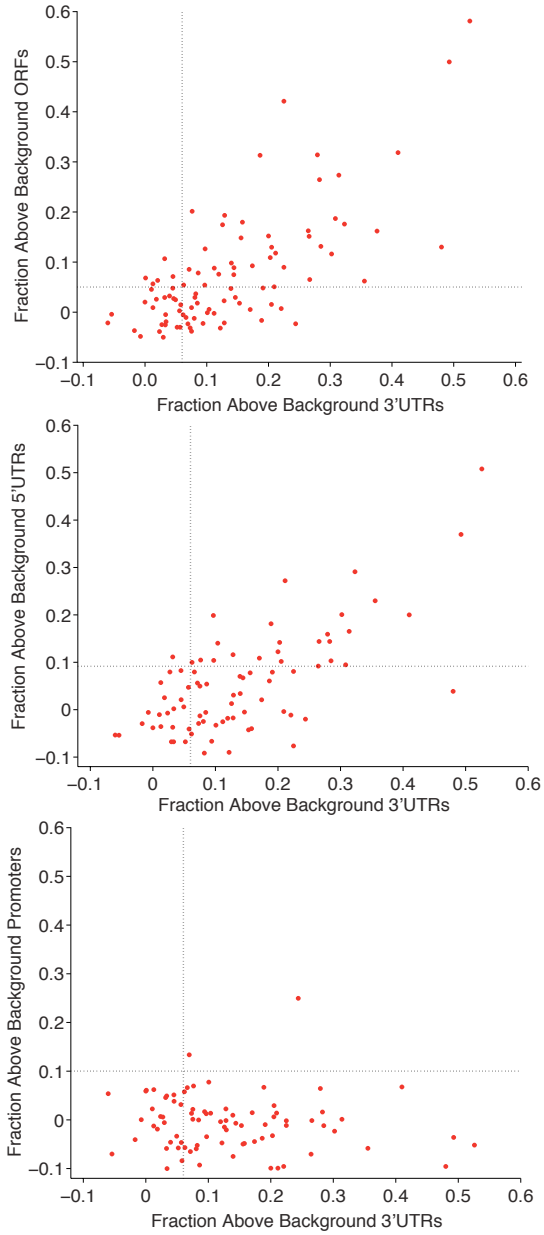

Figure 6-6: MicroRNA seeds showing the highest level of conservation in 3'UTRs tend also to be the most conserved in ORFs and in 5'UTRs, but not in promoter regions. Shown are the fractions of 8mer sites conserved above background at 60% confidence compared between 3'UTRs and ORFs (top panel), 3'UTRs and 5'UTRs (middle panel), and 3'UTRs and promoters (bottom panel). Dotted lines show the cutoff for conservation above background equal to the maximal amount by which any miRNA was conserved below background.

TargetScan website [65]. While most (> 97%) of the predicted ORF target genes were not predicted by TargetScan to be targeted in their 3'UTR by the same miRNA, genes with a predicted ORF target for one miRNA were significantly more likely to contain a predicted site in their 3'UTR for that miRNA than for any other miRNA (1.7 fold more likely, $p < 2 \times 10^{-7}$, Chi-Square test), providing evidence for some degree of simultaneous targeting by sites in both the 3'UTR and ORF. For each of the predicted ORF target sets, I searched for significantly enriched GO terms using Amigo Term Enrichment software [14]. I found that 37 out of 94 miRNAs had target sets with significantly enriched terms (versus 70 out of 94 for 3'UTRs), including 21 of the top 25 miRNAs with the most conserved ORF targets. Enrichment terms were significantly more likely to be shared by predicted targets of the same miRNA in ORFs and 3'UTRs, than by two different miRNAs (1.9 fold more likely, $p < 10^{-12}$ Chi-Square test).

6.1.3 Extent of Conserved Targeting in Mammalian ORFs

To see if the same results extended to mammals, I applied the MinoTar algorithm to a multiple alignment of vertebrate species with human. As in *Drosophila*, miRNA seed sites in ORFs accounted for most of the top-conserved 8mers, were highly conserved overall, and highly conserved sites could be discriminated above background with high (> 90%) confidence, while control sets all behaved similarly to the set of all 8mers (Fig 6-7, Fig 6-8, Fig 6-9; Table 6.2). I also compared the predicted conservation of miRNA seed sites in human ORFs to the conservation observed in multiple species alignments of human 3'UTRs and 5'UTRs (Fig 6-10). Interestingly, compared to *Drosophila*, there was a larger dropoff in the level of conserved targeting between 3'UTRs and ORFs and between ORFs and 5'UTRs. The fraction of conserved sites above background in ORFs was about 40% that in 3'UTRs while the fraction of

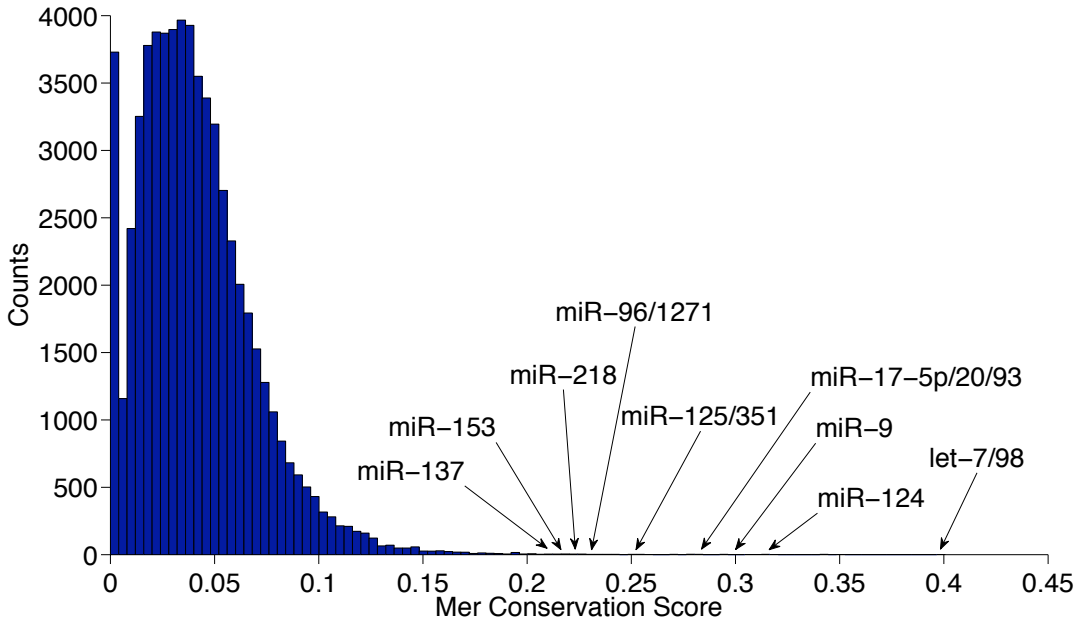

Figure 6-7: Histogram of conservation scores for all 65,536 8mers. A majority of the top conserved 8mers correspond to miRNA seed sites.

conserved sites above background in 5'UTRs was low, and not reliably above zero. The level of conservation was again highly correlated between miRNA sites in 3'UTRs and ORFs, and to a small extent between 3'UTRs and 5'UTRs, but sites in the promoter region showed no similar relationship (Fig 6-11). (Fraction reaching 60% confidence cutoff for the top 50% conserved miRNA seeds in 3'UTRs versus bottom 50%; ORFs: 0.10 versus 0.01, $p < 5 \times 10^{-9}$; 5'UTRs: 0.05 versus -0.003, $p < 0.003$; Promoters: 0.008 versus -0.003, $p = 0.1$; Mann-Whitney test).

The final target predictions from the MinoTar algorithm for both *Drosophila* and mammals are available at [53].

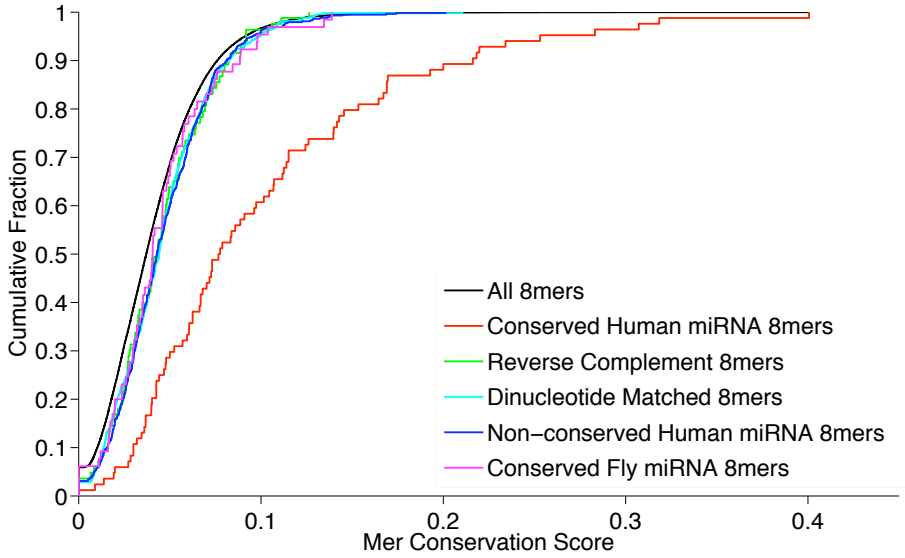

Figure 6-8: Cumulative plot of scores for different sets of 8mers. Only seeds for conserved microRNAs show preferential conservation while all the control sets show conservation similar to the set of all 8mers.

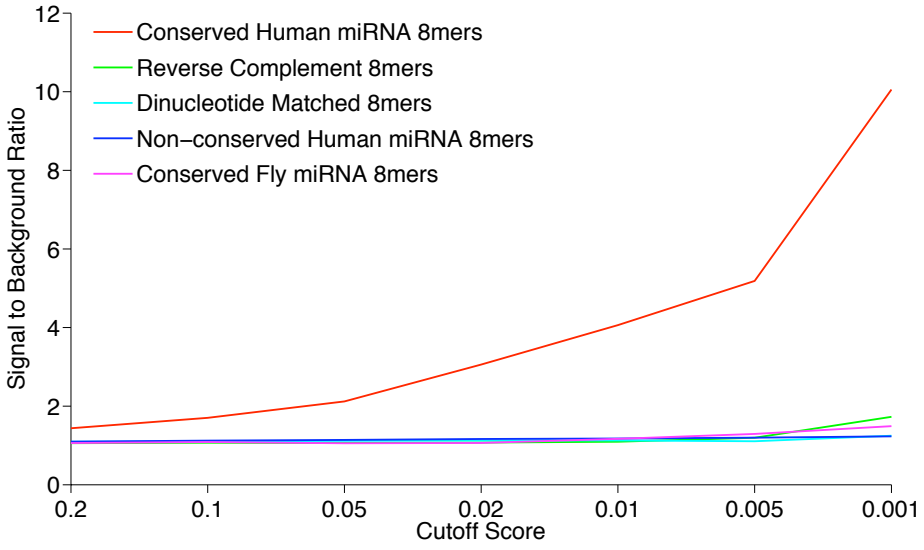

Figure 6-9: Imposing increasingly stringent conservation cutoff results in higher signal-to-noise ratios for the set of human conserved miRNA seeds, while control sets behave as background at all cutoffs.

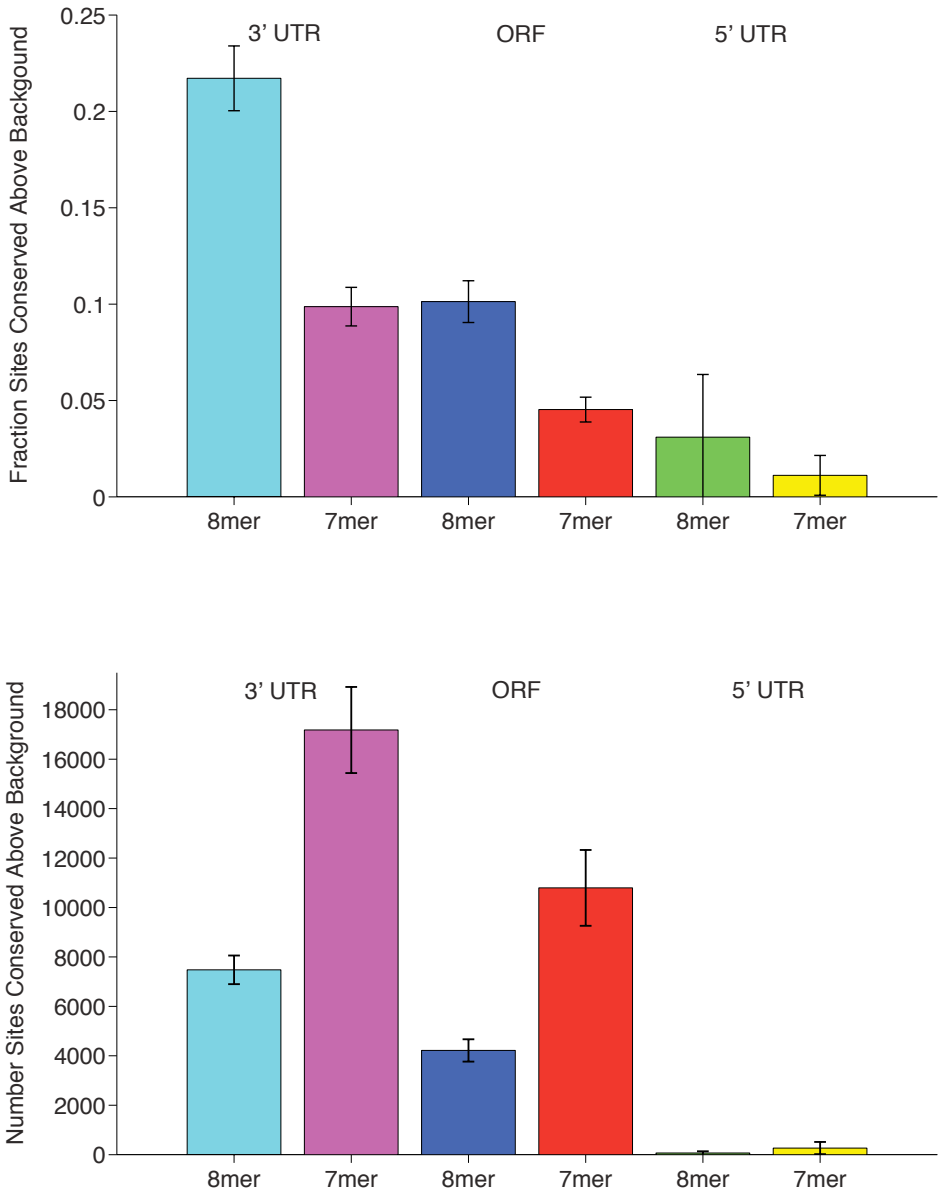

Figure 6-10: The scale of conserved miRNA targeting in 3'UTRs, ORFs and 5'UTRs in humans. Top: Fraction of sites conserved above background for both 8mers and 7mers in 3'UTRs, ORFs and 5'UTRs. Bottom: Number of predicted sites above background for 8mers and 7mers in 3'UTRs, ORFs and 5'UTRs. Error bars show standard deviation in the estimates obtained from sampling of background sets.

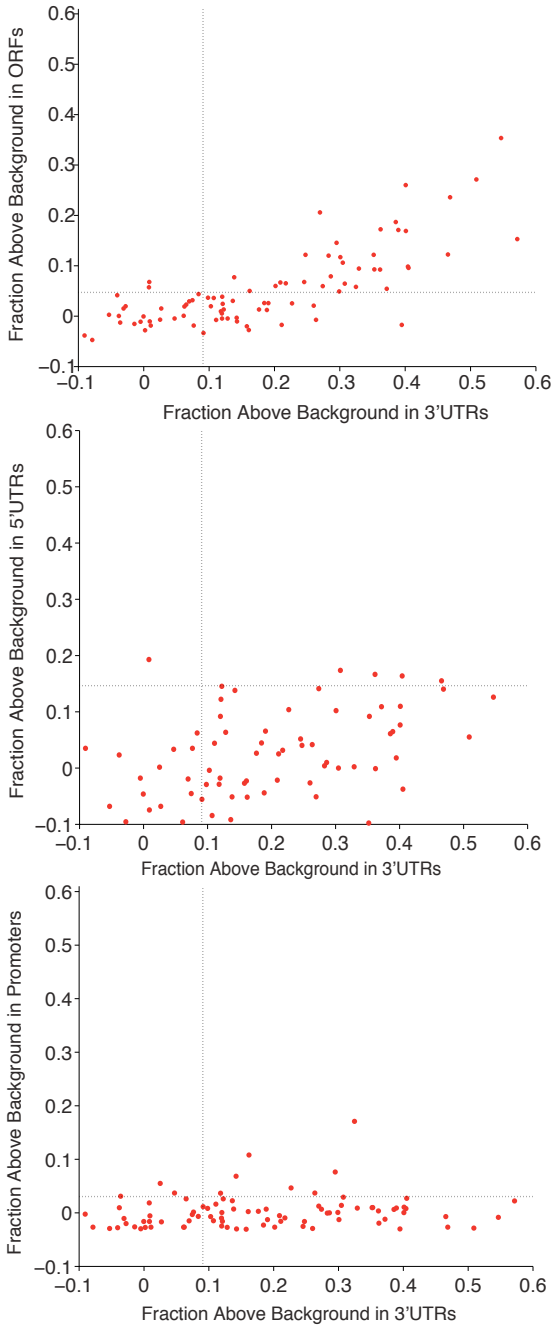

Figure 6-11: MicroRNA seeds showing the highest level of conservation in human 3'UTRs tend also to be the most conserved in ORFs and to a very small extent in 5'UTRs, but not in promoter regions. Shown are the fractions of sites conserved above background at 60% confidence cutoff between 3'UTRs and ORFs (top panel), 3'UTRs and 5'UTRs (middle panel), and 3'UTRs and promoters (bottom panel). Dotted vertical and horizontal lines show the cutoff for conservation above background equal to the maximal amount by which any miRNA was conserved below background.

Motif	Annotation
CTACCTCA	let-7/98
TTACCTCA	
CTACCTCC	
ACTACCTC	
CCTACCTC	
GCTACCTC	
TCTACCTC	
CTACCTCG	
TACCTCAG	
TACCTCAT	
ATGGCGGC	unknown
ATGGCGGA	
TGGCGGCG	
GGCGGCGG	
GGCGGCGC	
AGGCGGCG	
CGGCGGCG	
GTGCCTTA	miR-124
ATGCCTTA	
GTGCCTTG	
AGTGCCTT	
AAGTGCCT	
TGCCTTAA	

Motif Ctd...	Annotation
ACCAAAGA	miR-9
AACCAAAG	
TACCAAAG	
GCACTTTA	miR-17-5p/20/93
CGCCGCCG	unknown
CCGCCGCC	
GCCGCCGC	
GCGCCGCT	
GCCGTCGG	
TTAGCTCG	unknown
CTCAGGGA	miR-125/351
GCGCGCTT	unknown
TTTGATGA	unknown
CGCACGCG	unknown
CGCACTCG	
GTGCCAAA	miR-96/1271
TGTAAATA	unknown
AAGCACAA	miR-218
CTATGCAA	miR-153
GAGGTAGG	unknown
TCGCGCCG	unknown
AGCAATAA	miR-137

Table 6.2: A list of the top 18 highest scoring motifs in humans listed by descending score. miRNA seeds account for 4 of the top 5 and 10 of the top 18 motifs.

6.2 Experimental Results

6.2.1 Predictions Recover Targets with High Confidence

Having seen strong computational evidence for microRNA targeting in ORFs, I next investigated this targeting experimentally. To test whether predicted ORF target sites could confer substantial down-regulation, I selected 6 genes with highly conserved seed sites for 3 different miRNAs: miR-1, miR-8 and miR-6 (a member of the K-Box family). Since one of the genes, Arp87c was predicted to be targeted by both miR-1 and miR-8, in total I tested 7 miRNA-target pairs. For each of these genes, I cloned the ORF into a reporter plasmid and measured down-regulation upon co-expression with the targeting miRNA in S2R+ cells. Briefly, each ORF was fused with one of either a Myc or FLAG epitope tag, while the same ORF with synonymous point mutations in the miRNA seed site was fused with the other tag (Fig 6-12). This allowed for the ORF with wild type miRNA seed site (ORF-WT) and with the mutated site (ORF-Mut) to be co-transfected and simultaneously visualized on a western blot using 2 different secondary antibodies. For quantification, the ratio of ORF-WT to ORF-Mut when transfected with miRNA was compared to the ratio when transfected with a control plasmid. As an additional control, all ORFs were co-transfected with non-targeting miRNAs. All experiments were repeated at least twice under each epitope tag. More details on the experimental setup are given in Section A.1.

Down-regulation was detectable in 5 out of 7 miRNA-target pairs while no non-targeting miRNAs caused down-regulation of any of the genes. 4 out of 7 miRNA-target pairs showed down-regulation greater than 25% and 2 out of 7 showed greater than 50% down-regulation. The strongest effect was seen for one of the miR-1 targets (CG8494) that contained three seed sites for miR-1 (one 8mer and two 7mers). In

order to observe the effect of multiple sites within a single gene, I systematically mutated away all 3 sites for this gene and tested the down-regulation for all 8 possible mutated configurations. Regression on the observed log fold-change showed that down-regulation was largely additive between the targeting sites ($R^2 = 0.91$) with all 3 sites conferring significant down-regulation (Site 1: 13%±9%; Site 2: 36%±8%; Site 3: $45\% \pm 6\%$; errors give 95% confidence intervals). This suggests that as in 3'UTRs, genes with multiple sites are more likely to be strongly regulated by a miRNA.

6.2.2 Predicted Targets are Preferentially Down-regulated

In order to examine the scale of miRNA targeting in ORFs on endogenous targets, I transfected S2R+ cells with either miR-1 or a control plasmid and compared expression levels using a whole genome microarray. While microarray analysis only allows one to observe effects at the mRNA and not the protein level, recent results have suggested that changes at the mRNA level may capture a significant portion of the effects caused by miRNA targeting ([38], [29]). I looked at the effect of miR-1 overexpression on 4 categories of genes: (i) genes with a miR-1 ORF site predicted by the MinoTar algorithm, (ii) genes with any miR-1 seed site in the ORF, (iii) genes predicted to be targeted by miR-1 in the 3'UTR by TargetScan, and (iv) genes with any miR-1 seed site in the 3'UTR (Fig. 6-13). Predicted ORF targets were significantly down-regulated versus the set of all genes (12%, $p = 2 \times 10^{-16}$; K-S test), significantly more down-regulated than genes with a non-conserved seed in their ORF (12% versus 6%, $p < 2 \times 10^{-4}$; K-S test) and showed mean down-regulation about half as strong as predicted 3'UTR sites (12% versus 24%). These results suggest that while weaker than 3'UTR targeting, targeting in ORFs is widespread, and that conservation can preferentially recover functional ORF sites. Additionally, I looked at the down-regulation of genes with a miR-1 seed site in their 5'UTR. These genes

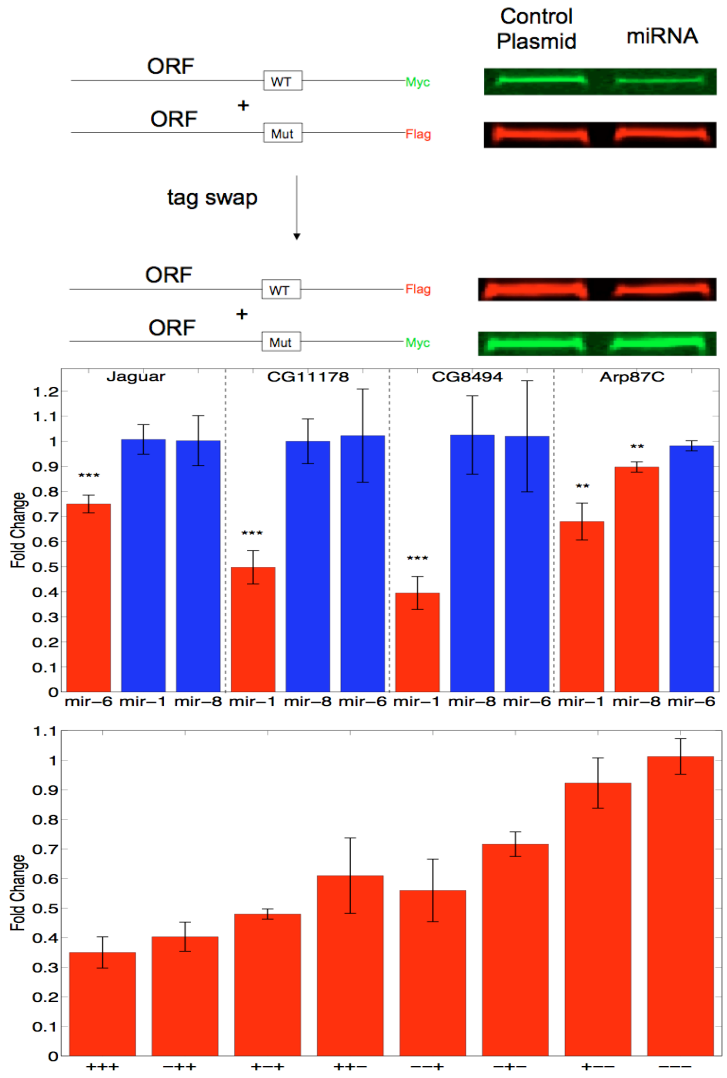

Figure 6-12: Top panel: Illustration of the experiment. Each ORF with wild type miRNA target site and the same ORF with mutated site were placed under different epitope tags and co-expressed with either a control plasmid or miRNA, and then run on Western blot. Quantification was made by comparing the ratio of the two channels under miRNA versus under control plasmid. Shown are the bands from a test of CG11178 targeting by miR-1. Middle panel: Down-regulation of target genes. Shown are the fold changes of targets under targeting miRNAs (red bars) as well as under miRNAs not predicted to target the genes (blue bars). Error bars show standard deviation, asterisks denote p-values (Students t-test; **: $p < 0.01$, ***: $p < 0.001$). Bottom panel: Effect of multiple target sites. Shown is down-regulation of CG8494 by miR-1 with all 8 combinations of the 3 predicted sites (WT sites are marked as + and mutated sites as -), averaged over 3 separate experiments. Error bars give standard deviation.

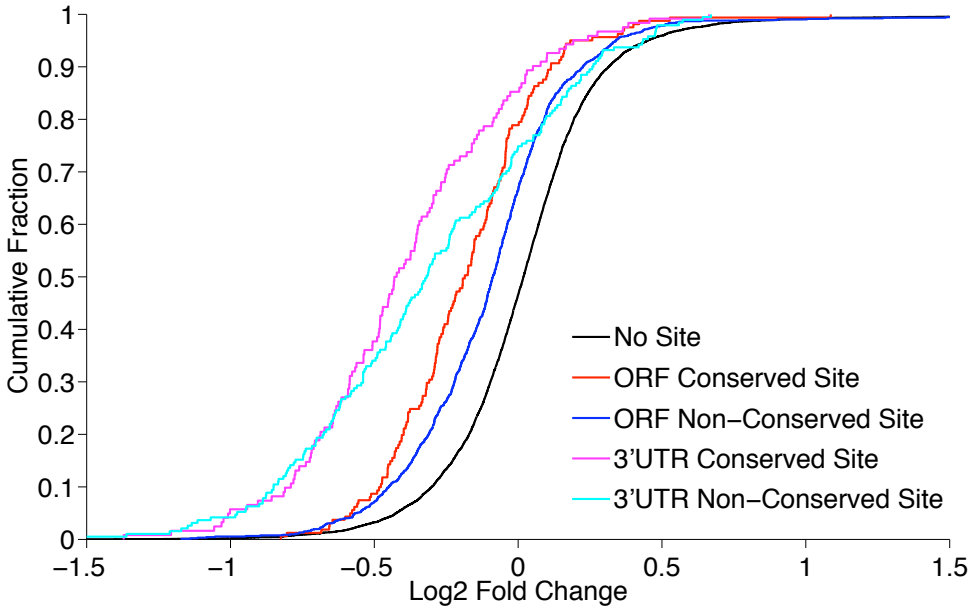

Figure 6-13: Cumulative distributions for genes of different categories. Genes with predicted ORF sites show down-regulation about twice as strong as those with any ORF site, and about half as strong as those with predicted 3'UTR sites.

showed mean down-regulation of a similar scale to those with non-conserved seed sites in their ORF (6%, $p < 4 \times 10^{-5}$; K-S test).

Chapter 7

Importance of ORF Targeting

The results in this section show that conserved miRNA targeting in *Drosophila* ORFs occurs at a similar scale to conserved targeting in 3'UTRs. As in 3'UTRs, not all of the contextual features that make some target sites more effective than others are known. However, by seeking highly conserved seed matches, the MinoTar algorithm can recover functional ORF sites with high confidence. Additionally, the results suggest that factors indicative of stronger targeting in 3'UTRs, particularly the presence of multiple seed sites, should also be helpful in finding the most effective ORF sites. And while most predicted ORF targets are not predicted to be 3'UTR targets, the set of genes targeted in both regions show significantly more overlap than expected, suggesting that simultaneous targeting of genes in both regions occurs in some cases.

In general, targeting in ORFs appears to be weaker than 3'UTR targeting. However, given the scale of conserved ORF targeting observed, it seems likely that a large number of important ORF targets remain to be discovered. For both 3'UTR and ORF targets, it remains an open question how to interpret the vast scale of conserved miRNA targeting observed. With tens to hundreds of preferentially conserved targets per miRNA in the 3'UTR alone, a similar number in ORFs, and the potential for

species-specific [19] as well as non-canonical targets [69], the scale of targeting by miRNAs is daunting. It has been suggested that a significant fraction of target sites may impart only modest down-regulation, and exist merely to finely tune expression levels [7]. Given the weaker strength of targeting in ORFs compared to 3'UTRs, it seems possible that for ORFs an even greater fraction of sites may serve such a purpose.

Interestingly, comparison of results in *Drosophila* and in human indicates that the relative importance of targeting in ORFs and 5'UTRs to 3'UTRs may be stronger in *Drosophila*. The discrepancy is particularly strong for 5'UTRs, where for humans the fraction of miRNA seed sites conserved is quite small, while in *Drosophila* the fraction is about 40% the fraction conserved in 3'UTRs. Indeed, in the miR-1 over-expression microarray experiment, genes with a 5'UTR site for miR-1 showed significant down-regulation, suggesting that many of these sites may be functional. One possibility is that in *Drosophila*, where 3'UTRs are significantly shorter than in humans, microRNAs have been forced to make greater use of ORFs and 5'UTRs for targeting. Such an interpretation must be taken with caution, however. Because of the far larger effective population size in *Drosophila* than in humans ($\sim 10^6$ versus $\sim 10^4$), features under far smaller selective pressures will be conserved in *Drosophila*. Therefore, it is possible that this difference merely reflects the difference in thresholding of conservation for the two clades, rather than a difference in biological importance. As ORF targeting in both species continues to be investigated, the answer should become clearer.

An interesting question is why targeting should be weaker in ORFs or 5'UTRs than in 3'UTRs. A number of experiments have suggested an answer. According to these results, the reduced efficacy of 5'UTR and ORF sites is attributed to displacement of the miRNA silencing complex within the 5'UTR by the scanning machinery, as it passes from the cap to the start codon, and within the coding region by translo-

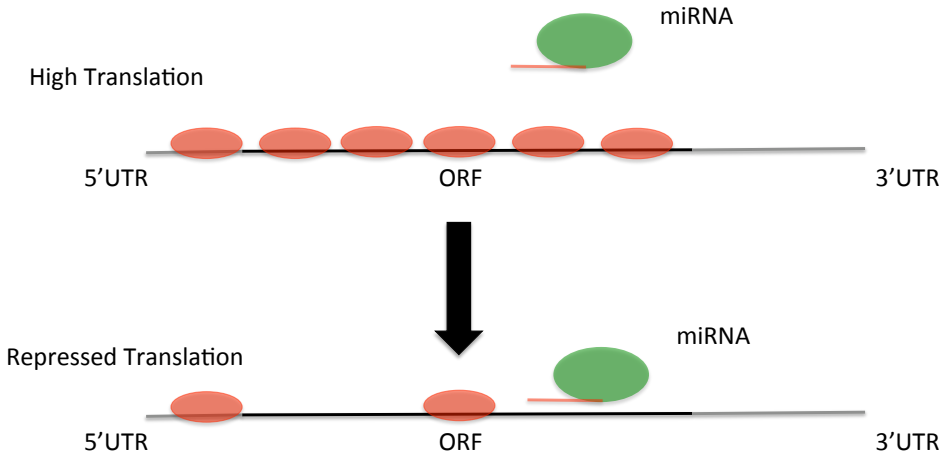

Figure 7-1: An intriguing possibility is that the strength of ORF targeting can be modulated according to cellular conditions. While miRNA targeting in ORFs is normally partially repressed by the passage of ribosomes, under conditions with repressed translation (such as starvation), this repression could be alleviated and the targeting made stronger.

cating ribosomes [38, 39]. Supporting this model, 3'UTR sites that lie within $\sim$ 15 nucleotides of the stop codon are no more effective than ORF sites, as expected if the silencing complex were displaced by the ribosome leading edge as the stop codon approached the ribosome A site [38].

If this is the reason for ORF targeting being less effective, an intriguing possibility is that the strength of ORF targeting could be modulated according to the global level of translation in the cell. A number of cellular conditions, such as starvation, are known to repress translation globally. Given this, it seems possible that ORF targets that are of only modest effect under normal cellular conditions, could have a significant regulatory role under such conditions (Fig 7-1). This could enable a whole class of regulatory relationships to be revealed, allowing for large numbers of genes to quickly be turned off under these conditions. I have made a significant effort to test this in *Drosophila* S2R+ cells using starvation conditions (as well as stimulation

of the insulin pathway which should have the opposite effect and make ORF targets less effective), but have not been able to see a repeatable effect so far. I am currently continuing to examine this possibility in humans in collaboration with the Bartel lab at the Whitehead Institute.

Part III

Repeat-Mediated MicroRNA Targeting

Chapter 8

Sequence Repeats

The human genome (like the genomes of most species), is filled with repeated sequences of various types [23]. These repeats range from multiple adjacent copies of simple nucleotide sequences (such as the CAG/CTG tri-nucleotides in the case of poly-Q repeats), to highly repeated protein domains, to sequences of several kilobases associated with transposons. Many of these regions are subject to mutational processes that can quickly change the number of repeats, leading to their rapid evolution [63]. As a result, these regions have particularly high variability between people, a useful property that has led to the widespread use of some regions in forensic tests [62], and which in the case of certain repeats can lead to important phenotypic consequences, including a number of diseases [79, 100].

Sequence repeats can exert phenotypic consequences through a number of mechanisms. These include effects at the protein level as well as at the DNA or RNA level. At the protein level, repeats can block the natural protein degradation process, cause aberrant protein accumulations, or interfere with protein function [79, 100]. At the DNA level, repeats may cause epigenetic changes that result in widespread changes in transcription, and can also function directly as promoter elements to re-

cruit transcriptional activators and silencers [79, 100]. At the RNA level, repeats may change RNA secondary structure, alter regulation of the translation rate through the recruitment of proteins, or lead to the formation of short silencing RNAs through the formation of double stranded RNA [79, 100].

In this section, I describe a novel relationship between repeated sequences and microRNA targeting (and most specifically targeting in ORFs). This relationship leads to a new mechanism under which repeats can exert phenotypic effect, and suggests a number of cases where microRNA targeting in ORFs can be far stronger than earlier work has indicated was possible. As briefly described in Section II of this thesis, work on microRNA targeting in mammalian ORFs has provided a mixed picture on the importance of such targeting. Large-scale studies examining the effects of introducing or deleting a miRNA have shown that sites in 3'UTRs generally are more effective than those in either 5'UTRs or open reading frames (ORFs) [6]. However, although ORF sites generally are less effective, enough ORF sites mediate repression to observe a signal above background in large-scale functional studies [74, 38, 5, 90], and even more sites appear to bind the silencing complex sufficiently to mediate enrichment of the mRNA (or a cross-linked fragment of the mRNA) during immunoprecipitation of the silencing complex [28, 46, 20, 41, 106]. Supporting the biological function of some of these ORF sites, bioinformatic approaches (including the work in this thesis) have shown that many ORF sites are preferentially conserved [73, 95, 33]. And reporter assays have also confirmed that sites in both 5'UTRs and ORFs can mediate repression [28, 77, 27, 33, 80, 91, 97, 29, 56].

The efficacy of miRNA-mediated repression increases with the number of sites, suggesting that targeting might be substantial if a gene contained many sites in its coding region, as could arise from coding-sequence repeats. While simple classes of repeats, such as microsatellites, are particularly widespread in the genome, of greater

interest in this work are repeated sequences of greater complexity, a striking case of which occurs within the C2H2 class of zinc-finger genes. C2H2 genes have undergone extensive expansion over the course of vertebrate evolution and constitute the largest group of human transcription factors [24]. Typically, C2H2 genes contain a significant number of tandemly-repeated C2H2 amino-acid domains, each of which coordinates a zinc ion and has the potential to bind DNA in a sequence-specific manner [105]. The emergence of highly repeated amino-acid domains creates instances of highly repeated short nucleotide sequences in a large fraction of such genes.

In this section of the thesis, I show that the ORFs of many repeat-rich genes contain strikingly large numbers of target sites of particular miRNAs. Moreover, these genes with many sites are frequently strongly repressed. For seven miRNA families, this type of targeting appears to be extensive. In the most notable cases, four miRNA families (miR-23, miR-181, miR-188 and miR-199) have seed sites that match repeated sequences within C2H2 zinc-finger genes. Three others (miR-370, miR-766 and miR-1248) have seed sites that match simpler repeats. Effective targeting of coding-region repeats is highly predictable, and, due to the large number of target sites within a single ORF, down-regulation observed in reporter assays can be stronger than that of many 3'UTR targets. For the C2H2 class of zinc-finger genes, targeting is shared among paralogous genes, suggesting the potential for their coordinate regulation. mRNAs of both RB1 and its accessory protein, RBAK, are targeted by miR-181 primarily through ORF sites, suggesting an unappreciated role for miR-181 in cancer development and underscoring the potential importance of ORF sites in understanding miRNA biology

Chapter 9

Repeat-Mediated MicroRNA Targeting

9.1 miR-181 Represses mRNAs with ORF Repeats

The investigation in this section of the thesis was pursued after a re-analysis of previously published microarray data from miR-181a transfection in HeLa cells [5] yielded a surprising result. In that experiment, some of the most strongly down-regulated mRNAs contained miR-181 seed sites in their coding regions. This result was intriguing because, although miRNA targeting has been observed in ORFs, it usually confers more subtle repression. Further investigation revealed that the strong down-regulation was the result of a group of mRNAs containing numerous miR-181 sites (Fig 9-1). Genes that contained a single 8mer site in their coding region were only slightly, though statistically significantly, down-regulated (mean log2 fold-change: -0.07; $p < 2 \times 10^{-4}$, Mann-Whitney U test) and significantly less so than those with a single 8mer site in their 3'UTR (mean log2 fold-change: -0.07 vs. -0.17; $p < 0.01$, Mann-Whitney U test). However, those genes with an increasing number

of ORF sites showed increasingly strong down-regulation. In particular, the 52 genes with at least four 8mer ORF sites were nearly all repressed and, even when restricted to the subset of 18 of 52 genes with no 3'UTR sites, showed significantly stronger repression than those genes with a single 8mer 3'UTR site (mean log2 fold-change: -0.51 vs. -0.17; $p < 5 \times 10^{-3}$, Mann-Whitney U test). Many of these genes had even more than four 8mers sites as well as equally large numbers of 7mer sites (Table B.2). Perhaps most surprising, considering the strikingly large number of sites, was not that such genes were strongly down-regulated, but that genes should have so many sites to a single mRNA. Indeed, miR-181 was exceptional in this regard (Fig 9-2). Whereas for most miRNAs few or no genes contained large numbers of 8mer sites, for miR-181 there were many such genes.

After observing the impact of miR-181 on a large class of endogenous transcripts, the first aim was to confirm the direct, miRNA-mediated repression of a number of specific genes. To do so, reporter proteins were generated in which firefly luciferase was fused in-frame to the C terminus of the protein product of each of five genes: ZNF573, ZFP37, ZNF20, ZNF791 and RBAK. These genes contained (9, 8, 6, 5, 9) miR-181 8mer sites and (2, 7, 2, 1, 10) miR-181 7mer sites, respectively. Repression by miR-181a was evaluated by comparing normalized luciferase values from cells co-transfected with miR-181a to those with miR-23a, a non-cognate miRNA. For each gene tested, significant and robust repression by miR-181a was observed (Fig 9-3 top panel; $p < 10^{-6}$ Mann Whitney U Test). To confirm that the measured signal in each case was coming from the full-length fusion proteins, the reading frame register was disrupted by inserting or deleting one nucleotide early in the zinc finger coding sequence of the mRNA. For all of the fusion constructs, these frame shift mutations significantly reduced luciferase activity (Fig 9-3 bottom panel).

To confirm that the observed repression was directly mediated by the ORF sites,

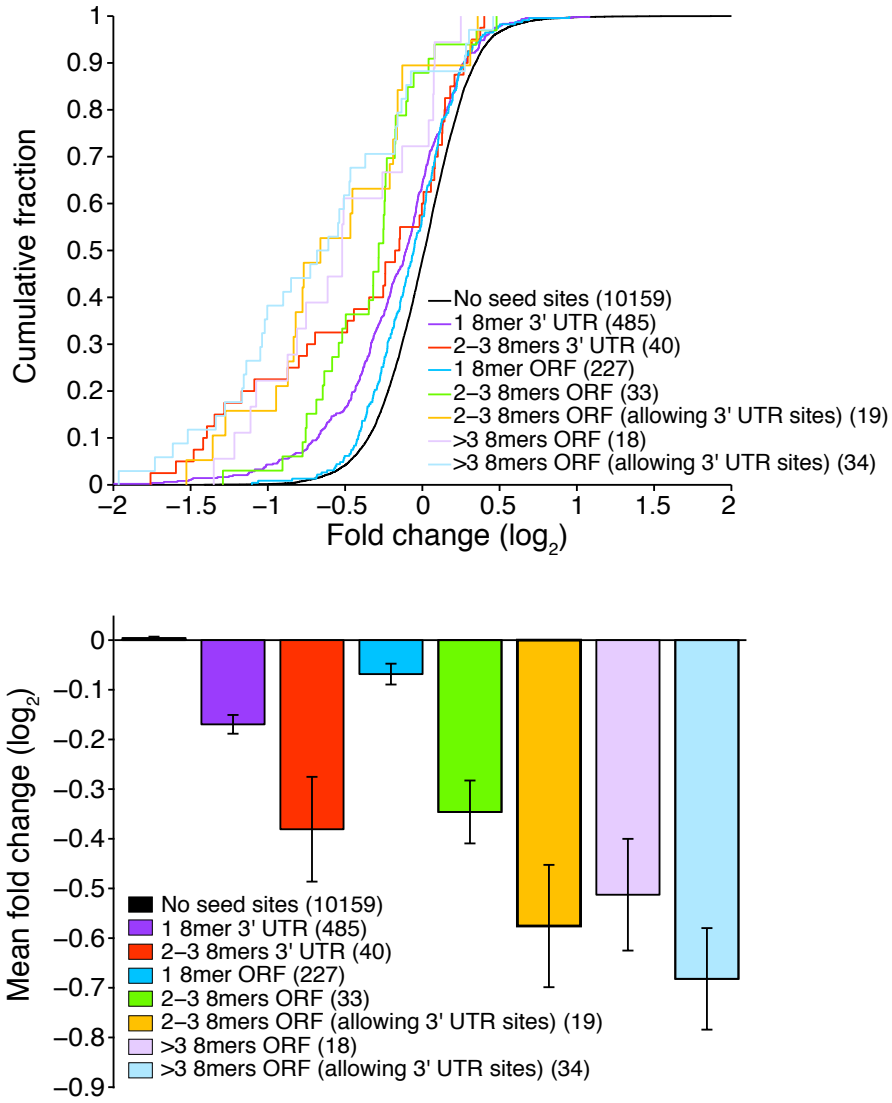

Figure 9-1: miR-181 targets genes with multiple coding-region sites. Top panel: Response of mRNAs to the introduction of miR-181a into HeLa cells. Plotted are cumulative distributions of fold changes for mRNAs with the indicated numbers and types of sites. Except for the two categories indicated, categories with ORF sites excluded mRNAs with 3'UTR sites, and those with 3'UTR sites excluded mRNAs with ORF sites. Bottom panel: Mean fold changes for the same categories of genes. Error bars show standard deviation from bootstrapping.

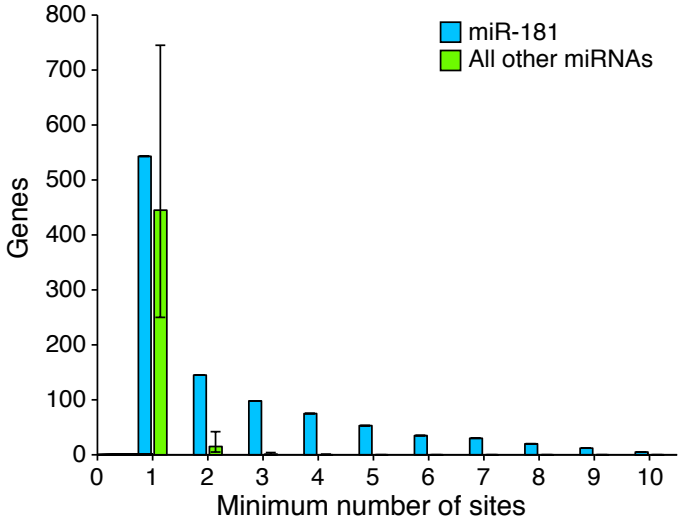

Figure 9-2: The propensity of miR-181 to have many ORF sites. Plotted are numbers of genes containing at least the indicated number of sites for either miR-181 (green) or the median across all miRNAs (blue). Error bars show the interquartile range.

a ZNF20-luciferase mutant in which each of the six miR-181 8mer and two miR-181 7mer seed sites within the ORF were mutated with two synonymous point substitutions was generated. Compared with this mutated construct, wildtype ZNF20 was significantly and specifically repressed by miR-181a (Fig 9-4 top panel; $p < 10^{-6}$). The repression attributed to these sites increased from 2.5-fold to 6.3-fold when the fragment containing the sites was incorporated as part of the reporter 3'UTR (Fig 9-5). This difference between the efficacy of ORF sites and 3'UTR sites was significant ($p < 10^{-4}$) and consistent with the general observation of ORFs being more refractory to miRNA targeting than are 3'UTRs [38].

Similarly, a test of whether the repression of RBAK, observed in the luciferase assays (Fig 9-3) and in the previous microarray study [5], was directly mediated by its miR-181 sites was performed. In addition to its nineteen ORF sites (nine 8mers

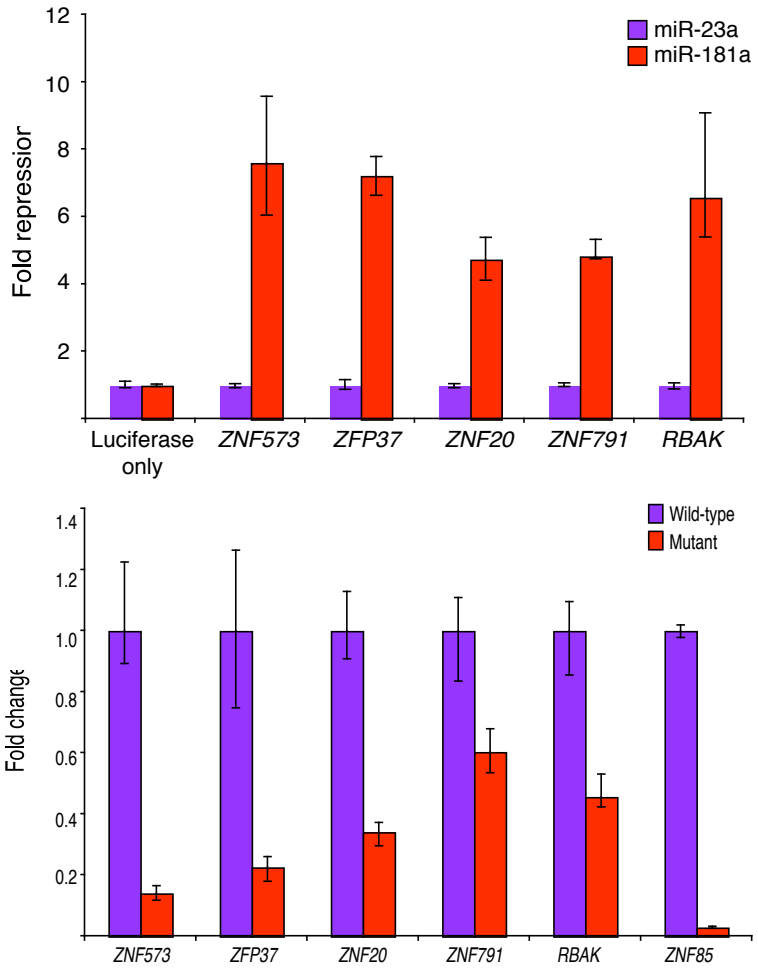

Figure 9-3: Top panel: miR-181a mediated repression of reporters with miR-181 ORF sites. Reporters included the luciferase ORF following the ORF of the indicated mRNA. Fold repression was calculated relative to that of the non-cognate miRNA, miR-23a. Plotted are the normalized values, with error bars representing the third largest and third smallest values ($n = 12$; $p < 10^{-6}$, except for the control, luciferase-only reporter, for which $p = 0.32$). Bottom panel: Zinc-finger-luciferase constructs as genuine fusion proteins. In order to verify that the ORF-luciferase proteins were being appropriately expressed, firefly luciferase expression, normalized to the Renilla luciferase transfection control, of wild-type constructs was compared with that of mutant constructs ($n \leq 5$; $p < 0.005$).

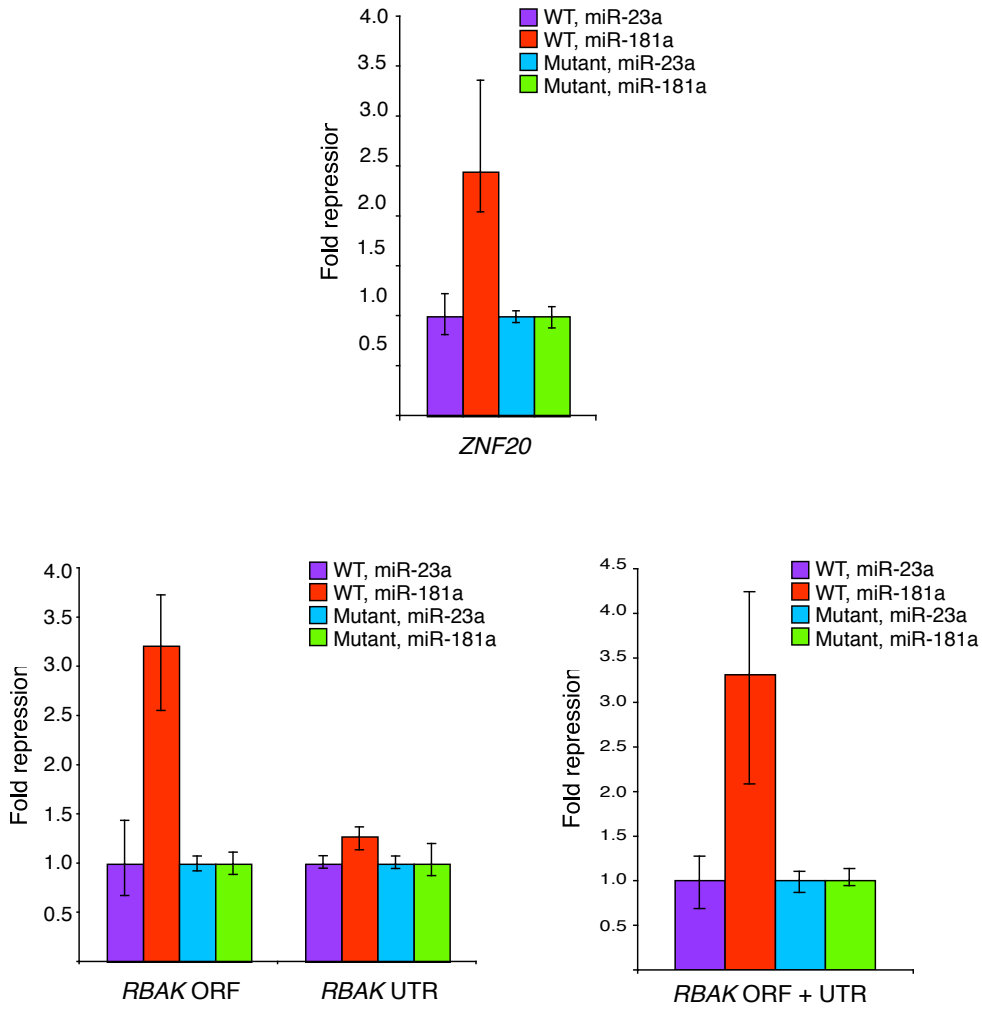

Figure 9-4: Top panel: Dependence of ZNF20 repression on miR-181 ORF sites. Repression was calculated and depicted as in (Fig 9-3), additionally normalizing repression of the reporter with wild-type sites (WT) to that of a reporter in which the eight ORF sites were mutated ($n = 12$; $p < 10^{-6}$). Left bottom panel: Direct repression of RBAK ORF and 3'UTR mediated by miR-181a ($n = 12$; $p = 7 \times 10^{-7}$ and $p = 0.0009$, respectively). Right bottom panel: miR-181a directly represses combined RBAK ORF and 3'UTR ($n = 12$; $p < 8 \times 10^{-7}$).

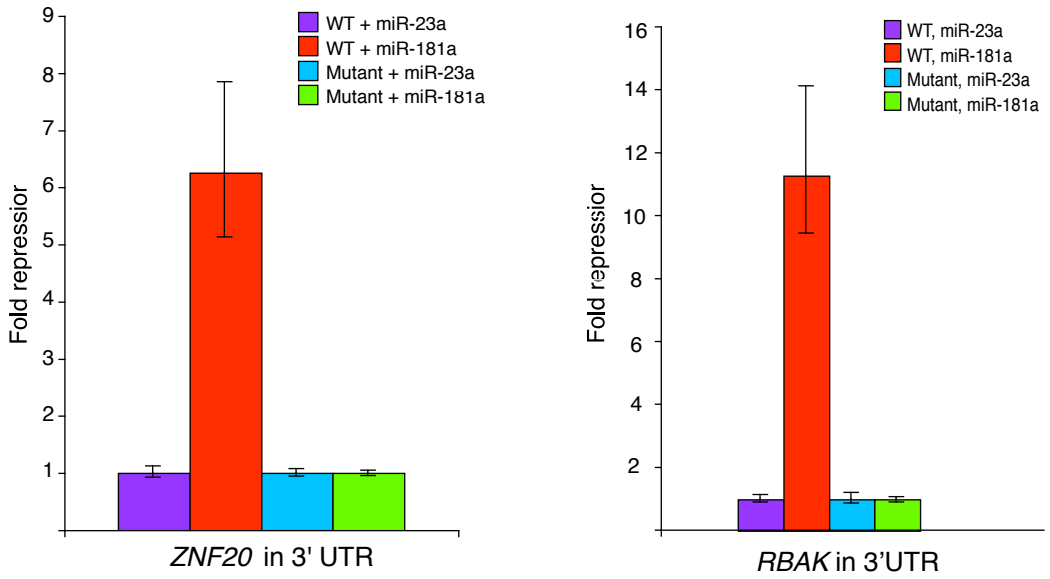

Figure 9-5: Repression of ZNF20 and RBAK increased when expressed as a 3'UTR. Direct repression of ZNF20 (left panel) and RBAK (right panel), expressed as a 3'UTR, mediated by miR-181a, as assayed by luciferase assay. Fold repression was calculated relative to that of the non-cognate miRNA, miR-23a. Plotted are the normalized values where the repression of the reporter with wild-type sites (WT) was normalized to that of the mutant reporter in which the eight miR-181 sites were mutated. Error bars represent the third largest and third smallest values ($n = 15$; $p < 2 \times 10^{-8}$).

and ten 7mers), the 3'UTR of RBAK contains a conserved 7mer-m8 site and a poorly conserved 7mer-A1 site [34]. A panel of constructs was generated, expressing either the RBAK ORF with a C-terminal luciferase tag or the RBAK 3'UTR following the luciferase reporter (Fig 9-4 bottom left panel). Although the 3'UTR sites mediated statistically significant repression (1.3-fold; $p = 0.0009$), the ORF sites gave far stronger repression (3.2-fold repression; $p < 10^{-6}$). When both the ORF and 3'UTR sites were monitored in combination, 3.3-fold repression was observed (Fig 9-4 bottom right panel; $p < 10^{-6}$), which was not significantly different from that observed with the ORF sites alone ($p = 0.59$). These data indicate that the repression of RBAK by miR-181a was direct and that the majority of the miR-181a mediated repression of RBAK was due to targeting through its ORF sites. Again, analogous to the results on ZNF20, the repression attributed to the ORF sites increased from 3.2-fold to 11.2-fold when the fragment containing the sites was incorporated as part of the reporter 3'UTR (Fig 9-5).

9.2 Additional miRNA Seeds Match ORF Repeats

To find other miRNAs that might affect ORF targets similarly, a search was performed for all 8mers highly repeated within human ORFs. For each 8mer, the number of non-overlapping occurrences within each ORF was counted. Because four ORF sites mediated robust repression by miR-181, this was chosen as a threshold, and for each 8mer the number of genes with four or more instances of that 8mer was recorded (Fig 9-6). The majority (77%) of 8mers did not occur four or more times in any coding region, and the vast majority (97%) occurred four or more times in only five or fewer coding regions. At the tail of the distribution, 334 8mers appeared at least four times in at least 25 genes, among which were seven miRNA 8mer seed matches (Table 9.1). For each of the seven corresponding miRNA seed families, potential target sets

Family	Seed Complement	Conservation	Number Targets	Target Family
miR-23	AATGTGAA	Vertebrates	43	C2H2 Zinc Fingers
miR-181	TGAATGTA	Vertebrates	75	C2H2 Zinc Fingers
miR-188-3p	TGTGGGAA	Mammals	210	C2H2 Zinc Fingers
miR-199-5p	ACACTGGA	Vertebrates	82	C2H2 Zinc Fingers
miR-370	CAGCAGGA	Mammals	25	Varied
miR-766	GCTGGAGA	Human	27	Varied
miR-1248	AAGAAGGA	Human	34	Varied

Table 9.1: Information on microRNAs with repeated coding-region sites in many genes

containing genes with at least four 8mer sites were compiled (these are given in Table B.1, Table B.2, Table B.3, Table B.4, Table B.5, Table B.6, and Table B.7).

Based on the sets of predicted targeted genes, these miRNA families split into two main groups (Fig 9-6). For four of the miRNAs (miR-23, miR-181, miR-188 and miR-199), the target sets were almost entirely C2H2 zinc-finger genes, and sites in these genes mostly occurred within tandemly repeated C2H2 amino-acid domains. For the other three annotated miRNAs (miR-370, miR-766, and miR-1248), the predicted targets were more varied, and sites largely occurred within highly prevalent amino-acid pairs or triplets, though in some cases they occurred within long stretches of simple nucleotide repeats. Because of their limited conservation, low expression levels and questionable status as authentic miRNAs, two of these three miRNAs (miR-766 and miR-1248) were not considered further.

The most common form of the C2H2 amino-acid domain is $XCX_2CX_{12}HX_3H$ (where X represents any amino acid and the subscripts represent the number of amino acids [31]). Typically, C2H2 zinc-finger genes contain many tandem repeats of the zinc-finger motif (8.5 on average in humans), which are connected by a specific linker sequence most commonly of the form TGEKPY [31]. The 8mer sites for the four miRNA families each occurred within specific amino-acid realizations at specific lo-

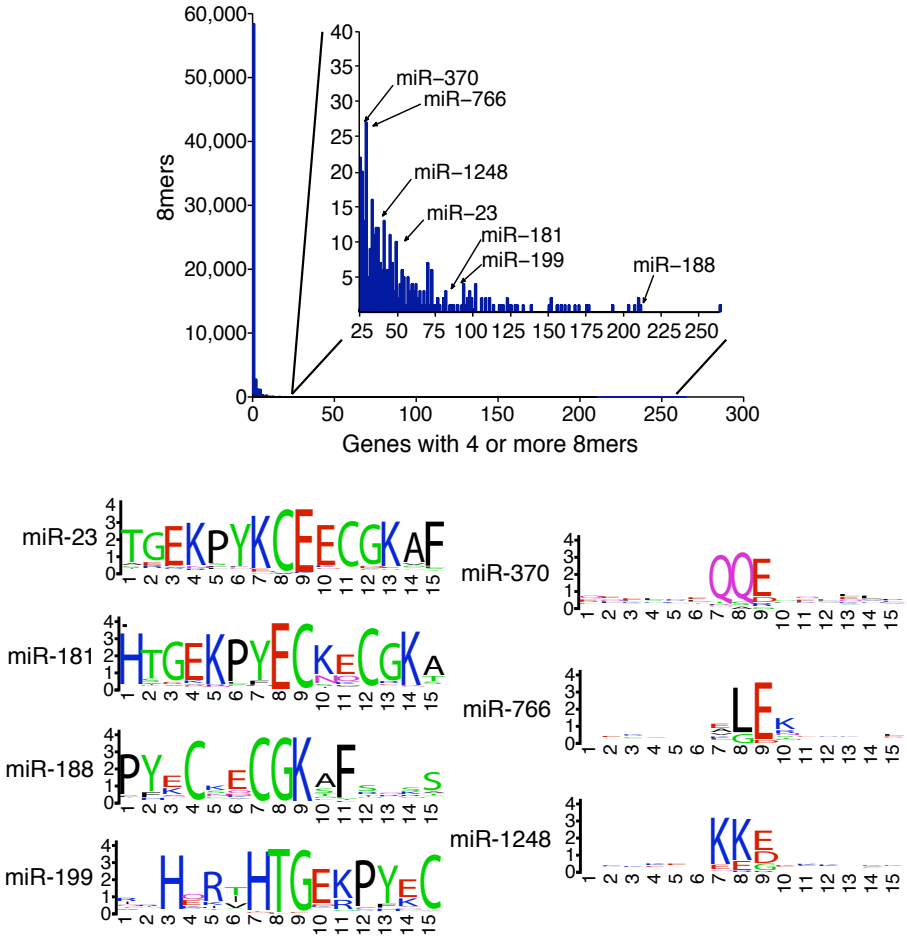

Figure 9-6: Potential targeting of frequent ORF repeats by additional miRNAs. Top panel: Motifs frequently repeated in ORFs of many mRNAs. The histogram considers all 65,536 possible 8mer motifs and plots the number of these 8mers that match the indicated number of unique ORF at least four times. Also indicated are the seven miRNAs with 8mer sites among the 334 motifs that appeared at least four times in at least 25 genes. Bottom panel: The amino-acid sequence coded by regions flanking 8mer ORF sites corresponding to the miRNA indicated. Sites overlap codons 7–9. Letter size indicates enrichment, visualized using WebLogo.

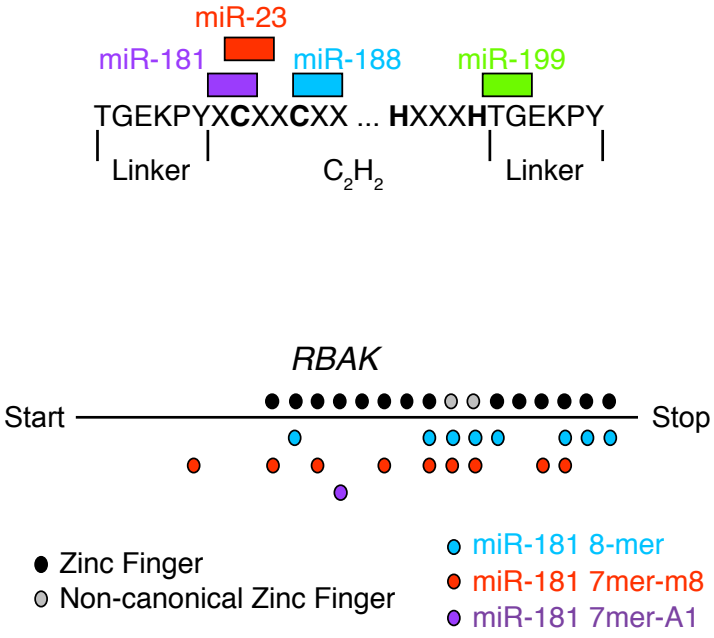

Figure 9-7: Top panel: Locations of 8mer sites within the repeated C2H2 domain. Bottom panel: Locations of C2H2 domains and miR-181 sites within the RBAK gene. Two C2H2 domains in which one of the histidines has been lost are shown as non-canonical zinc fingers. In cases where a single zinc finger contains both 8mer and 7mer sites, the 7mer overlaps the second cysteine in the motif

cations in this motif (Fig 9-7 top panel). Because of the large numbers of paralogous motifs within a single gene, the number of seed sites can be large. A typical example is the gene RBAK, which contains fourteen C2H2 domains, eight miR-181 8mers and ten miR-181 7mers (Fig 9-7 bottom panel).

9.3 Predicted Targets are Repressed by miRNAs

From the list of predicted miR-23 targets (Table B.1), three genes (ZNF225, ZNF486 and ZNF85) were chosen for experimental follow-up. Fusion constructs with a C-

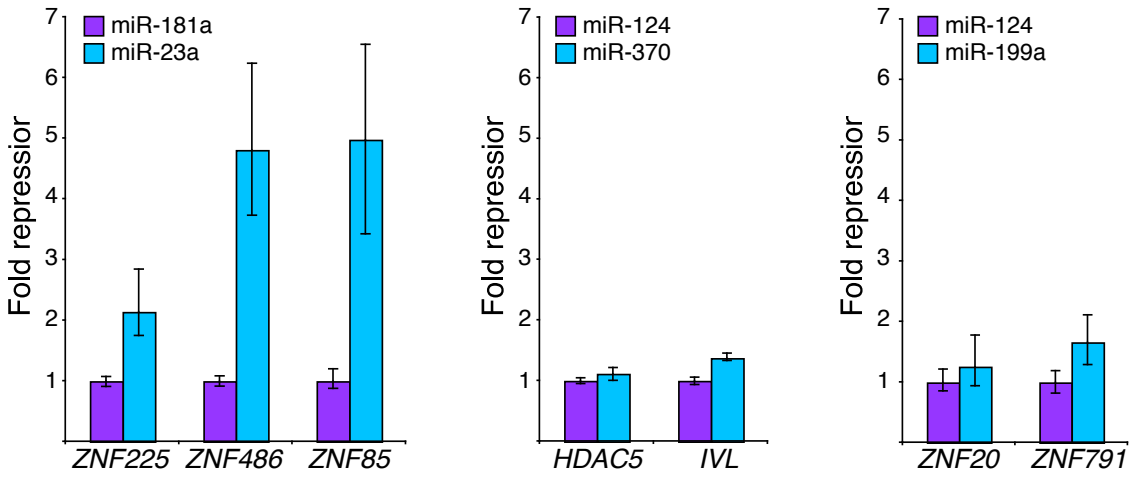

Figure 9-8: ORF target predictions recover functional targets for additional miRNAs. Left panel: miR-23a mediated repression of reporters with miR-23 ORF sites ($n = 12$; $p < 10^{-6}$). Middle panel: miR-370 mediated repression of reporters with miR-370 ORF sites ($n = 15$; $p = 0.0009$ and 10^{-8}, for HDAC5 and IVL, respectively). Right panel: miR-199a mediated repression of a reporter with miR-199 ORF sites ($n = 12$; $p = 2 \times 10^{-6}$ and 0.08, for ZNF791 and ZNF20, respectively).

terminal luciferase tag were made as before, and disruption of the reading frame by a nucleotide insertion substantially decreased luciferase expression, confirming that the majority of signal came from the full-length protein (Fig 9-3 bottom panel). The constructs were transfected in the presence of miR-23a or a non-cognate miRNA, miR-181a. For all three targets, normalized luciferase values were significantly reduced in the presence of miR-23a when compared with those with the non-cognate miRNA (Fig 9-8 left panel; $p < 10^{-6}$). The magnitudes of repression were particularly notable because miR-23 has high target abundance and thus generally weaker targeting efficacy [4].

This analysis was extended to predicted targets of both miR-370 and miR-199. In the case of miR-370, the response of two targets was probed, IVL and HDAC5, both

of which were significantly repressed ($p < 10^{-7}$ and $p = 0.009$, respectively; Fig 9-8). Similarly, ZNF20 and ZNF791, experimentally supported targets of miR-181, also contained multiple seed matches to miR-199 and were examined for their response to miR-199a (Fig 9-8). ZNF791, which contains six miR-199 8mer sites and one 7mer site, was significantly repressed ($p < 10^{-5}$). In contrast, the luciferase control was either unaffected by these miRNAs or, in the case of miR-199a, significantly less repressed than the ZNF791-luciferase fusion ($p = 0.0003$). Taken together, although individual ORF sites are less effective than 3'UTR sites [38], these results indicate that highly repetitive ORFs containing many miRNA sites can generally be subject to significant and, in some cases, substantial repression by the cognate miRNA.

9.4 Targets are Paralogous Families of C2H2 Genes

Having observed the extent and, in some cases, surprising magnitude of targeting arising from ORF repeats, I next considered the evolutionary processes giving rise to this phenomenon. The focus was on C2H2 zinc-finger genes because these formed the most dramatic and frequent instances of this phenomenon. Among the predicted targets, most C2H2 genes (>80%) contained the general transcriptional repressor KRAB domain (Fig 9-9 top panel). KRAB-containing C2H2 genes display particularly interesting patterns of evolution, having high rates of gene duplication and loss as well as a dramatic expansion over the course of vertebrate and mammalian evolution [57]. To understand the role that these duplications played in forming miRNA targets sets, I collected sequences of all KRAB domains annotated in human and used these to create a multiple alignment of KRAB domains. From this alignment, the inferred phylogeny of KRAB-containing genes provided a context for considering the four miRNA target sets (Fig 9-10).

The inferred phylogeny revealed that each of the four miRNAs target multiple

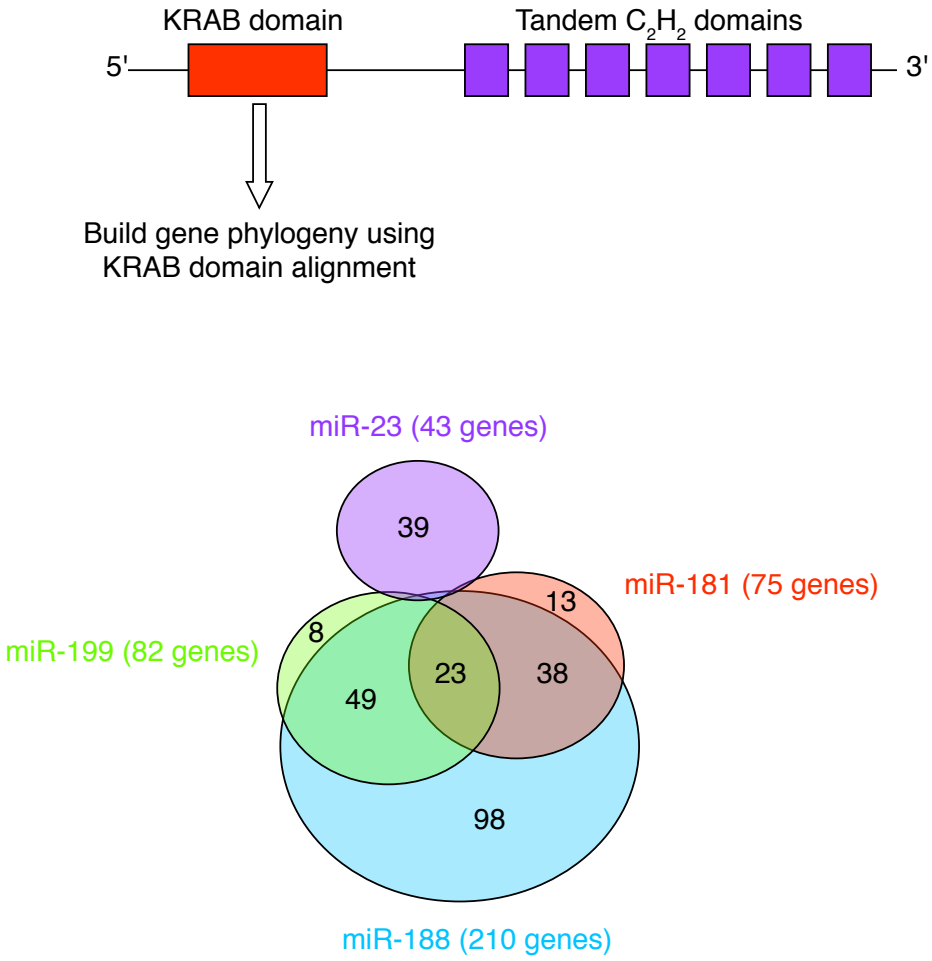

Figure 9-9: Top panel: Diagram of domain structure of C2H2 zinc-finger genes. Bottom panel: Overlap between the predicted targets of the indicated miRNAs.

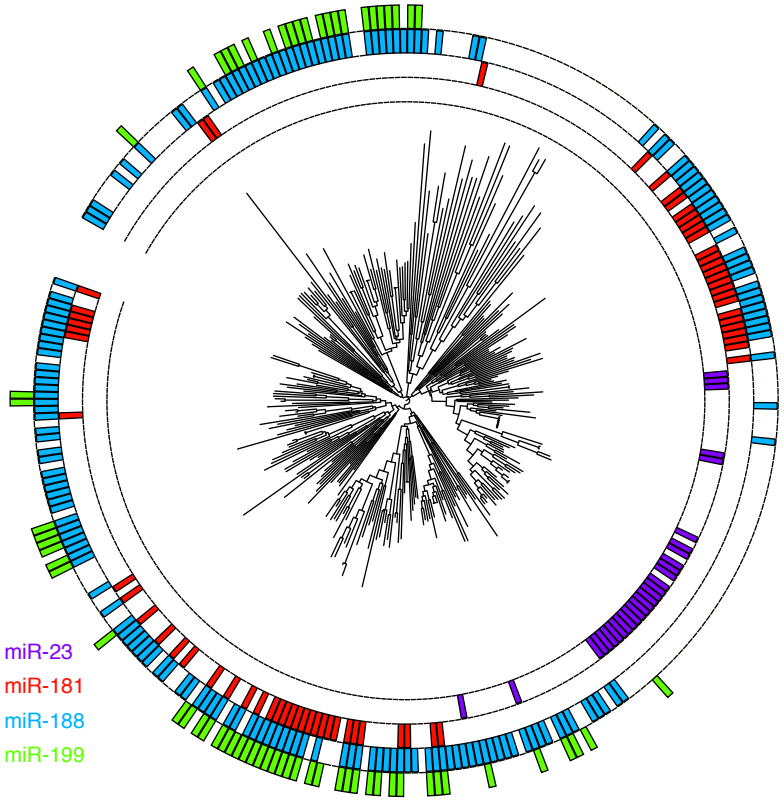

Figure 9-10: The relationship between shared ancestry of KRAB-containing C2H2 genes and shared miRNA sites. The phylogeny inferred from alignment of the KRAB domains is shown, marking at the perimeter those with at least four 8mer sites to the indicated miRNA.

families of genes that have undergone significant expansions through gene duplication. Combined over the four miRNA families, the target sets covered the majority of KRAB-containing C2H2 genes with varying amounts of overlap between each set (9-9 bottom panel). Whereas the miR-188 target set spanned nearly the entire phylogeny, the other three miRNAs each targeted more specific families of genes. For instance, the miR-23 targets were nearly all members of one family: the recently duplicated ZNF91 family, which has undergone a significant expansion in the primate lineage [42].

Sequence analysis suggested that two duplication processes contributed to the creation of such extensive repeat-containing families. First, individual C2H2 domains were duplicated multiple times within a single zinc-finger gene. Next, the gene itself was duplicated to form an extensive gene family, and this family underwent subsequent diversification, with occasional intragenic domain duplication or loss. Due to the initial intragenic duplication process, the sequences of individual zinc fingers within a gene were far more similar to each other than expected by chance. To verify the importance of this effect, I implemented a randomization procedure for nucleotide sequences within C2H2 domains that preserved amino acid sequences and average codon usage over all domain instances. Even when fixing the observed amino acid sequences, real instances of C2H2 domains from within the same gene showed significantly higher nucleotide similarity than expected by chance (Fig 9-11). When C2H2 domains with randomized sequences were mapped back to genes, far fewer of these genes contained large numbers of miRNA sites than did the real genes (Fig 9-12). It was tempting to speculate that the similarity in nucleotide sequences across C2H2 domains within a gene represented an additional selective pressure to maintain miRNA seed sites in these genes. I observed marginal evidence for this model but unfortunately could not detect such an effect at high confidence.

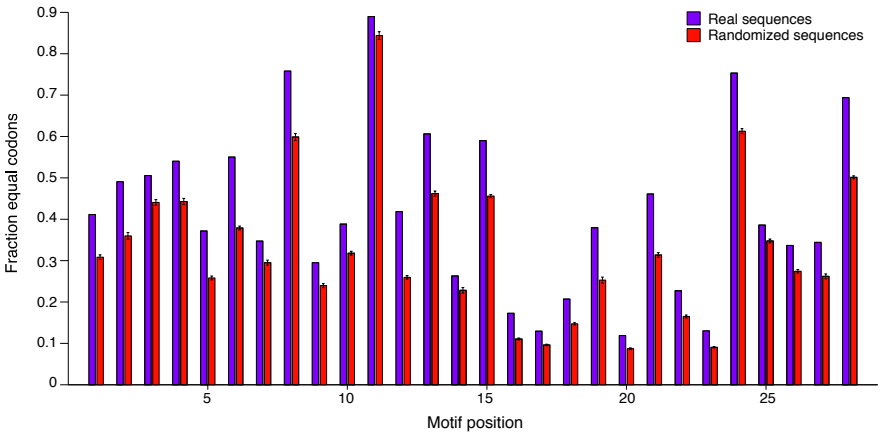

Figure 9-11: Similarity of codons across C2H2 domains for real zinc-finger genes compared to those with randomized nucleotide sequences. Shown are the fractions of equal codon sequences across all pairs of C2H2 domains within the same gene for each position in the C2H2 domain for both real sequences and randomized sequences. Error bars show standard deviation from 50 codon randomization trials.

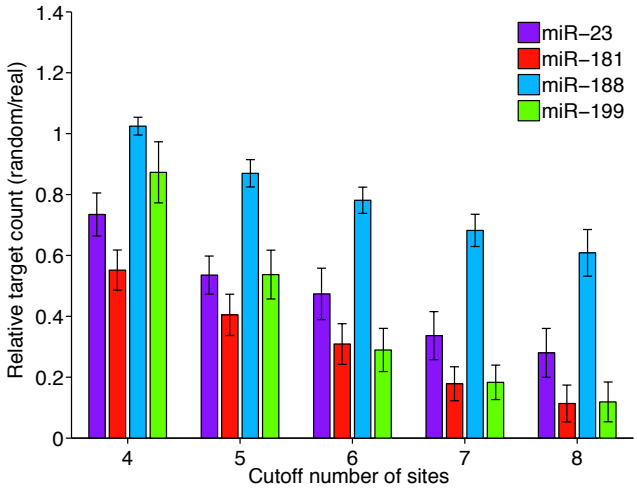

Figure 9-12: Increased targeting due to C2H2 domain similarity. Shown are the mean sizes of gene sets containing the indicated minimum number of 8mers in codon-randomized genes relative to the size of the set for the actual gene sequences. Error bars show standard deviation from 50 randomized trials.

The initial intragenic duplication of individual C2H2 domains allowed for the founding nucleotide-sequence choice to be amplified. After repeated duplication, a gene would ultimately include either many sites or no sites at all, depending on the presence or absence of a target site within a founding C2H2 domain. Such duplication contributed to the modularity of miRNA target sets; even for families of genes with C2H2 domains containing similar amino acid sequences, genes from one family often contain many target sites, whereas those from another contain almost none. For example, although miR-188 sites are prevalent throughout the set of KRAB genes, they are almost entirely absent from the ZNF91 family of genes, despite these genes encoding comparable numbers of instances of the amino acid triplet (CGK) within which the miR-188 seed site appears.

The presence of large numbers of miRNA sites within coding regions also provided an opportunity to gain miRNA sites in the 3'UTR through the acquisition of nonsense mutations. For all four miRNAs, there was clear evidence of this process. For each miRNA, the set of predicted target genes were far more likely to contain 3'UTR target sites than were the overall set of genes (Fig 9-13; $p < 10^{-7}$ for all comparisons except miR-199 7mers; binomial test). Moreover, regions flanking 3'UTR sites of predicted ORF targets had high similarity to the regions flanking the corresponding ORF sites, which suggested that many of these 3'UTR sites resided in the remnants of zinc-finger domains that had been lost to the 3'UTR (Fig 9-14). Although the average number of 3'UTR sites in these genes was modest compared to the number of ORF sites (mean total number of 7mer and 8mer sites: miR-23, 2.7; miR-181, 1.2; miR-199, 0.6; miR-188, 0.6), due to the greater efficacy of the 3'UTR sites, their presence presumably enhances the targeting of some genes.

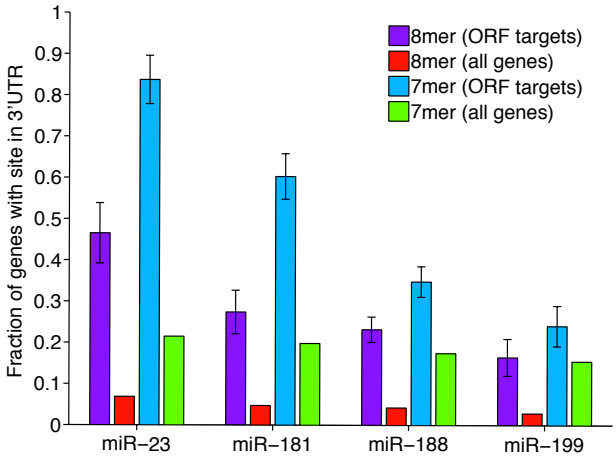

Figure 9-13: Propensity of ORF targets to also be targeted in 3'UTRs. Shown are fractions of genes containing either 8mer or 7mer sites for the indicated miRNA within their 3'UTRs, comparing the ORF target sets with the background of all genes. Error bars show standard deviations from 100 bootstrapping trials.

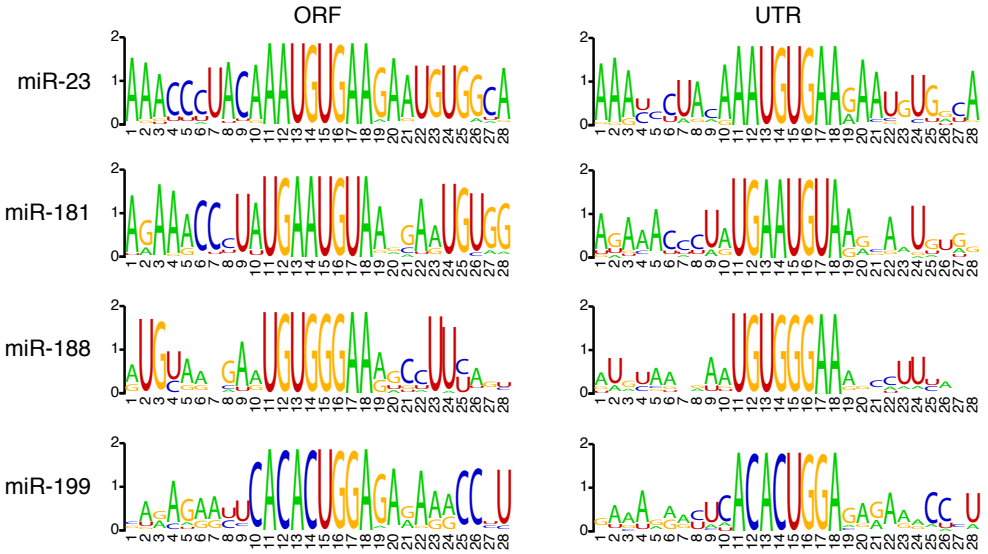

Figure 9-14: Evidence that 3'UTR sites derived from ORF sites. Shown are nucleotide compositions flanking miRNA sites in both ORFs and 3'UTRs of mRNAs with ORF sites.

Chapter 10

Importance of Repeat-Mediated Targeting

In animals, miRNAs target many genes through sites in their 3'UTRs but cause only modest repression of most of these targets. Compared with this generally modest targeting within 3'UTRs, most targeting within ORFs is substantially weaker still. Nonetheless, the work in this section has demonstrated that significant numbers of substantially repressed ORF targets exist and that such targets can be easily identified by the presence of large numbers of sites to the same miRNA. Recently, a similar observation of ORF targeting was independently observed to occur in the case of a single miRNA [56]. The work here, going significantly further, provides a systematic examination of this phenomenon and gives the first bioinformatic evidence and experimental verification of the widespread nature of this type of targeting. Indeed, this targeting involves multiple miRNAs and likely hundreds of genes. While some of the target sites identified show the potential for supplemental 3' base-pairing to the miRNA, none exhibit the extensive complementarity required for cleavage of the mRNA. It therefore appears likely that the mechanism of repression for these genes

is identical to that for most 3'UTR targets.

Interestingly, although there are examples of both 5'- and 3'UTRs containing repeats, none of the highly repeated motifs matched known miRNA seeds, which indicates that repeat-mediated targeting is largely ORF specific. Moreover, although repeats in coding regions are prevalent in numerous animal clades, the presence of miRNA sites within these repeats appeared to be vertebrate-specific. This work focused on those miRNAs with at least four sites in large sets of genes, but there were many other miRNAs with at least this many sites in a handful of genes. In addition, many genes contained multiple ORF sites to multiple miRNAs. The eventual determination of expression patterns for these miRNAs at cellular resolution should enable prediction of additional targets in which ORF sites combine to achieve substantial repression.

The most striking cases of repeat-rich targeting occurred within the KRAB-domain C2H2 zinc-finger genes, which constitute the largest collection of human transcription factors. The results here indicate that a single miRNA can target entire families of these genes, thereby simultaneously regulating large numbers of evolutionarily-related transcription factors. The targeting of such a large number of transcription factors by a miRNA has the potential to cause significant and widespread downstream effects. Through phylogenetic analysis I have shown how repeat-mediated targeting arises in these genes under an extensive evolutionary duplication process. Analysis of KRAB-domain containing families indicates positive selection towards diversification of DNA-binding residues [31]. Hence, following duplication, individual genes frequently gain new downstream regulatory roles, while most members of the family retain the potential for upstream miRNA-mediated regulation. With a few exceptions, the functions of KRAB-domain genes remain unknown [57]. While a few show tissue-specific expression, most are widely expressed [101], and I have found

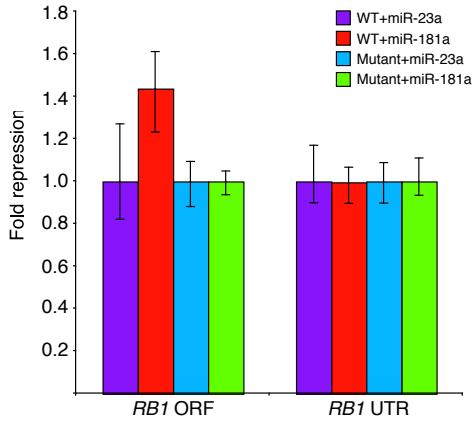

Figure 10-1: Direct repression of RB1 ORF mediated by miR-181a, as assayed by luciferase assays. Fold repression was calculated relative to that of the non-cognate miRNA, miR-23a. Plotted are the normalized values where the repression of the reporter with wild-type sites (WT) was normalized to that of the mutant reporter in which the three ORF sites were mutated. Error bars represent the third largest and third smallest values.

that most predicted target genes appear to have expression patterns overlapping that of the corresponding miRNA, suggesting that the targeting described here is likely to affect their in vivo expression.

One ORF target of miR-181, RBAK, is reported to function as an RB1-associated transcriptional repressor [94]. Suggestive of coordinate regulation, RB1, which encodes the well-characterized tumor-suppressor protein [43, 22], also contains sites to miR-181 in both its ORF (two 8mers and one 7mer) and 3'UTR (one nonconserved 7mer). Although no repression was observed for the 3'UTR site, significant repression by miR-181a was observed for the RB1 ORF-luciferase fusion (Fig 10-1; 1.4-fold repression, $p < 10^{-6}$). This newly recognized miR-181 targeting of RB1 and RBAK had not been appreciated from previous analyses focusing only on 3'UTR sites and is intriguing when considering that miR-181 is up-regulated in some cancers and is

important for maintaining cancer stem cells in hepatocellular carcinoma [87, 15, 61]. Even transient induction of miR-181b is sufficient to mediate an epigenetic switch to cancer, and inhibition of miR-181b reduces colony formation in several cancer cell lines, observations proposed to result from direct targeting of the tumor-suppressor CYLD [59]. Because both RB1 and RBAK repress activation of E2F-dependent promoters and decrease DNA synthesis [94], the ability of miR-181 to repress RB1 and RBAK might provide an additional mechanism by which this miRNA mediates transformation.

The work here also suggests a general role of coding-sequence repeats in post-transcriptional regulation. For transcriptional regulation, the accumulation and clustering of multiple transcription factor binding motifs has been used for some time to predict functional regulatory relationships [44]. Given the large number of post-transcriptional regulatory processes that exist beyond miRNAs [35, 82] and the vast extent of sequence repeats within many protein-coding genes, miRNAs might not be the only regulatory process utilizing this phenomenon.

Part IV

Widespread Non-Coding Function in Coding DNA

Chapter 11

Evolution of DNA in Coding Regions

11.1 Selection on Synonymous Mutations

Due to the degeneracy of the genetic code, the same amino acid sequence can be encoded by many different nucleotide sequences. Classically, synonymous substitutions– those nucleotide changes in a protein coding sequence that don't change the amino acid sequence– were viewed as having no functional significance. Such sites were therefore thought to evolve neutrally [16], an assumption that continues to underly many of the tests for selection at the amino acid level. However, as sequence data became available, it became clear that codons are used in unequal frequencies in many species, a phenomenon termed 'codon bias' [47]. This observation suggested that many synonymous sites are not completely neutral, but are in fact under at least weak purifying selection. While the full causes of codon bias still aren't known, the most common explanation is that the bias reflects pressure to reduce the depletion of tRNAs [16]. One line of evidence supporting this explanation is that in most organ-

isms, including *Drosophila*, the level of codon bias increases with gene expression [92].

In addition to the selective pressures that maintain codon bias, any functional role at either the DNA or RNA level of a gene would be expected to cause selection on synonymous sites. Indeed, specific instances of regulation within coding regions have been previously identified. A number of studies (including the work described in Section II of this thesis) have demonstrated that microRNAs can target sequences in the open reading frame of genes, and cause a significant impact on the evolution of such sequences in both *Drosophila* and mammals. In addition, sequence motifs have been identified that either enhance or silence splicing [32] when found in exons near the splice junction. Finally, a small number of transcriptional enhancers have been found to be located within the coding regions of genes in mammalian Hox genes [70, 98].

These studies suggest that there may be a significant amount of non-coding regulation in coding DNA that awaits discovery, and that a global analysis of conservation in protein coding genes should help guide such discovery. In this section, I undertake such an analysis in *Drosophila melanogaster*. My goals for this undertaking were twofold. First, to estimate the extent of non-coding regulation within coding DNA that appears to be under purifying selection in *Drosophila*. Second, to provide a map of conservation across *Drosophila* protein coding genes that can be used by researchers investigating regulatory elements within specific coding regions. Analogous maps of conservation in non-coding regions have found widespread use in identifying regulatory elements in non-coding DNA [93], and it is my hope that the results here will prove similarly useful.

From this analysis, I have found evidence that non-coding selection on coding DNA in *Drosophila* is widespread, with perhaps 10% of coding DNA under such selection (beyond that expected due to codon bias). I have also found evidence suggesting that regulation of a number of processes, particularly splicing and translation, are a

substantial cause of such selection. Finally, I have identified a number of large regions that exhibit particularly high conservation of their nucleotide sequences, suggesting these regions may be particularly dense in regulatory signals.

11.2 The Effects of Background Selection

Much of the analyses in this section involve comparisons of conservation among regions near different genomic features. The results of such analyses can be influenced in subtle ways by an effect known as background selection. Under this effect, selection at one locus effects the evolution of other loci that are in linkage disequilibrium (LD) with the locus. A substantial literature has been written on this subject, and it remains an active area of investigation. For the purposes of this section, the effect is largely an alternative interpretation for a number of results that must be ruled out. I therefore give a brief description here of this phenomenon, and the impact it can have on the evolution of a locus under selection.

Figure 11-1 gives a basic illustration of the effect in a simple two-locus model. Consider multiple loci that are nearby on the genome and therefore under linkage disequilibrium, where all the loci are under purifying selection. One of these (the red locus in Fig 11-1) is the locus of interest, while the others are interfering loci. If there were no other loci under selection, then the probability of fixation of a new allele subject to selection coefficient s would be $\frac{1-e^{\frac{-sN_e}{N}}}{1-e^{-2N_e s}}$, where N_e is the effective population size (N is the population size at the time the mutation occurs). However, with the other loci present, a number of the alleles at the red locus are linked to deleterious alleles at the other loci and therefore are bound to be lost from the population (save for a recomination event). Because of this, there is a reduction in the effective population size N_e. In the case of no recombination, this reduction is simply by a factor of $1-q$, where $1-q$ is the fraction of alleles that contain deleterious alleles at the other loci.

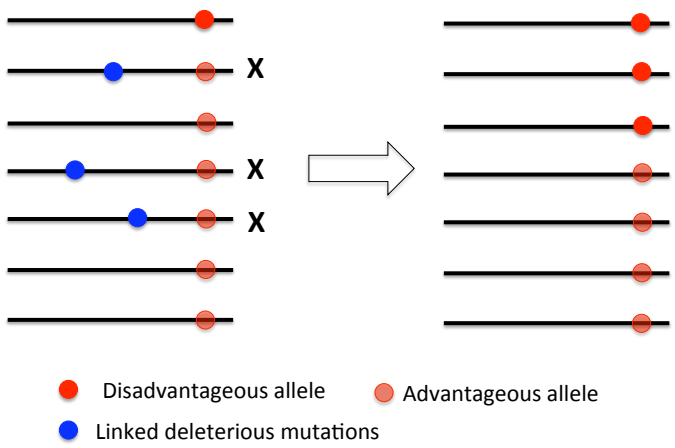

Figure 11-1: An illustration of background selection. The effective population size for the locus of interest (red) is reduced due to the presence of deleterious mutations at linked loci.

This effect is important when comparing codon conservation within different genomic regions. In particular, a codon far from the boundary of an exon will on average be in linkage disequilbrium with a greater number of amino acid loci than one lying near the boundary. Therefore, such a codon will evolve as if there were a smaller effective population size. And for a locus under weak selection (on the order of the inverse of N_e, as many synonymous changes are thought to be) even a small local change in N_e will have a noticeable effect in fixation probability.

A number of approximate models have been derived for quantifying the effect of interacting loci on population size under different scenarios [17]. Under some simplifying assumptions, it can be shown that for two loci with a recombination rate r between them, the reduction in N_e is given by $1 - q\frac{(hs)^2}{(r+hs)^2}$, where s is the selective coefficient at the linked locus and h the heterozygous effect [96]. The dynamics in more complex cases have been worked out as well, but frequently simulations are necessary. Of greatest importance for this section is whether position in an exon can

have a significant influence on the evolution of a synonymous change under realistic assumptions in *Drosophila*. The answer appears to be positive: under realistic assumptions, purifying selection on codons near coding region boundaries can be more effective and therefore lead to an increased conservation rate in such regions [75]. Further, a number of studies have shown empirical results that agree with such an effect, such as a reduced rate of codon bias within long genes in *Drosophila* [78]. Therefore, in the results of the section below, I consider background selection as at least a plausible alternative explanation that must be disproved.

Chapter 12

Analysis of Conservation in Coding Regions

12.1 Conservation in Feature Neighborhoods

In Section II, I showed how the MinoTar algorithm could be used to provide evidence for widespread microRNA ORF targeting and to make high confidence predictions of individual targets. In this chapter, I adapt the algorithm to search for more general elements that are preferentially conserved within coding regions. The basic procedure is similar to the analysis for miRNAs: the starting point for the analysis is the randomized multiple-species coding region alignments simulated by the algorithm. These alignments can be used to score the conservation of any general feature compared to the level expected by chance. In this chapter, the analysis utilizes conservation scores of individual codons and short motifs. In this case, the number of species a codon (or motif) is conserved to in the real alignment N_{real} is compared to the number of species the same codon (or motif) is conserved to in each of S random alignments N_i giving $p = \frac{\sum_i (N_{real} \leq N_i)}{S}$. A low p means the codon (or motif) is highly conserved

compared to the conservation level expected by chance.

As a first application, I used this procedure to evaluate conservation levels in the vicinity a number of genomic features, particularly splice sites and translation start/end sites. Previous work has shown the importance of exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs) in the regulation of splicing [32] in a number of genes. And an analysis in *Drosophila* has also shown evidence for a trade-off between codon usage and splice signals in the vicinity of splice junctions [104]. If such regions contain a greater fraction of functional sites, than a greater level of conservation would also be expected in these regions.

A first step in this analysis was to produce annotations of overlapping genomic features at all positions of all protein-coding transcripts. To do this, the genomic coordinates of all protein coding transcripts were downloaded from FlyBase [99] and compared against the genomic coordinates of other relevant features (Fig 12-1). Every position in all protein coding genes was annotated for overlap with any 5'UTR, intron, 3'UTR, or ORF of another transcript. Additionally, every position was annotated by its distance to the nearest transcription start and end sites, translation start and end sites, and splice junctions. Distances to splice junctions were further characterized by whether the splice junction was a 5' or 3' splice junction, and whether the junction was constitutive or alternate. Definitions of constitutive or alternate depend on annotations of transcripts in FlyBase and some fraction will be subject to change as either new transcript variants are discovered or currently annotated transcript variants are found to be spurious.

To judge preferential conservation in a particular class of sequences, the p-values for these codons were grouped and compared to the grouped p-values of a background set. The level of conservation above chance was judged by the height of the empirical cumulative distribution of p-values for the group of codons above the cumulative

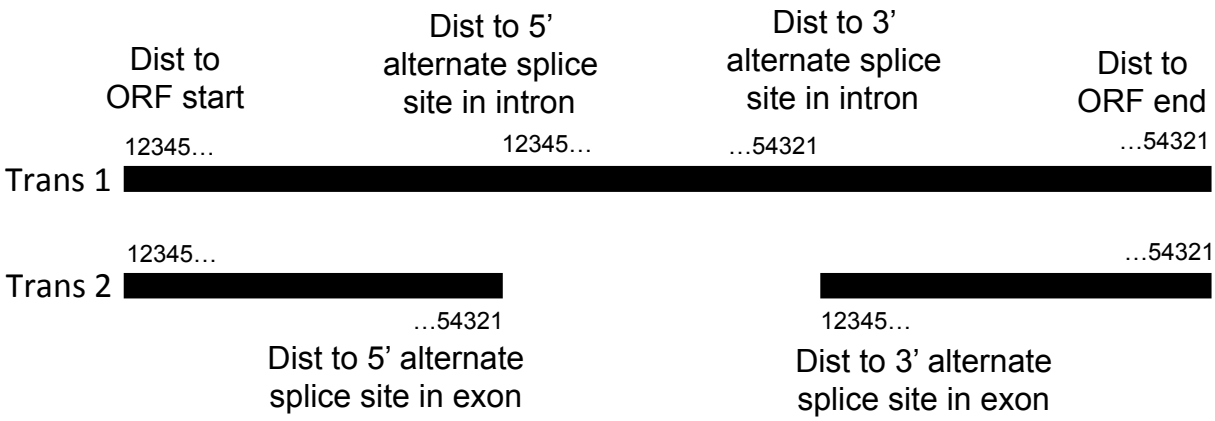

Figure 12-1: An illustration of the annotation of features. Every position of every protein coding transcript was annotated for overlap with and distance to a number of genomic features, including those related to transcription, translation and splicing.

distribution of the background set. Specifically, the fraction of p-values preferentially conserved was taken as the mean of the difference between these two curves between p-values of 0 and 0.4. Results were independent of the choice of these exact values- they were chosen simply to get a robust average over cutoffs where the codons were more conserved than by chance.

As a test, I first ran this procedure on codons in regions where at least 2 ORFs overlap in different reading frames (Fig 12-2). While known examples of this are rare in *Drosophila* (making up only $\sim 21,000$– about 0.1% – of protein coding nucleotides), these regions provide a good test-case for the procedure, as such regions would likely be under considerable additional pressure to conserve codons (a synonymous change in one reading frame would frequently be a non-synonymous change in another frame). The p-values for codons shared by transcripts in at least 2 different reading frames were compared to those of codons in all regions of all transcripts. As expected, a significant fraction of codons ($\sim 30\%$) were preferentially conserved.

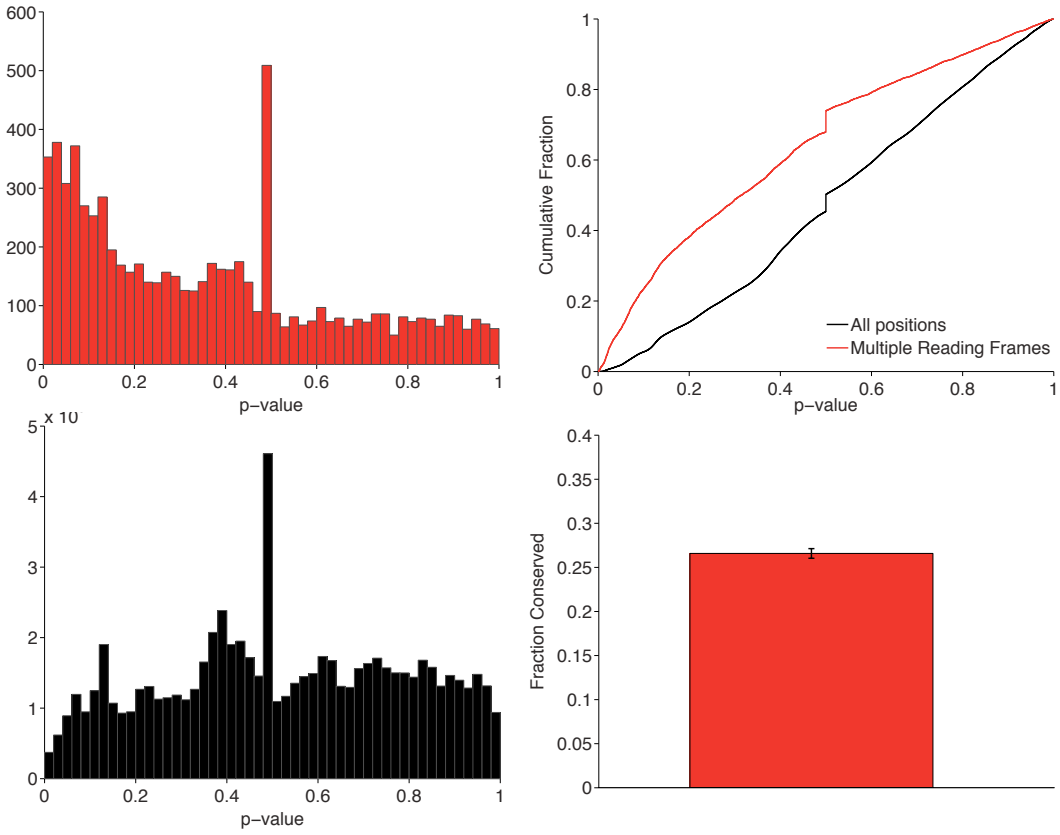

Figure 12-2: An illustration of the procedure for estimating the fraction of sites under preferential conservation for regions that encode proteins in at least 2 different reading frames. Such regions show significant additional constraint, with an estimated 30% of such bases under selection

Having seen the power of the procedure, I next applied it to codons at different distances from various genomic features. The p-values of codons were collected and binned by their distances to a number of features. For each of these features and all distances considered, a careful randomization procedure was followed to form a suitable background set. For every codon a given distance away from a feature, 10 codons were chosen from the same transcript to be part of the background set. This ensured that for every distance from a given feature, different transcripts were equally represented in both the feature set and the background set. This is important for controlling for conservation effects that occur on a transcript-wide basis, but are correlated with the presence of certain features (for example, an increased nucleotide conservation overall in transcripts with many splice junctions). Additionally, to avoid seeing effects that result from the non-random co-location of multiple features, codons from any one feature set (as well as it's background set) only included those at least 200 nucleotides away from all other features.

The results demonstrated widespread preferential conservation of codons near all features examined, but with highly varying effect strengths. Of weakest strength, were the constituent splice junctions (Fig 12-3). The 5' junctions showed high conservation of the few nucleotides directly adjacent to the splice junction (consistent with the highly biased nucleotide composition of these few nucleotides) but a quick dropoff in preferential conservation after about 5 nucleotides from the junction. Results for 3' junctions were similar but of even weaker effect. Of much stronger effect, however, were the alternate splice junctions (Fig 12-4). These showed stronger conservation extending over a longer range from the splice junction (the results are noisier due to the smaller counts). The strongest effect seen was for codons near translation start and ends.

In order to reduce noise and provide a comparison over longer distances, I redid

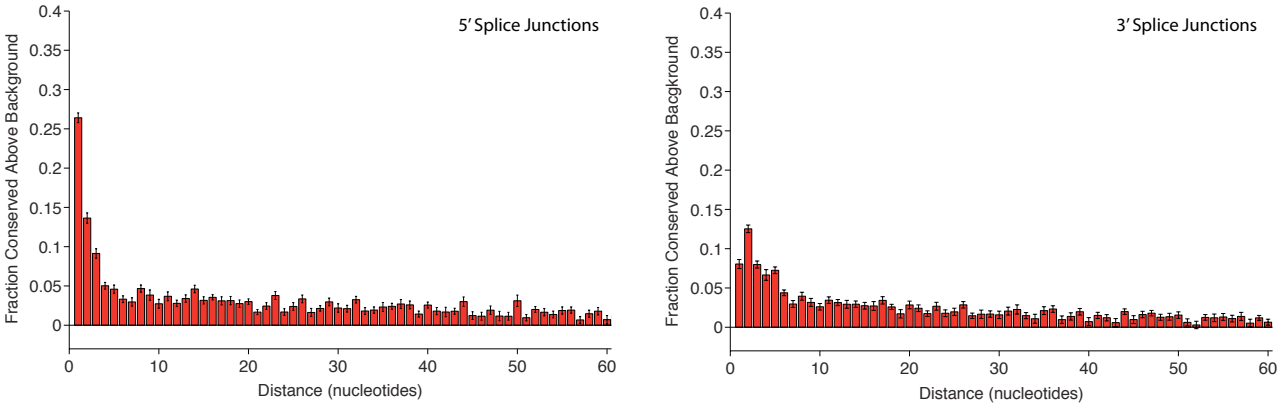

Figure 12-3: Plots of estimated fraction of sites under additional selection by distance to constitutive splice junctions. For constitutive splice junctions, the number of sites under additional conservation is only strong in the first few bases from the junction.

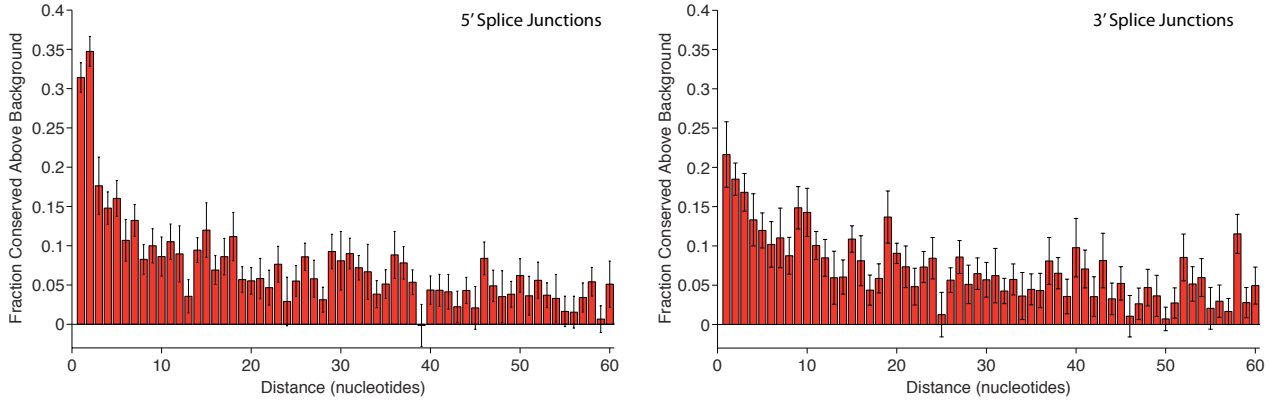

Figure 12-4: Plots of estimated fraction of sites under additional selection by distance to alternate splice junctions. The regions around alternate splice junctions show significantly higher conservation than those around constitutive splice junctions. This is in agreement with the prediction of a model where alternate splice junctions are under greater regulation than constitutive junctions.

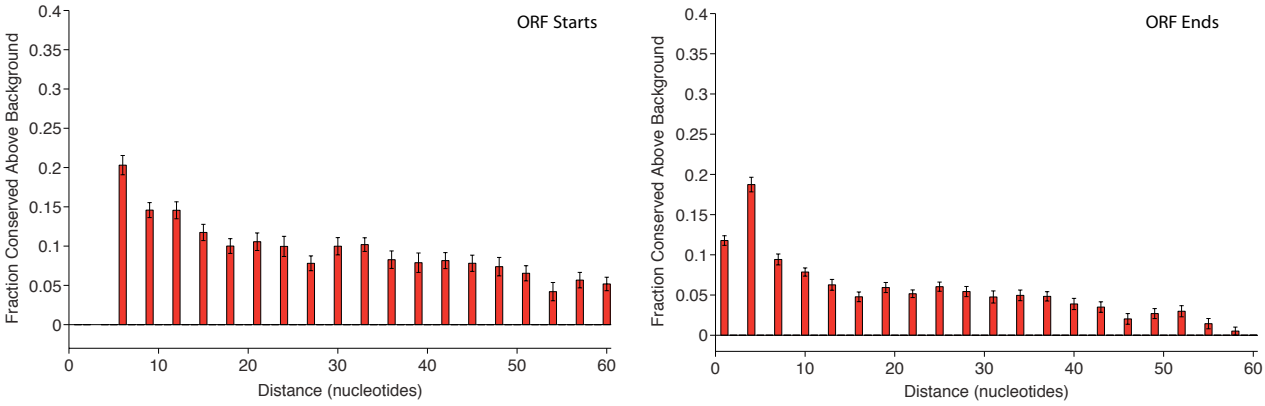

Figure 12-5: Plots of estimated fraction of sites under additional selection by distance to translation start and end sites. Such regions show strong additional conservation extending for 10s of bases from the site.

the analysis where codons were binned by distances of 20 from each feature, and conservation above background was assessed within each bin (Fig 12-6; Fig 12-7; Fig 12-8). These plots show clearly the dropoff of the effect with distance from the feature in each case. They also show clearly the stronger effect of preferential conservation in alternative versus constitutive splice junctions. This effect is consistent with a model in which alternate splice junctions are differentially regulated to a greater extent than constitutive splice junctions, which seems highly plausible. Under such a model, the regions near alternative splice junctions would be more likely to contain regulatory sequences than those near constitutive splice juncitons, resulting in a higher level of conservation.

12.2 Evidence against Background Selection

A potential concern with the above analysis is that all features considered are near the edges protein coding region. Therefore, an alternative explanation for the results

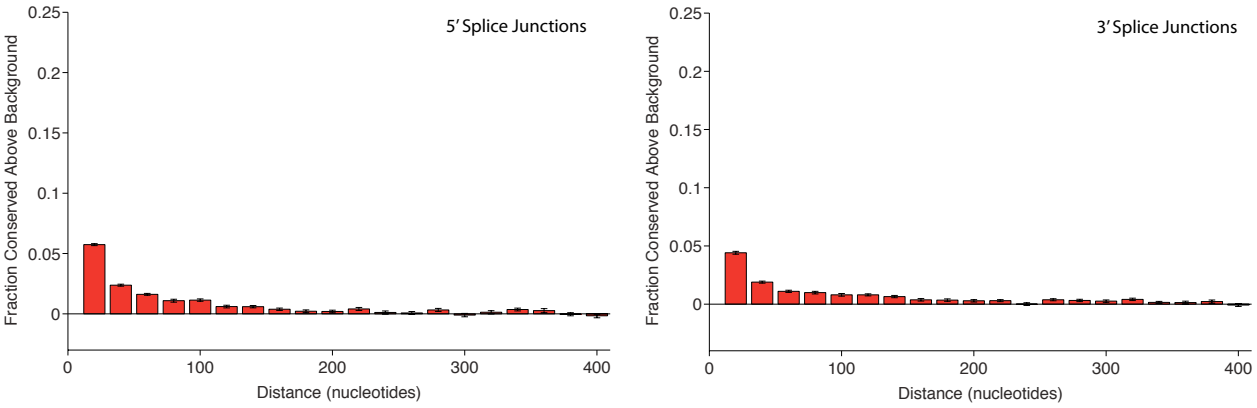

Figure 12-6: Plots of estimated fraction of sites under additional selection by distance to constitutive splice junctions, binned in distances of 20 basepairs.

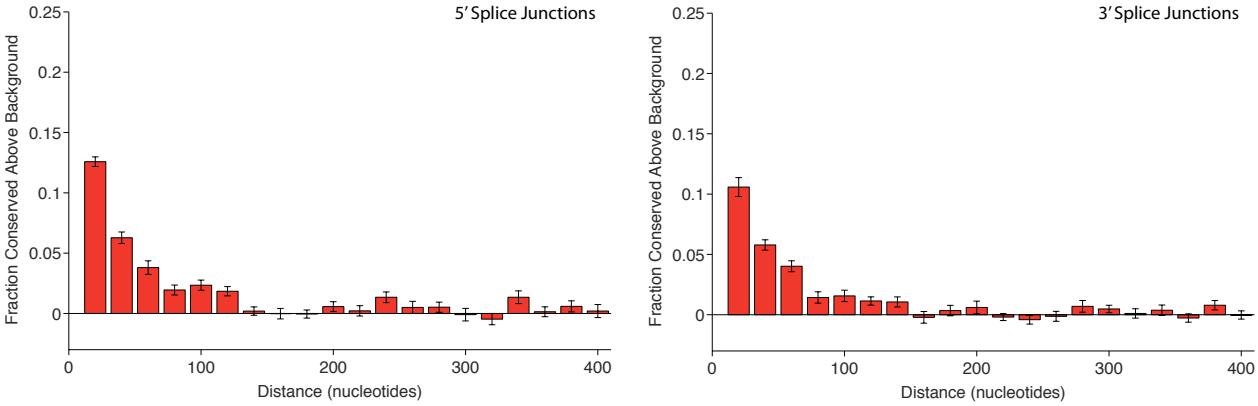

Figure 12-7: Plots of estimated fraction of sites under additional selection by distance to alternate splice junctions, binned in distances of 20 basepairs.

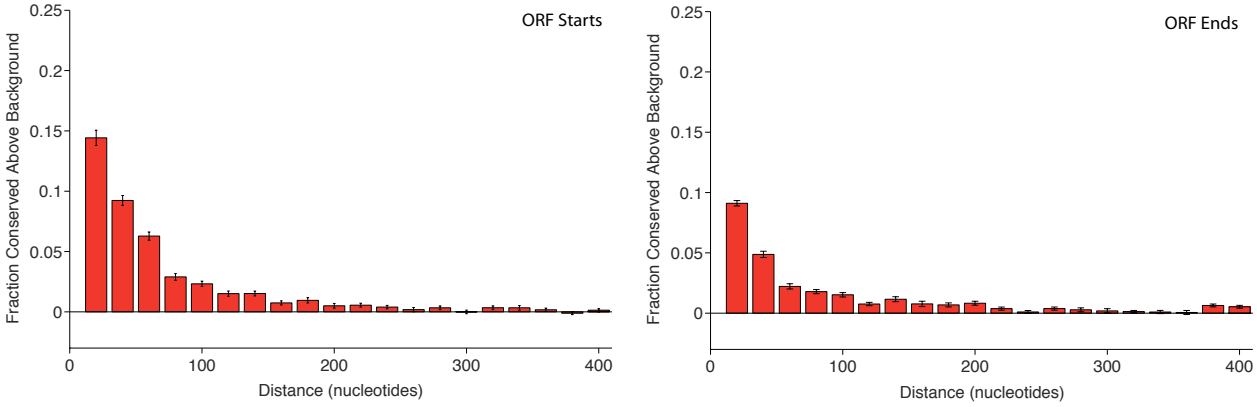

Figure 12-8: Plots of estimated fraction of sites under additional selection by distance to translation start and end sites, binned in distances of 20 basepairs.

is that preferential conservation in these regions isn't due to an increase in selection, but rather due to decreased background selection. Under such a model, if a codon is near the edge of a protein coding region, it will experience linkage with fewer sites under amino acid selection, and therefore be in a region of higher effective population size. In the case of weak selection, such as that causing codon bias, the change of thresholding due to such a change in population size can be important and result in an increased fraction of preferentially conserved sites.

I put forward two arguments why the above model is unlikely to explain the effects near different edges described above. First, I examined the conservation of codons that lay on the intronic (rather than exonic) side of an alternate splice junction (Fig 12-9). Again, excess conservation was seen in the regions around the splice junction, extending for many tens of nucleotides out from each junction. These results are consistent with selection on intronic splicing regulatory signals. And because there is no reduction in interference near the boundaries, the results cannot be explained by a decrease in background selection.

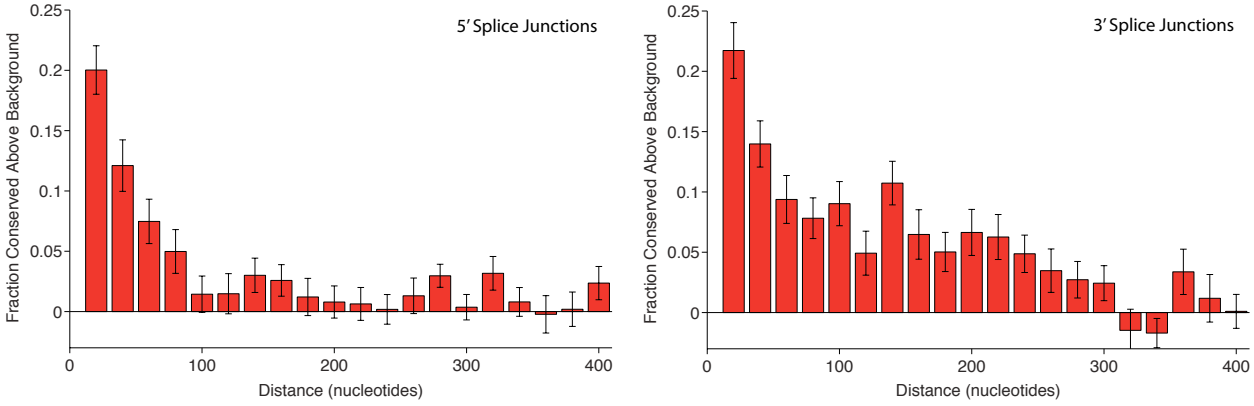

Figure 12-9: Plots of estimated fraction of sites under additional selection by distance to alternate splice junctions, binned in distances of 20 basepairs. Here, the codons in question lie on the intronic side of the splice junction in question. Unlike codons on the exonic side of genomic features, these codons wouldn't be expected to be show greater conservation due to reduced background. The conservation seen provides evidence that results aren't a result of background selection.

Second, if the increased conservation were the result of a decrease in background selection alone, one would expect an increased in optimal codon usage in these regions. However, exactly the opposite effect is seen. In all of the regions examined, codon usage was less optimal in the immediate vicinity of the feature. This suggests that the same selective pressure causing increased conservation in these regions competes with selection on codon choice, resulting in a reduction in the use of optimal codons.

Additionally, I examined evidence for region-specific conservation of motifs, which would indicate selection for sequences particular to each region. I looked at conservation of 6mers within neighborhoods of each of the different features. Each 6mer was evaluated both overall (in all regions) as well as within 25 nucleotides within translation starts/ends and 5' and 3' constitutive splice junctions. These neighborhoods were mutually exclusive: any sequence within 50 nucleotides of more than one genomic feature was excluded from the analysis. 6mers were chosen because they were

long enough to allow for conservation to be differentiated among different motifs, but short enough to allow for sufficient counts for accurate assessment of conservation of each 6mer. Similarly, constitutive rather than alternate splice junctions were used because there was only enough sequence to get sufficiently accurate conservation estimates for the constitutive junctions. In order to evaluate the conservation specific to each region, a region-specific conservation score was formed for all 6mers in each different region. This was given by the fraction of 6mers with p-value below a given cutoff (results shown are for $p = 0.1$) when in a given region minus the fraction overall. The sequences for each region were repeatedly randomly divided into 2 equal sets and the region-specific motif conservation was scored for each 6mer in each of these sets.

The results provided some evidence for the existence of region-specific conservation patterns, though of modest and varying strengths in the different regions. The regions that had the most consistently identified motifs were those near translation ends followed by those near 5' splice junctions. High-scoring motifs near translation starts also showed some self-consistency, but those near 3' splice junctions were around as consistent as expected by chance. However, it was difficult to completely remove all sources of bias from this analysis or to obtain high-confidence specifically conserved motifs from any of the regions.

12.3 Many Genes Contain Ultraconserved Regions

In the course of the above investigations, I observed a number of large regions with nearly complete codon conservation across all 12 *Drosophila* species. Here, I show that such regions are surprisingly prevalent. In order to compare conservation in larger regions, I first devised a simple way of scoring larger windows by their conservation. Conservation in a window of codons W was scored as the mean of the p vales for the

Window Size	FDR = 0.001	FDR = 0.01	FDR = 0.05
30	888	2218	3686
60	1422	2194	3377
120	1191	1976	2618
300	698	1232	1757

Table 12.1: Number of genes with at least one region of the given length (in nucleotides) and false-discovery rate. Thousands of genes have highly-conserved regions at very high confidence.

codons contained in that window: $p_W = \frac{1}{N}\sum_{i=1}^{N} p_i$. If large regions of conservation didn't exist and codon conservation was largely independent, such scores would be centered around 0.5 with a roughly normal distribution.

Instead of this distribution, the distribution of scores had a heavy left tail, indicating a significant number of regions with much higher conservation than expected by chance. This was evaluated for all non-overlaping windows of size 30, 60, 120 and 300 nucleotides (Fig 12-10). Scores above the median were well fit by a normal distribution, but those below the median weren't (Fig 12-11). The normal distribution fit on the scores above the median was used to determine a false discovery rate (FDR) for scores below the median. For each score the FDR was estimated by comparing the fraction of real scores below a given cutoff to the fraction expected given the fit distribution. Using this, it was possible to estimate the total number of windows conserved above chance for each of the different window sizes. This gave a relatively consistent but growing estimate of the number of windows with conservation greater than expected by chance (6%, 8%, 9% and 11% for 30, 60, 120 and 300 nucleotide windows, respectively). These windows occurred in many different genes and so thousands ($\sim 25\%$) of genes contained at least one window with high conservation (Table 12.1).

In order to illustrate the striking conservation of some of the genes with the

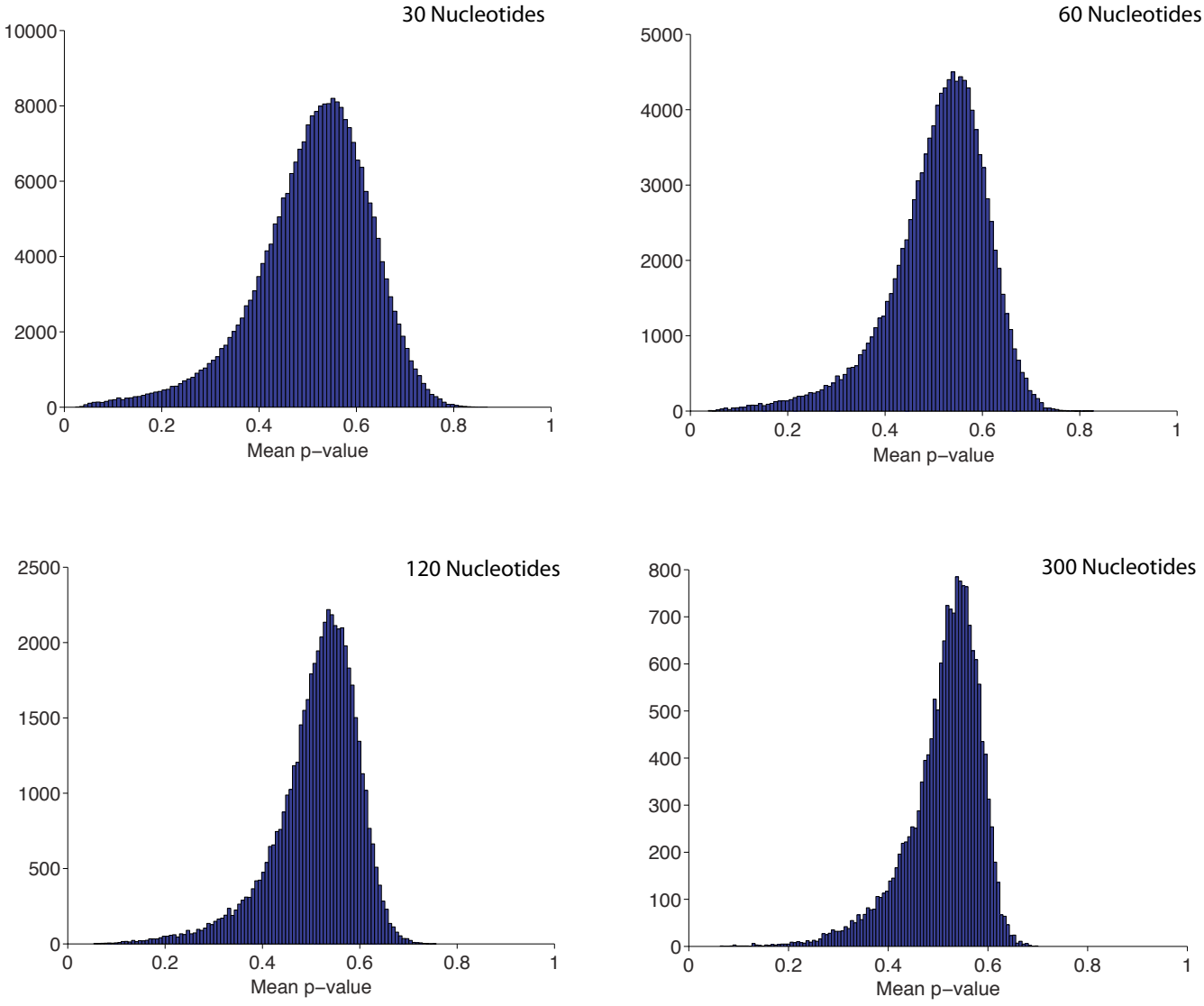

Figure 12-10: Histograms of mean p-value in windows of increasing sizes. All windows show a heavy left-tail of regions with higher than expected conservation.

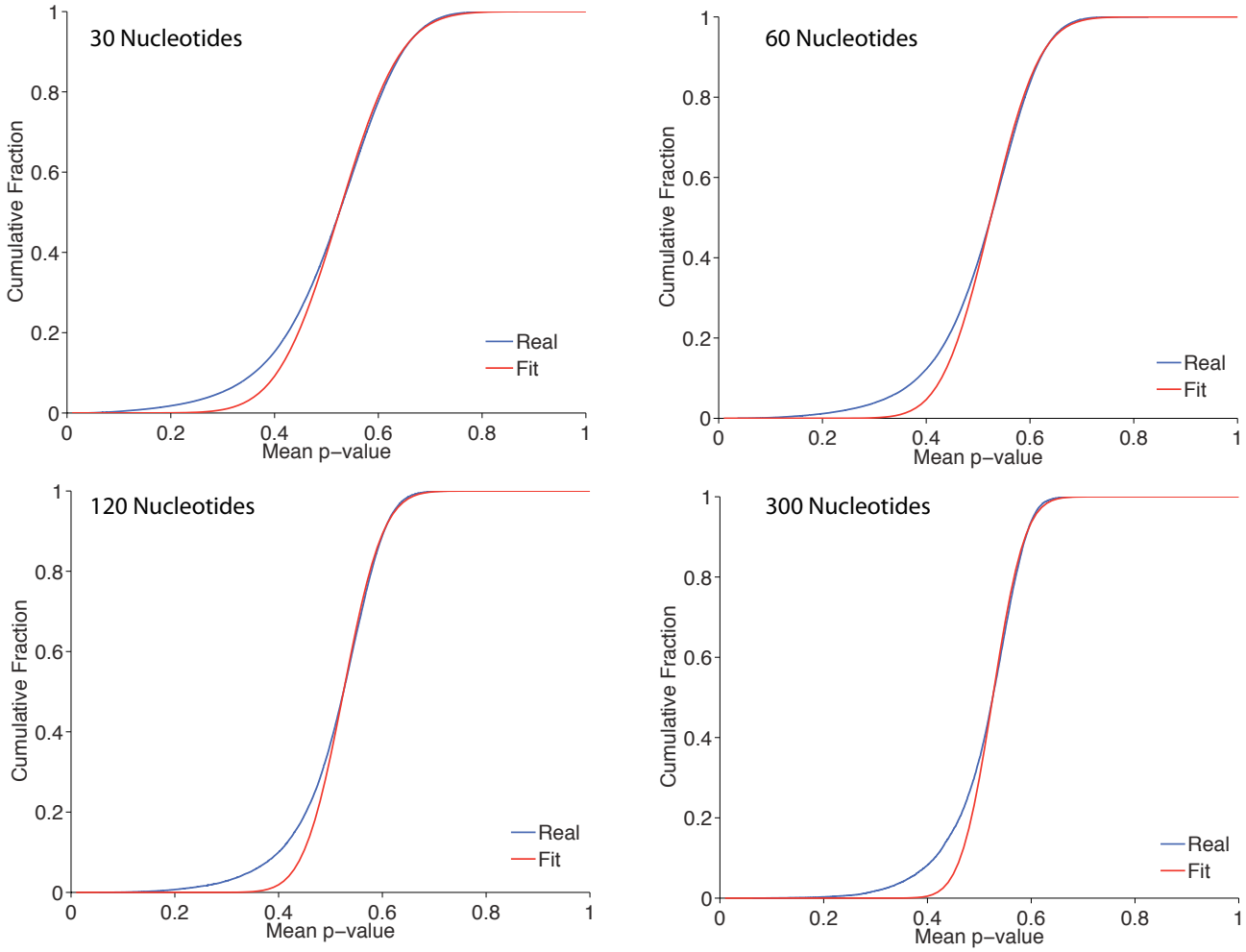

Figure 12-11: Cumulative distribution of mean p-values in windows of increasing sizes. Blue curves show the real distributions and red show a normal distribution fitted on the right side of the distribution. The area above the red curve and below the blue represents regions conserved beyond that expected by chance.

Gene	Ultraconserved Region			Overall Gene		
Name	Length	Mean p-value	Mean Codon Cons	Length	Mean p-value	Mean Codon Cons
Sh	342	0.07	10.8772	1716	0.148	10.8112
exba	144	0.078	11.5625	1269	0.3254	9.3239
kuz	138	0.0798	11.6957	3717	0.3882	8.6223
SK	132	0.0811	10	1575	0.2223	9.8267
bru-3	132	0.0814	11.8409	1269	0.2574	10.4539
A2bp1	174	0.0823	11.7069	2598	0.3826	8.0439
tutl	135	0.0824	11.7778	2709	0.3995	8.6633
mmd	75	0.0831	11.72	2514	0.5343	7.9833
Ca-beta	129	0.0835	11.7442	1593	0.3285	9.3804
orb	135	0.0852	11.6	2748	0.5038	6.9061
para	177	0.0855	11.7119	6396	0.2756	9.7223
slo	156	0.0856	11.8846	3552	0.285	10.1672
Arf79F	135	0.0868	11.8222	549	0.2946	10.2131
CG34400	141	0.0884	10.2979	2973	0.3316	8.8789
Ih	144	0.0886	11.8958	3873	0.3376	9.1325

Table 12.2: Features of the genes containing the top 15 ultraconserved windows of length 120 nucleotides. Listed lengths indicate the largest region that contains a mean p-value smaller than the cutoff (here a cutoff of 0.15 was used).

most conserved windows, Table 12.2 shows a list of 15 genes with the most conserved windows of 120 nucleotides. A number of these genes, such as orb (oo18 RNA-binding protein), mmd (mind-meld) and tutl (turtle) show only specific exons or portions of exons that are very highly conserved (Fig 12-12). Others, such as Arf79F (ADP ribosylation factor 79F) show more extensive conservation with regions of particularly high conservation, and still others such as sh (short winged) show very high nucleotide conservation across almost the entire gene (Fig 12-13).

12.4 The Function of Ultraconserved Regions

Having seen the large number of regions with extensive nucleotide conservation, I next sought to investigate the characteristics of these regions. As a first step, I examined the expression of those genes with highly conserved regions of different lengths. If ultraconserved regions were due to the same selective pressures leading to codon bias, one would expect such regions to be enriched for highly expressed genes. Indeed, genes

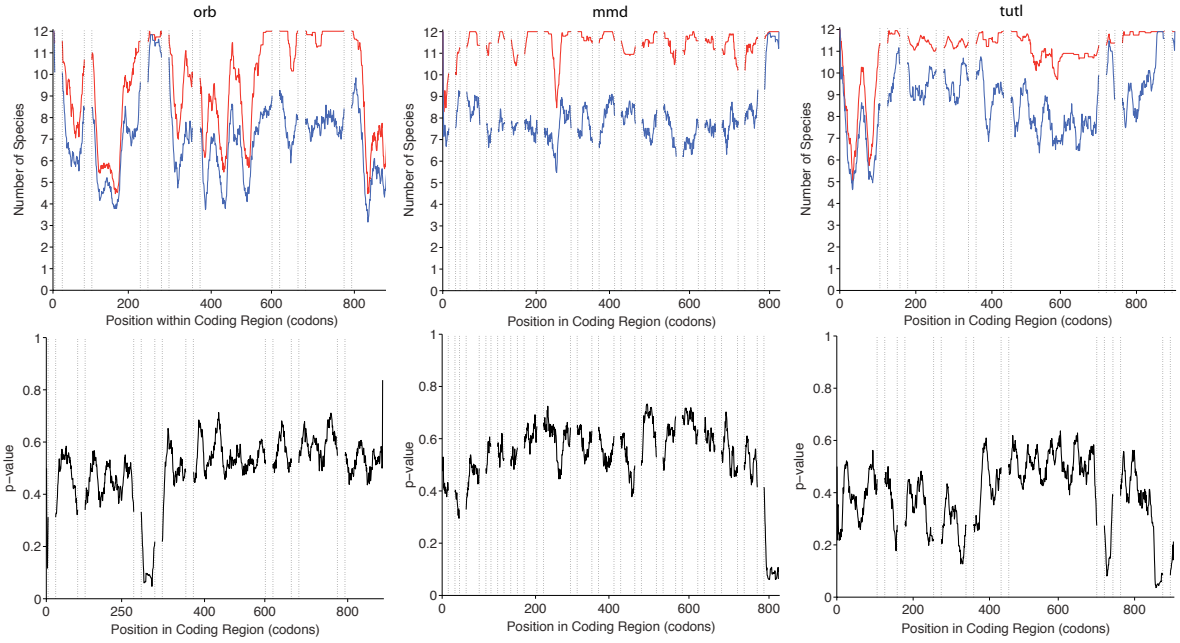

Figure 12-12: Conservation plots for orb, mmd and tutl genes. Top: Average number of species a feature is conserved to (red: amino acid, blue: codon) in a sliding window of 120 nucleotides. Bottom: average p-values in a sliding window of 120 nucleotides. These genes show only modest codon conservation overall but contain regions of exceptionally high codon conservation. Black dashed-lines show exon boundaries.

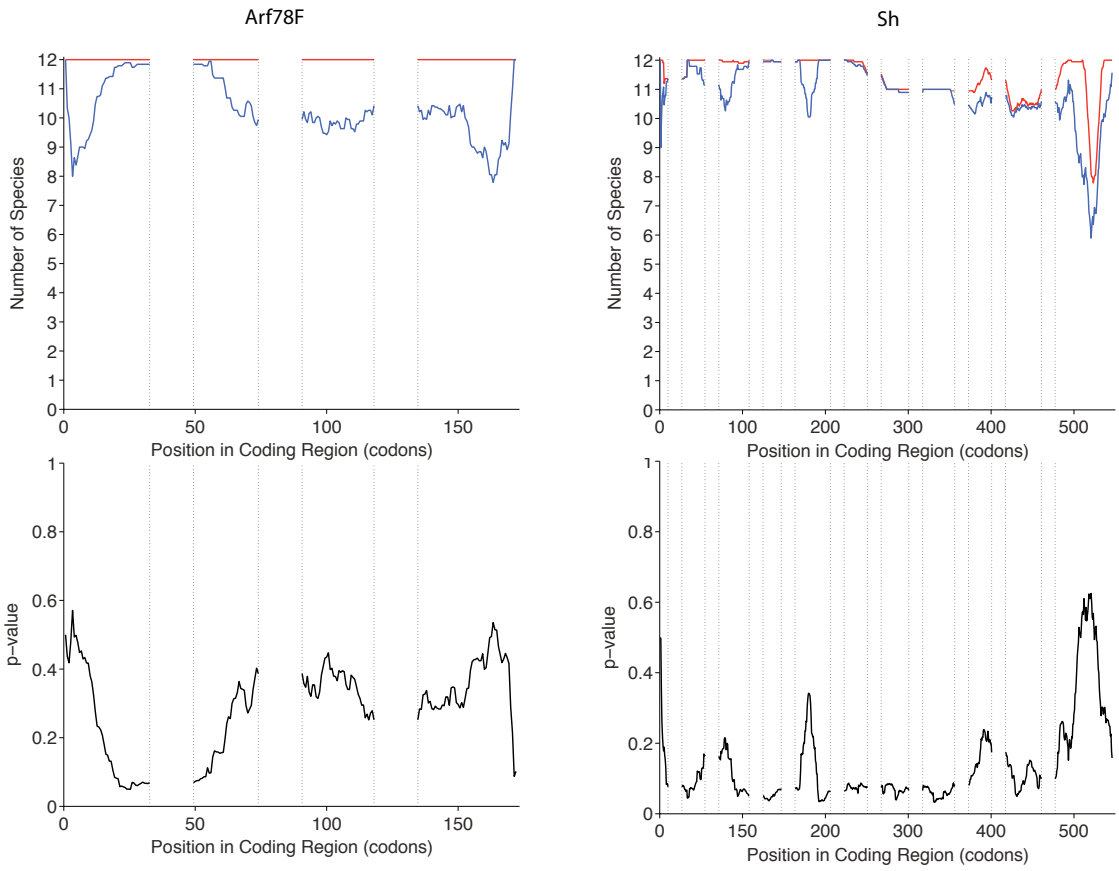

Figure 12-13: Conservation plots for Arf79F and Sh genes. Top: Average number of species a feature is conserved to (red: amino acid, blue: codon) in a sliding window of 120 nucleotides. Bottom: average p-values in a sliding window of 120 nucleotides. Arf79F shows high conservation overall, but contains a single region of exceptionally high conservation. Sh shows exceptionally high conservation along almost the entire gene. Black dashed-lines show exon boundaries.

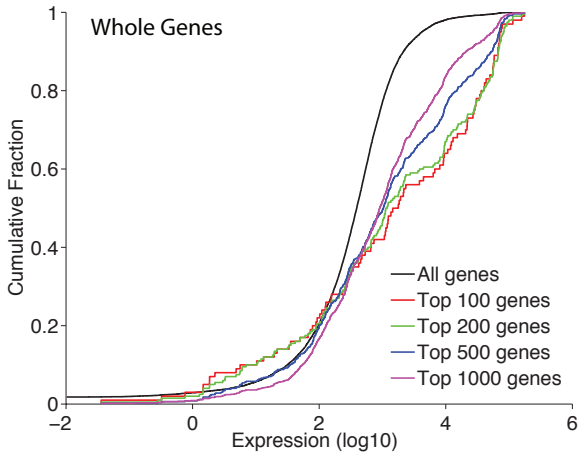

Figure 12-14: Cumulative distribution for expression of genes with the highest codon conservation overall. Highly conserved genes overall show a strong preference to be highly expressed.

with high overall codon conservation (low overall mean p-value) showed significantly elevated expression levels (Fig 12-14). In contrast, genes containing ultraconserved regions show showed little or only modest enrichment for high expression (Fig 12-15). Expression values for all genes were derived from the developmental expression survey conducted in [37]. An overall expression value for all genes was derived by summing expression across all developmental time points measured.

This was further confirmed by an analysis of GO term enrichment [14]. Enrichment was evaluated for sets of genes with both overall high conservation and with the most conserved regions of size 120 nucleotides. In order not to confound the analysis by any correlation with amino acid conservation, genes were restricted to those with average amino acid conservation across the whole gene of at least 11 out of 12 species. The background set consisted of all genes with average amino acid conservation of 11 out of 12 species. The set of genes with highest conservation overall were only enriched for terms related to highly expressed genes: largely ribosomal proteins and

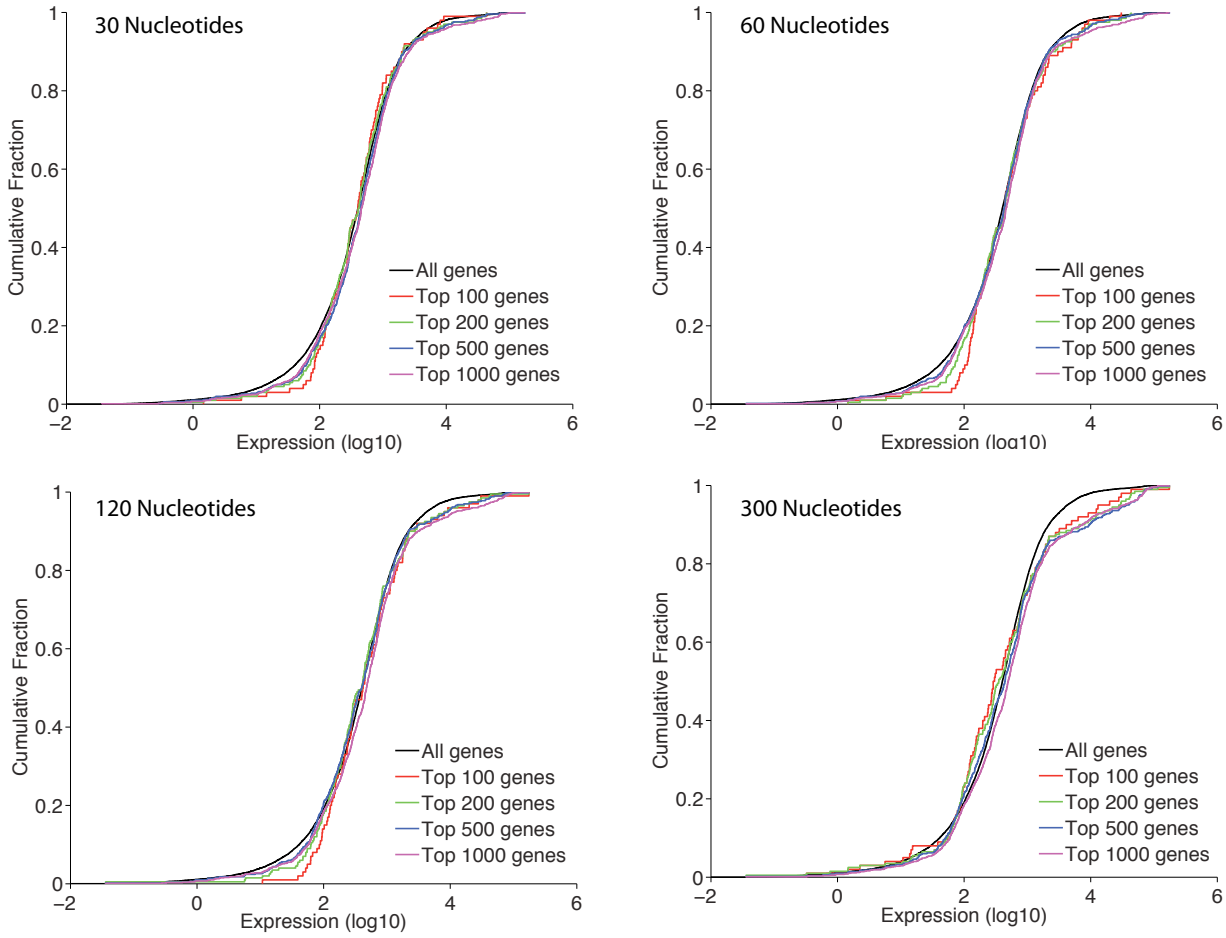

Figure 12-15: Cumulative distribution for expression of genes with the highest codon conservation in windows of different sizes. Expression levels for genes are similar to those for genes overall.

components of the cytoskeleton (Table 12.3). In contrast, terms for the genes with the most conserved 120 nucleotide windows showed enrichment for a variety of terms, particularly a number of developmental terms, and a number of neural processes (Table 12.4). Post-transcriptional regulation, including extensive alternate splicing and sub-cellular localization, is known to be particularly widespread in neurons [76]. Because the list of genes containing ultraconserved windows was slightly biased toward those with greater length (as it would be for genes chosen containing any feature), I conducted a test to make sure this bias couldn't account for the enrichment of GO terms seen. When examining a list of genes chosen to have the same length distribution and constraint of average amino acid conservation of at least 11 out of 12 species, no GO terms were found to be significantly enriched.

An investigation into features of the ultraconserved regions suggested that a significant fraction of them may be involved in splicing (Fig 12-16). When compared to all regions, ultraconserved regions were signficantly more likely to be close to splice junctions, particularly alternate splice junctions (within 100 nucleotides: 5' constitutive: 0.41 ± 0.02 vs 0.17; 3' constitutive: 0.36 ± 0.02 vs 0.16; 5' alternate: 0.16 ± 0.02 vs 0.02; 3' alternate: 0.16 ± 0.02 vs 0.02, all p-values $< 10^{-16}$ binomial test). In contrast, ultraconserved regions were little or no more likely to be close to translation start or end sites. In addition, genes with ultraoconserved sites were far more likely to be highly spliced (average number of coding region exons: 9.2 ± 0.3 vs 3.9, $p < 10^{-100}$ Mann-Whitney U Test), and the exons within which the ultraconserved regions lay were significantly smaller (530 ± 110 vs 1220, $p < 10^{-100}$ Mann-Whitney U Test). Ultraconserved regions showed evidence of being involved in other functions as well. For example, ultraconserved regions were far more likely to contain a region under multiple reading frames (0.04 ± 0.01 vs 0.001, $p < 10^{-16}$ binomial test). It seems likely that more unannotated regions under multiple reading frames may be

GO Term	P Value
Biological Processes	
GO:0006412 translation	9.62e-15
GO:0051231 spindle elongation	4.81e-09
GO:0000022 mitotic spindle elongation	4.81e-09
GO:0044267 cellular protein metabolic process	3.88e-08
GO:0007010 cytoskeleton organization	2.96e-07
GO:0007052 mitotic spindle organization	1.87e-06
GO:0034645 cellular macromolecule biosynthetic process	4.72e-06
GO:0019538 protein metabolic process	4.75e-06
GO:0009059 macromolecule biosynthetic process	5.23e-06
GO:0007051 spindle organization	5.45e-06
GO:0044249 cellular biosynthetic process	1.26e-05
GO:0000279 M phase	1.99e-05
GO:0000226 microtubule cytoskeleton organization	3.11e-05
GO:0000278 mitotic cell cycle	3.19e-05
GO:0009058 biosynthetic process	4.82e-05
GO:0022402 cell cycle process	7.77e-05
GO:0022403 cell cycle phase	7.82e-05
GO:0010467 gene expression	3.57e-04
GO:0007049 cell cycle	4.80e-04
GO:0007017 microtubule-based process	5.35e-03
GO:0044260 cellular macromolecule metabolic process	6.54e-03
Cellular Components	
GO:0022626 cytosolic ribosome	3.39e-24
GO:0044445 cytosolic part	2.21e-21
GO:0005840 ribosome	1.16e-19
GO:0005829 cytosol	2.38e-18
GO:0022627 cytosolic small ribosomal subunit	1.08e-11
GO:0022625 cytosolic large ribosomal subunit	1.01e-09
GO:0015935 small ribosomal subunit	1.01e-09
GO:0030529 ribonucleoprotein complex	1.72e-09
GO:0044444 cytoplasmic part	6.21e-09
GO:0015934 large ribosomal subunit	4.93e-08
GO:0043228 non-membrane-bounded organelle	5.81e-08
GO:0043232 intracellular non-membrane-bounded organelle	5.81e-08
GO:0005737 cytoplasm	1.91e-06
GO:0043292 contractile fiber	6.40e-04
GO:0044449 contractile fiber part	4.78e-03
GO:0005811 lipid particle	6.26e-03
Molecular Functions	
GO:0003735 structural constituent of ribosome	4.52e-20
GO:0005198 structural molecule activity	3.60e-18

Table 12.3: The enriched GO terms for genes with highest overall codon conservation reflect genes of high expression level.

GO Term	P Value
Biological Processes	
GO:0009605 response to external stimulus	4.57e-08
GO:0032501 multicellular organismal process	6.99e-07
GO:0050896 response to stimulus	2.84e-06
GO:0007275 multicellular organismal development	1.53e-05
GO:0030154 cell differentiation	2.18e-05
GO:0048869 cellular developmental process	6.39e-05
GO:0042221 response to chemical stimulus	1.28e-04
GO:0032502 developmental process	1.47e-04
GO:0042330 taxis	2.29e-04
GO:0048468 cell development	3.97e-04
GO:0006811 ion transport	4.73e-04
GO:0006935 chemotaxis	5.95e-04
GO:0007411 axon guidance	5.95e-04
GO:0048731 system development	8.20e-04
GO:0040011 locomotion	8.36e-04
GO:0007610 behavior	1.41e-03
GO:0007399 nervous system development	3.10e-03
GO:0032989 cellular component morphogenesis	3.18e-03
GO:0008037 cell recognition	3.20e-03
GO:0008038 neuron recognition	3.20e-03
GO:0007409 axonogenesis	3.58e-03
GO:0048856 anatomical structure development	4.76e-03
GO:0008344 adult locomotory behavior	4.89e-03
GO:0048589 developmental growth	5.79e-03
GO:0003008 system process	8.90e-03
GO:0030030 cell projection organization	8.98e-03
Cellular Components	
GO:0071944 cell periphery	8.75e-09
GO:0005886 plasma membrane	1.76e-07
GO:0034702 ion channel complex	2.50e-07
GO:0044459 plasma membrane part	1.41e-06
GO:0031226 intrinsic to plasma membrane	1.09e-04
GO:0005887 integral to plasma membrane	1.09e-04
GO:0016020 membrane	3.80e-04
GO:0034703 cation channel complex	7.60e-04
GO:0016021 integral to membrane	4.60e-03
GO:0031224 intrinsic to membrane	5.86e-03
Molecular Functions	
GO:0022836 gated channel activity	5.51e-09
GO:0005216 ion channel activity	3.28e-07
GO:0022838 substrate-specific channel activity	3.28e-07
GO:0022803 passive transmembrane transporter activity	7.36e-07
GO:0015267 channel activity	7.36e-07
GO:0005230 extracellular ligand-gated ion channel activity	3.75e-05
GO:0005231 excitatory extracellular ligand-gated ion channel activity	1.25e-04
GO:0015075 ion transmembrane transporter activity	1.05e-03
GO:0005261 cation channel activity	1.46e-03
GO:0022834 ligand-gated channel activity	1.61e-03
GO:0015276 ligand-gated ion channel activity	1.61e-03
GO:0022891 substrate-specific transmembrane transporter activity	2.35e-03
GO:0022857 transmembrane transporter activity	2.80e-03
GO:0060089 molecular transducer activity	4.51e-03
GO:0004871 signal transducer activity	4.51e-03

Table 12.4: The enriched GO terms for genes with highest window codon conservation (window size of 120 nucleotides) represent diverse processes, particularly those involved in neural development and function.

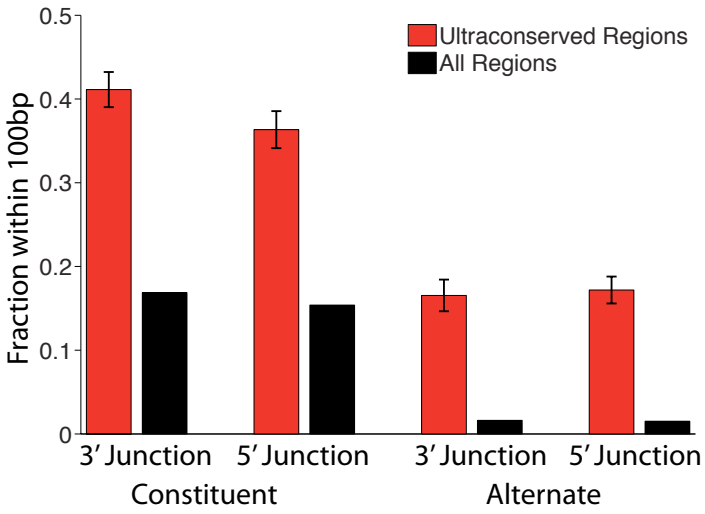

Figure 12-16: Ultraconserved regions are far more likely to be near splice junctions than expected by chance. The effect is stronger for alternate than for constitutive splice junctions. Ultraconserved regions are no more likely to lie near translation start or end sites.

represented within the ultraconserved regions. However, it was not possible to identify any high confidence candidates from the sequence alone. Similarly, an attempt was made to look for additional secondary structure in ultraconserved regions, but was unsuccessful in finding such a signal.

As with conservation in feature neighborhoods, an alternative explanation for ultraconserved regions is that these regions are merely under reduced background selection due to lying within shorter exons, and so weak selection towards optimal codons is more effective in these regions. To show that this isn't the case, I evaluated the codon frequency usage in ultraconserved regions. For each codon C, I defined a codon bias score by $B_C = M_{AA(C)} \frac{N_C}{N_{AA(C)}}$, where $AA(C)$ is the amino acid encoded by C, $M_{AA(C)}$ is the multiplicity of the amino acid (the number of codons encoding that amino acid), and N_C and $N_{AA(C)}$ give the number of counts over all genes for

the codon and its associated amino acid. A codon bias score higher than 1 means a codon tends to be favored overall, and a score lower than 1 means that a codon tends to be disfavored. Codon bias in a region was taken as the mean of the codon bias scores for all codons in that region.

I first verified that codon usage bias increases with gene expression, as has been previously observed [92]. Then I looked at the codon bias of 120 nucleotide regions ranked by their conservation. These regions were binned into groups of 50, and the codon bias score for each bin was taken as the mean over all regions included in each bin. In all cases, only regions with average amino acid conservation greater than 11 were considered in order to prevent any confounding of the results.

While codon bias rose at first with rising conservation, the regions with the highest conservation actually had lower codon bias than average (Fig 12-17). In contrast to the result expected under decreased background selection, this suggested that such regions are under additional selection that interferes with the selection to maintain optimal codon choice. Because lower frequency codons would be more surprising when highly conserved, it was possible that this observation merely reflected the fact that the most surprisingly conserved regions would consist of low-frequency codons. Therefore, I repeated the analysis using average number of species of codon conservation as the measure of conservation (rather than p-values) and verified that the effect was the same (Fig 12-17). Additionally, the results were the same when restricted to the Sopophora clade, among which codon bias is highly consistent, and when restricting to four-fold degnerate sites.

12.5 Open Questions

Traditionally, genomes have been largely viewed as divided into coding and non-coding DNA and, with a small number of exceptions (for example [18], [68]), most

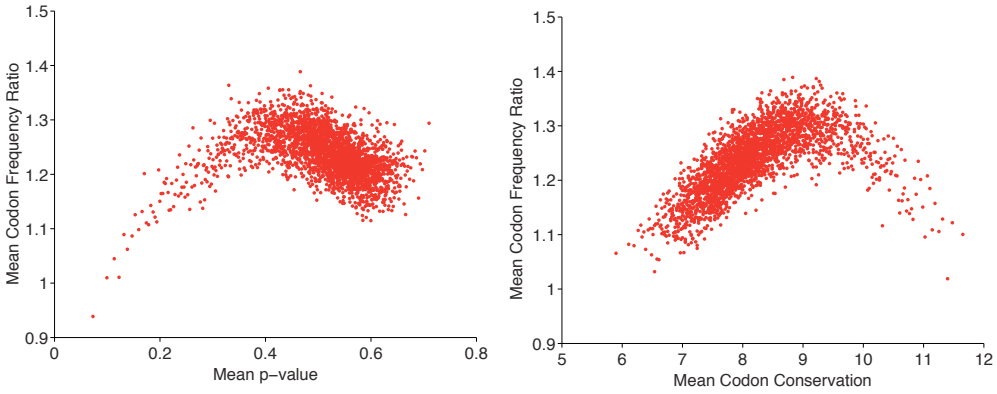

Figure 12-17: The use of preferred codons is greater for regions with moderately elevated conservation but not for those with very high conservation. Shown are the mean codon frequency ratios versus conservation (Left: mean p-value; Right: mean number of species of codon conservation) for windows 120 nucleotides in length. This result suggests highly conserved regions are under selection that interferes with selection on codon usage.

computational tools for analyzing regulatory signals in DNA have been designed to work only within non-coding DNA. Such tools will frequently fail to work within coding DNA. The techniques developed in this section, however, have been expressly derived for application to coding DNA. The results derived under these techniques have shown that a surprisingly large fraction of coding DNA appears to be under selection for additional function. It appears that regulatory signals are denser in coding DNA than has previously been appreciated. Indeed, it has recently been suggested that the genetic code is, in fact, optimal for encoding additional non-coding information [60].

The work described here is only a first step, and while it has been possible to show that conserved regulatory signals appear common, it is unknown what most of the regulatory signals do. A large number of regulatory events happen in the lifetime

of a gene, from the recruitment of transcription factors and chromatin modifications at the DNA level to regulation of splicing, degradation, localization and translation at the RNA level. It seems likely that different regulatory regions in coding DNA are involved in many of these processes. As more detailed hypotheses about addtional regulatory processes are developed, it will become possible to test these specific hypotheses using the tools developed in this section.

Part V

General Conclusion

Chapter 13

The Future

Where do we go from here? In this section, I attempt an answer. First, I describe the implications of a number of results from Sections II - IV along with a number of open questions stemming from that work. Then I take a step back and offer a view on what the future holds for the broader field of comparative genomics.

One of the major results presented in this thesis is that non-coding regulation within coding regions is widespread. Further, much of this regulation can be predicted with high confidence, even for short signals such as microRNA binding sites. (This is perhaps surprising as the number of synonymous sites that are free to mutate within such regulatory elements is small– in some cases only 2 or 3 bases. This demonstrates the importance of having genomic sequences for a large number of related species. With a sufficient number of species, high power can be achieved even on elements with limited evolutionary freedom.) The evidence presented here suggests that greater attention should be paid to regulation within coding regions. Such regulation, while often neglected, has the potential to cause important phenotypic consequences. Indeed, in a handful of recent studies, synonymous mutations have been implicated in diseases [16]. In one, a microRNA binding site in the coding

region of the IRGM gene has been suggested to play a role in Crohn's disease [58, 10]. Tools such as those developed here may prove pivotal in guiding the search for other cases of such regulatory relationships.

A number of questions remain in the study of regulation in coding regions. First, while this thesis has implicated specific regulatory processes in the function of some conserved elements, the function of most regions remains unknown. Of perhaps greatest interest are the ultraconserved elements in *Drosophila*– it is still unknown what precise function the vast majority of these serve. Building off of the work in this thesis, a number of experiments could be done to investigate these regions. The simplest would be to introduce new versions of the genes containing these regions mutagenized to change the nucleotide sequence while leaving the protein sequences unchanged. Then changes in the processing (for example, changes in the splicing or localization) of the mutant versions of such genes could be studied.

The results in Section III demonstrate a novel role for sequence repeats in microRNA regulation. An open question is the extent to which such an effect may occur for other regulatory processes as well. Given the extent of repeats in the genome and the large number of regulatory processes utilizing short motifs, it seems likely that other instances of such an effect exist outside of microRNA regulation. As more instances of regulatory motifs are discovered for these processes, it will be interesting to investigate if this is the case.

A more general question concerns the relative advantages and disadvantages for a gene to harbor regulatory signals within coding rather than non-coding DNA. Coding DNA is already under strong selective pressure at the protein level, which should make it more difficult to quickly evolve regulatory signals within such DNA. Additionally, coding regions may provide additional restrictions on regulation. This appears to be the case for microRNA targeting, where regulation can be inhibited by translating

ribosomes. On the other hand, as discussed in Section II, this restriction suggests the possibility for unique and exciting regulatory relationships (such as the induction of ORF targeting under starvation conditions). Additionally, regulation in coding DNA may provide other advantages as well. Many genes contain alternative exons with functional domains only included in certain variants of the protein. By harboring a regulatory element within an alternate exon, a regulatory relationship can be confined to act only on the variants of the protein containing that domain.

Additionally, the prevalence of conserved non-coding elements in coding DNA suggests an interesting problem that has, to my knowledge, not been adequately addressed. A number of the tests for positive selection at the amino acid level (as well as many techniques in population genetics) depend on the assumption that synonymous changes are neutral [86]. However, it isn't clear how such tests will perform when this assumption breaks down. In regions where a large fraction of synonymous changes are under purifying selection, it's possible that the results of such tests will be significantly distorted.

All of this work has been done within the early years of comparative genomics. Many of the results presented here were only possible given the relatively large numbers of sequenced species genomes. However, the number of species already sequenced pales in comparison with the number that remain. While there are currently 12 Drosophila and 29 mammalian sequenced genomes, there are thousands of species in both clades that can eventually be sequenced. What kinds of analysis can be done with thousands of species that can't be done with ten of species? I won't claim to have a complete answer. However, it seems clear that many of the problems that people work on today (such as motif discovery) will not require anything near thousands of species before techniques are no longer gaining anything from more species.

I suggest that one way to identify fruitful avenues for such research will be to

look for areas where one must choose from among many multiple hypotheses. In such situations, multiple hypothesis correction would make it impossible to identify the true hypothesis without a large number of independent evolutionary time points. I see potential for this in any application where there is interaction between multiple loci, and compensatory mutations can occur. One example is the identification of the detailed rules of cis-regulatory modules, where multiple transcription factors interact both with DNA and each other in order to activate or repress transcription. Another is the identification of pairs of interacting proteins (and nucleic acids) as well as the identification of the particular residues involved in mediating an interaction. These suggestions likely only hint at what will be possible. As our effort to record and map out the biological world continues, many capabilities will likely arise that we can't even imagine today.

Part VI

Appendices

Appendix A

Additional Methods

A.1 Methods in Drosophila Studies

This section describes a number of both the computational and experimental techniques used in Section II more detail.

A.1.1 Binning Gene Regions by Conservation Level

In order to also account for and remove gene-region specific variations in codon conservation rates, the MinoTar algorithm first bins regions of coding DNA by conservation level and learns a different background model for each bin. A background model for codon conservation was first trained on all ORF sequences, which was used to produce a p-value at every codon instance in all genes. Every codon instance was then assigned a region conservation score given by the mean of the p-values in a window of 120 nucleotides (40 amino acids) centered at that codon. Codons were then sorted by their region conservation scores, and placed according to these scores into 5 equally spaced bins. A new background model was learned separately for each of these bins. When evaluating k-mer conservation, a region conservation score was evaluated for

the 120 nucleotides centered at that k-mer, and the bin corresponding to that score was used as the background model.

A.1.2 Alignments and miRNA Sequences

Multiple species alignments and genome annotations for Drosophila (12-way) and Human (17-way) were downloaded from the UCSC genome browser [50]. Three fish species were excluded from analysis because of significant non-aligned sequence. For coding regions, genome annotations were used to exclude all regions overlapping a 3'UTR or 5'UTR for any transcript. In all statistical calculations, regions overlapping more than one transcript were only used once to avoid over-counting. For analysis of 5'UTRs, annotations were used to remove all regions overlapping ORFs or 3'UTRs. Alignments of 3'UTRs for both Drosophila and Humans were taken from the TargetScan website [55]. Alignments of promoters (taken as the 500 bp upstream of Transcription Start Sites) were downloaded from UCSC genome browser. Regions overlapping ORFs, 3'UTRs or 5'UTRs of any transcripts were removed. Mature miRNA sequences as well as annotations of miRNA conservation were downloaded from the TargetScan website.

A.1.3 Assessing Motif Conservation

In ORFs, the conservation rate of a set of k-mers for a given cutoff value p_{cutoff} was determined by the fraction of instances achieving $p < p_{cutoff}$ among those instances with $p_{min} < p_{cutoff}$. Similarly, in non-coding regions the conservation rate of a set of k-mers was assessed by the fraction of instances conserved to at least M out of N species among those instances with aligned sequence in at least M out of N species. In both cases, background sets of k-mers, consisting of all k-mers with identical nucleotide content as each of the k-mers in the true set, were used to judge expected background

levels of conservation. Conservation above background was measured by the signal-to-background ratio, defined by $SB = \frac{Fraction\ Conserved\ True\ Set}{Fraction\ Conserved\ Background\ Set}$. Confidence at a given cutoff was calculated as $\frac{SB-1}{SB}$. Errors for fractions and numbers of sites above background were estimated by repeating the analysis 50 times with background sets of size equal to the set of miRNA seeds.

A.1.4 3' Binding to miRNAs

miRNA seed sites were grouped based on their potential binding to the 3' end of corresponding miRNAs following (13) (defined by contiguous base-pairing starting at positions 9 –16 within the miRNA, and allowing for shifts of up to 2 nucleotides of such matches within the mRNA). In cases of seed families with multiple miRNAs, the member with greatest potential base-pairing was chosen. Signal to background was defined as the fraction of seed sites with given 3' binding conserved to the 60% confidence threshold divided by the expected fraction conserved to this threshold. Expected conservation was found by repeating the above procedure while swapping miRNA seed sites with the 3 ends of all other miRNAs.

A.1.5 Final Target Prediction

For each of the microRNAs, a set of target predictions was made for both 8mers and 7mers by using p-value cutoffs derived on all microRNAs that gave predictions at 60% confidence. In addition, following a procedure first developed by [34] an additional confidence score was predicted for each individual site, the p_{ct} (probability of conserved targeting). This score reflects the Bayesian posterior probability that a target was preferentially conserved given its conservation level. This score was derived by the following procedure. First, microRNAs were binned into 10 groups according to their conservation levels. Then within each of these bins, the confidence

was assessed within windows of p-value scores. For a given site, the p_{ct} was calculated by averaging the confidence scores derived for the two neighboring p-value windows. For a transcript, an aggregate probability of conserved targeting q_{ct} was calculated according to the following formula: $q_{ct} = 1 - \prod(1 - p_{ct})$, where the product is over all of the sites in the transcript.

A.1.6 GO-term Enrichment

The AmiGO GO-term enrichment tool [49] was used to search for statistically enriched GO terms in sets of genes. Thresholds were: corrected p-value of 0.05, minimum of 2 gene products.

A.1.7 Cell Transfections

S2R+ cells were maintained in Schneiders medium (Invitrogen), supplemented with 10% FBS and 1% pen-strep. Cells were transfected in 12-well plates using the Effectene Transfection Kit (Qiagen) according to the manufacturers instructions. For each ORF, cells were co-transfected with 3 plasmids: (i) ORF-WT fused to a Myc tag (or FLAG tag in tag-swap), (ii) ORF-MUT fused to a FLAG tag (or Myc tag in tag-swap), and (iii) either mCherry (Control) or mCherry-microRNA, all under the Actin promoter. Cells were cultured for 3 days, lysed and analyzed by Western Blot using LI-COR reagents. Imaging and quantification were performed using the LI-COR Aerius Infrared Imaging System.

A.1.8 Microarrays

S2R+ cells were transfected with either mCherry (Control) or mCherry-miR-1, both under the Actin promoter. Cells were cultured for two days before harvest and total

RNA was extracted using TriZol reagent (Invitrogen) and further purified using Qiagen RNeasy column. Recovered RNA was quantified using a Nanodrop ND-1000 spectrophotometer (Nanodrop Technology). RNA integrity was assessed using an Agilent 2100 bioanalyzer. Samples were labeled following Agilents Two-color microarray-based gene expression analysis (Quick Amp labeling) protocol. Gene expression profiles were generated using a customized Agilent 8 x 15k whole Drosophila Genome Microarray, processed in duplicate, and expression levels were extracted using Agilent Feature Extraction software. Log-ratios were averaged over multiple probes and over technical replicates. Only probes with signal above the median were included in the analysis.

A.1.9 Mutagenesis

Mutagenesis was carried out with the QuikChange II Site-Directed Mutagenesis Kit (Stratagene). All miRNA seed sites were disrupted with 2 synonymous point mutations. Below I show the microRNA seed sites in their context and the mutagenesis primers used. All nucleotide sequences displayed begin in-frame.

1. Jaguar (FBgn0011225)

(i) K Box 8-mer site:

WT: TCCTGTGATATT

AA: Ser Cys Asp Ile

Mut: TCCTGCGACATT

Forward primer:

GATGCTATCAACACGTCCTGCGACATTGAGCTGCTGGAGGCCTG

Reverse primer:

CAGGCCTCCAGCAGCTCAATGTCGCAGGACGTGTTGATAGCATC

2. CG11178 (FBgn0030499)

(i) Mir-1 8-mer site:

WT: ACATTCCAG

AA: Thr Phe Gln

Mut: ACTTTTCAG

Forward primer:

CAGCCCAAGAGATCTCAGTTACTTTTCAGAATCATAAGGACGTCGAAG

Reverse primer:

CTTCGACGTCCTTATGATTCTGAAAAGTAACTGAGATCTCTTGGGCTG

3. CG8494 (FBgn0033916)

(i) Mir-1 8-mer site:

WT: ACATTCCAG

AA: Thr Phe Gln

Mut: ACTTTTCAG

Forward primer:

GTCCACACGCGAGGAAACTTTTCAGGATCTCTCGCTGCCC

Reverse primer:

GGGCAGCGAGAGATCCTGAAAAGTTTCCTCGCGTGTGGAC

(ii) Mir-1 7-mer site 1:

WT: CATTCCAAG

AA: His Ser Lys

Mut: CACTCGAAG

Forward primer:

CCCATGGTGGGATTTTGCACTCGAAGGCGGACGTAATCAGC

Reverse primer:

GCTGATTACGTCCGCCTTCGAGTGCAAAATCCCACCATGGG

(iii) Mir-1 7-mer site 2:

WT: CCATTCCAA

AA: Pro Phe Gln

Mut: CCCTTTCAA

Forward primer:

CAGCTTACCCATTTCGATACCCTTTCAAAGCGACAATTTCCAGGTG

Reverse primer:

CACCTGGAAATTGTCGCTTTGAAAGGGTATCGAAATGGGTAAGCTG

4. Smaug (FBgn0016070)

(i) K Box 8-mer site:

WT: CTCTGTGATAAT

AA: Leu Cys Asp Asn

Mut: CTCTGCGACAAT

Forward primer:

GGGTCGATCAATCCACTCTGCGACAATCTTAATGGTATTACCC

Reverse primer:

GGGTAATACCATTAAGATTGTCGCAGAGTGGATTGATCGACCC

5. Arp87C (FBgn0011745)

(i) Mir-1 8-mer site:

WT: CACATTCCA

AA: His Ile Pro

Mut: CATATCCCA

Forward primer:

CCGGCTTTGCTGGTGAGCATATCCCAAAATGCAGGTTTCCC

Reverse primer:

GGGAAACCTGCATTTTGGGATATGCTCACCAGCAAAGCCGG

(ii) Mir-8 8-mer site:

WT: CACAGTATTATG

AA: His Ser Ile Met

Mut: CACAGCATCATG

Forward primer:

GATTCGCCATGCCTCACAGCATCATGCGCGTGGACATCGCC

Reverse primer:

GGCGATGTCCACGCGCATGATGCTGTGAGGCATGGCGAATC

6. Act88F (FBgn0000047)

(i) Mir-8 8-mer site:

WT: CCAGTATTA

AA: Pro Val Leu

Mut: CCCGTTTTA

Forward primer:

GTGGCCCCCGAGGAGCATCCCGTTTTATTGACCGAGGCTCCACTG

Reverse primer:

CAGTGGAGCCTCGGTCAATAAAACGGGATGCTCCTCGGGGGCCAC

A.2 Methods in Mammalian Studies

This section describes a number of both the computational and experimental techniques used in Section III more detail.

A.2.1 Luciferase Assays

HEK293 cells (ATCC) were plated in 24-well plates and transfected 24 hours later using Lipofectamine 2000 (Invitrogen) and Opti-MEM (Sigma) with 50 ng Renilla luciferase control reporter plasmid pIS1 [38], 400 ng firefly luciferase reporter plasmid and 25 nM miRNA duplex (Supplemental Table 9) per well. After 12 hours, transfection media was replaced with DMEM containing 10% fetal bovine serum and penicillin-streptomycin. Cells were harvested 48 hours post transfection. Luciferase activities were measuring using dual-luciferase assays (Promega), as described by the manufacturer. 4-5 biological replicates, each with 3 technical replicates, were performed. Firefly activity was first normalized to Renilla activity to control for transfection efficiency, and then normalized values were analyzed as described in [38].

Briefly, fold-repression was calculated relative to that of the non-cognate miRNA. When applicable, repression of the reporter with wild-type sites was additionally normalized to that of a reporter in which the miRNA sites were mutated. Statistical significance was determined using the Mann-Whitney U Test. Plasmids (deposited at Addgene) were constructed as described below.

A.2.2 Gene Sequences

RefSeq sequences for human ORFs were downloaded from the UCSC genome browser [50], version: hg19, Feb 2009. In cases of multiple transcript variants for a single gene, one variant was chosen at random as a representative. 3'UTR sequences were downloaded from the TargetScan website [55], version 5.1. For analysis of k-mers and generation of target sets, only non-overlapping k-mer instances were considered, and all transcripts with coding region length greater than 10 kilobases were excluded.

A.2.3 Microarray Data

The microarray data examining the response of introducing miR-181a into HeLa cells [5] (GSM302995) were analyzed using Agilent Feature Extraction Software. Log2 fold-change values for genes were obtained by taking the median value of log2 fold-change for all probes against that gene (excluding any probes not flagged by the Feature Extraction Software as Well Above Background). Down-regulation of a group of genes was taken as the mean of log2 fold-changes across the genes, with errors in these means estimated using 100 bootstrap trials.

A.2.4 Phylogenetic Reconstruction

The sequences of all KRAB-A domain instances as well as the UniProt identifier for the protein in which each domain occurred were downloaded from Pfam website [54], version 24: October 2009. UniProt identifiers were mapped to the corresponding genes using conversion files from the HUGO Gene Nomenclature Committee [51]. Of the resulting 311 genes containing KRAB-A domains, all but 4 (ZNF862, ZNF560, ZNF333, ZFP28) contained a single copy of the domain. For those 4 genes, a single instance of the domain was chosen at random to use in the phylogenetic analysis. A multiple alignment was inferred on the amino acid sequences of the 311 KRAB-A domains, based on which a phylogenetic tree of the corresponding genes was reconstructed, using ClustalX software [71] (version 2.0.12) under default parameter settings. Visualization of the tree and overlap with miRNA target sets was done using the Interactive Tree of Life software tool [52].

A.2.5 Randomization of C2H2-domain Sequences

Instances of a C2H2 domain of the form $XCX_2CX_{12}HX_3H$, as well as the six residues N-teriminal of this motif (to capture the linker sequence) were recorded from within all KRAB-containing genes. For each position in this 28 amino-acid motif, and each possible amino acid within this position, empirical codon frequencies were recorded. Motifs with randomized nucleotides were generated by maintaining the amino-acid sequences but randomly sampling with replacement from the empirical codon frequencies for each amino acid at each position in the domain. Genes with randomized sequences were generated by mapping randomized C2H2 domains back to the original gene sequences.

A.2.6 List of DNA Oligonucleotides Used

LinkerInsertF

CTAGCTGATGGGTACCGGCGGTTCTGGAGGG

LinkerInsertR

GTGACCCCTCCAGAACCGCCGGTACCCATCAG

ControlPrimerF

ATAGGCTAGCCACCATGGTCACCGACGCCAAAAACATAAAG

ControlPrimerR

TGTTGAGCTCATTACACGGCGATCTTTCCGCC

Znf573_Fw

GTTGAGCTAGCCACCATGACCTGTTTTCAGGAATTAG

Znf573_Rev

CTTGAGGTACCCACTTTTATGCTCCTATGAATTC

Zfp37_Fw

GTTGAGCTAGCCACCATGTCGGTCTCCAGCGGC

Zfp37_Rev

CTTGAGGTACCCTCATGAGATTTATCTTCTGAATGAG

Znf20_Fw

GTTGAGCTAGCCACCATGATGTTTCAGGATTCAGTGG

Znf20_Rev

CTTGAGGTACCTCTATTAATGGTATGAGTTCTTTCA

Znf791_Fw

GTTGAGCTAGCCACCATGGACTCAGTGGCTTTTGAGG

Znf791_Rev

CTTGAGGTACCTCGATTGTGCATTCTCATATG

RBAK_Fw

GTTGAGCTAGCCACCATGAACACATTGCAGGGGC

RBAK_Rev

CTTGAGGTACCGAGATTTTCCACATCAAGTACATTC

Znf225_Fw

GTTGAGCTAGCCACCATGACCACGTTGAAGGAGG

Znf225_Rev

CTTGAGGTACCTGTGTCATTTAAAAATAATGACAAA

Znf486_Fw

GTTGAGCTAGCCACCATGCCGGGACCCCTTAGAA

Znf486_Rev

CTTGAGGTACCCGTTCTTGGTTTCTGTCCAA

Znf85_Fw

GTTGAGCTAGCCACCATGAGCCTCAGCGCCCAG

Znf85_Rev

CTTGAGGTACCTATTTGTAATTTTTCTCCGGTATGA

Znf573_ins_1nt_after_622

CCCTAATCAGAGGGACTTATACACGGGATGTGATGT

Znf573_ins_1nt_after_622-antisense

ACATCACATCCCGTGTATAAGTCCCTCTGATTAGGG

Zfp37_ins_1nt_after_622

CGGGCGACCACTGGAGACTGGCTGTGT

Zfp37_ins_1nt_after_622-antisense

ACACAGCCAGTCTCCAGTGGTCGCCCG

Znf20_ins_1nt_after_622

GAAGAATCTCTACAGGGCATGTGATGCAGGAAACC

Znf20_ins_1nt_after_622-antisense

GGTTTCCTGCATCACATGCCCTGTAGAGATTCTTC

Znf791_ins_1nt_after_622

CTCTACAGAGATGTGATGCCAGGAAACATTCAAGAACCT

Znf791_ins_1nt_after_622-antisense

AGGTTCTTGAATGTTTCCTGGCATCACATCTCTGTAGAG

RBAK_ins_1nt_after_622

GACCCTGATGAGAAGATAACTTACACGGGATGTGATGTT

RBAK_ins_1nt_after_622-antisense

AACATCACATCCCGTGTAAGTTATCTTCTCATCAGGGTC

Znf85_ins_1nt_after_754

GTAATGGACTTAACCAATCGTCTCACAGCTACCCAG

Znf85_ins_1nt_after_754-antisense

CTGGGTAGCTGTGAGACGATTGGTTAAGTCCATTAC

Znf20_a963g_t966c

AGAGAATCCATATAGAAATAAGGAGTGCAAGAAAGCCTTCAGTTATCTTGAC

Znf20_a1047g_t1050c

CTAAAGAGAAACCCTATGATGGTAAAGAGTGCACAGAAACCTTCATTTCC

HDAC5_Forward

GTTGAGCTAGCCACCATGTCCCAGCAACACACACTG

HDAC5_Reverse

CTTGAGGTACCTTTATGTTTGGGTGGCCACTGC

IVL_Forward

GTTGAGCTAGCCACCATGAACTCTCCCAACGAGTCG

IVL_Reverse

CTTGAGGTACCCAGGGCAGGCTCCTGCTC

RBAK 3'UTR_Fw

CTTGAACCGGTAGTCAGATCTCAATTTTTAGAAAACTCTCTGAA

RBAK 3'UTR_Rev

GTTGGCGGCCGCTCCAGCAAGAAATGGAGCGAG

RBAK UTR site 1

CACAGAGAAGAATCCCGAAGTTTGTAACAAGAAGCAAAGCCT

RBAK UTR site 2

CAAAGGCAAAATCTGTCAATATGGTGTTTGTGGAAAATATATTGTCTTGGAAAT

Znf20_Into3UTR_Forward

GTTGATCTAGAATGATGTTTCAGGATTCAGTGG

Znf20_Into3UTR_Reverse

CTTGAGCGGCCGCTCTATTAATGGTATGAGTTCTTTCA

RB1_ORF_For

GTTGAACTAGTCACCATGCCGCCCAAAACCCC

RB1_ORF_Rev

CTTGAGGTACCTTTCTCTTCCTTGTTTGAGGTATCCAT

RB1_UTR_For

CTTGAACCGGTGGATCTCAGGACCTTGGTGGA

RB1_UTR_Rev

GTTGGCGGCCGCTTGTAGAAAATAGTAACATAGCAATTTTAAATGTACAGTT

RB1_a1368g_t1371c

AATTATTGAAGTTCTCTGTAAAGAACATGAGTGCAATATAGATGAGGTGAAAAATGTTTATTTC

RB1_g1620c_t1623c

GTAACCTTGATGAAGAGGTCAACGTAATTCCTCCACACACTC

RB1_g2076c_t2079c

AGATTTGTCTTTCCCATGGATTCTCAACGTGCTTAATTTAAAAGCCTTTGAT

RB1_a5691g_t5694c

CTACTGAAACAGATTTCATACCTCAGAGTGCAAAAGAACTTACTGATTATTTTCTTCA

RB1_a5691g_t5694c_antisense

TGAAGAAAATAATCAGTAAGTTCTTTTGCACTCTGAGGTATGAAATCTGTTTCAGTAG

RBAK_a4543g_t4546c

GGACAAAACCTGATGAGTGTAATGAGTGCGGGAAAACATATCATGGAG

RBAK_a4837g_t4840c

GAAATGAAGCCCTTTGAATGCAGTGAGTGCGGAAAATCCTTCTGTAAA

RBAK_a5005g_t5008c

TAGAGGAGAAGCCCTATAAATGTAATGAGTGCGGGAAAACCTTTTGTC

RBAK_a4912g_t4915c

CACAGGAGAGAAACCTTATGAGTGCAATGTATGTGGGAAATCCTTC

RBAK_a5080g_t5083c

CATTCAGGAGAGAAACCCTACGAGTGCAGCGAATGTGGG

A.2.7 List of RNA Oligonucleotides Used

miR-23a

AUCACAUUGCCAGGGAUUUCC

miR-23a Passenger Strand

AAAUCCCUGGGGAUGGGAUUU

miR-124

UAAGGCACGCGGUGAAUGCCA

miR-124 Passenger Strand

GCAUUCACCGCGUGCCUUAAU

miR-181a

AACAUUCAACGCUGUCGGUGAGU

miR-181a Passenger Strand

UCACCGACAGCGUUGAAUGAUAU

miR-199a

CCCAGUGUUCAGACUACCUGUUC

miR-199a Passenger Strand

ACAGGUAGUCUGAACACUGGGUU

miR-370

GCCUGCUGGGGUGGAACCUGGU

miR-370 Passenger Strand

CAGGUUCCACCCCAGCAGGCUU

A.2.8 List of Plasmids Used

pIS1

Control Renilla luciferase construct

pIS7L

Empty vector firefly luciferase construct

pIS7L-Znf573

ZNF573-luciferase fusion construct

pIS7L-Zfp37

ZFP37-luciferase fusion construct

pIS7L-Znf20

ZNF20-luciferase fusion construct

pIS7L-Znf791

ZNF791-luciferase fusion construct

pIS7L-RBAK

RBAK ORF-luciferase fusion construct

pIDTSMART-mZnf20

Minigene construct with half of Znf20 mutated

pIS7L-mZnf20

ZNF20-luciferase fusion construct with miR-181 sites mutated

pIS7L-Znf20-in-UTR

Luciferase followed by ZNF20 in 3'UTR

pIS7L-mZnf20-UTR

Luciferae followed by ZNF20 in 3'UTR with miR-181 sites mutated

pIS7L-fs Znf573

ZNF573-luciferase fusion construct with frame-shift

pIS7L-fs Zfp37

ZFP37-luciferase fusion construct with frame-shift

pIS7L-fs Znf20

ZNF20-luciferase fusion construct with frame-shift

pIS7L-fs Znf791

ZNF791-luciferase fusion construct with frame-shift

pIS7L-fs RBAK

RBAK ORF-luciferase fusion construct with frame-shift

pIS7L-Znf225

ZNF225-luciferase fusion construct

pIS7L-Znf486

ZNF486-luciferase fusion construct

pIS7L-Znf85

ZNF85-luciferase fusion construct

pIS7L-fs Znf85

ZNF85-luciferase fusion construct with frame-shift

pIS7L-HDAC5

HDAC5-luciferase fusion construct

pIS7L-IVL

IVL-luciferase fusion construct

pIDTSMART-mRBAK

Minigene construct with half of RBAK mutated

pIS7L-mRBAK

RBAK ORF-luciferase fusion construct with miR-181 sites mutated

pIS7L-RBAK UTR

Luciferase followed by RBAK 3'UTR construct

pIS7L-RBAK mUTR

Luciferase followed by RBAK 3'UTR construct with miR-181 sites mutated

pIS7L-RBAK O+U

RBAK-luciferase fusion followed by RBAK 3'UTR construct

pIS7L-RBAK mO+U

RBAK-luciferase fusion followed by RBAK 3'UTR construct with miR-181 sites mutated

pIS7L-Rb1

RB1-luciferase fusion construct

pIS7L-mRb1

RB1-luciferase fusion construct with miR-181 sites mutated

pIS7L-Rb1 UTR

Luciferase followed by RB1 3'UTR construct

pIS7L-Rb1 mUTR

Luciferase followed by RB1 3'UTR construct with miR-181 sites mutated

A.2.9 Plasmid Construction

pIS7L In order to generate the control empty vector luciferase plasmid, we first amplified the firefly luciferase from pSP-luc+NF fusion vector (Promega) with ControlPrimerF and ControlPrimerR. The resulting product and pIS1 (Grimson et al. 2007) was with NheI and SacI and ligated to create pIS7. This construct was then digested with NheI and BstEII and ligated with an annealed duplex of LinkerInsertF and LinkerInsertR to generate pIS7L.

C-Terminal Luciferase Fusions. In order to generate C-terminal luciferase fusions, the following ORFs (ZNF573, ZFP37, ZNF20, ZNF791, RBAK, ZNF225, ZNF486, ZNF85, HDAC5 and IVL) were amplified from cDNA clones (BC042170, BC126390, BC036714, BC106938, BC136676, BC108912, BC117268, BC051824 and BC046391, respectively) with the following oligounucleotides: Znf573_Fw, Znf573_Rev; Zfp37_Fw, Zfp37_Rev; Znf20_Fw, Znf20_Rev; Znf791_Fw, Znf791_Rev; RBAK_Fw, RBAK_Rev; Znf225_Fw, Znf225_Rev; Znf486_Fw, Znf486_Rev; Znf85_Fw, Znf85_Rev; HDAC5_Forward, HDAC5_Reverse; IVL_Forward, IVL_Reverse. The resulting products and pIS7L were digested with Kpn1 and NheI in order to generate the luciferase fusion constructs. To generate pIS7L-RB1, the RB1 ORF was amplified from the cDNA clone BC039060 genomic DNA with RB1_ORF_Fw and RB1_ORF_Rev. The resulting product was digested with KpnI and SpI and ligated into pIS7L, which had been digested with KpnI and NheI.

Frame-shift mutant luciferase fusions. In order to generate frame-shifts (1 nucleotide insertion about 100 bp downstream of the start codon), QuikChange II (Stratagene) was used, following the manufacturers instructions and the following oligonucleotides: ZNF573: Znf573_ins_1nt_after_622; Znf573_ins_1nt_after_622-

antisense ZFP37: Zfp37_ins_1nt_after_622; Zfp37_ins_1nt_after_622-antisense ZNF20: Znf20_ins_1nt_after_622; Znf20_ins_1nt_after_622-antisense ZNF791: Znf791_ins_1nt_after_622; Znf791_ins_1nt_after_622-antisense RBAK: RBAK_ins_1nt_after_622; RBAK_ins_1nt_after_622-antisense ZNF85: Znf85_ins_1nt_after_622; Znf85_ins_1nt_after_622-antisense

ZNF20 mutant ORF luciferase fusion. Using QuikChange-Multi kit (Stratagene) and two oligonucleotides (Znf20_a963g_t966c, Znf20_a1047g_t1050c), two 7mers were mutated to generate pIS7L-Znf20int. The ZNF20 minigene (see below) was excised from pIDTSMART-mZnf20 by StuI digested. This was ligated into pIS7L-Znf20int, which had also been digested by StuI, to generate pIS7L-mZnf20.

ZNF20 into 3'UTR luciferase construct. Wild-type or mutant ZNF20 was amplified with Znf20_Into3UTR_Forward and Znf20_Into3UTR_Reverse. The resulting product as well as pIS7L was digested with XbaI an NotI to generate pIS7L-Znf20-in-UTR and pIS7L-mZnf20-in-UTR, respectively.

RBAK mutant ORF luciferase fusion. The N-terminal section of RBAK was excised from pIS7L-RBAK using HindIII. This fragment was ligated into digested pIS0 (Grimson et al. 2007) to generate pRBAKPreMutate. To generate pRBAKMutate1, pRBAKPreMutate was then mutated using QuikChange-Multi (Stratagene) and the following oligonucleotides: RBAK_a4543g_t4546c, RBAK_a4837g_t4840c and RBAK_a5005g_t5008c. To generate pRBAKMutate2, pRBAKMutate1 was mutated using QuikChange Multi and the following oligonucleotides: RBAK_a4912g_t4915c and RBAK_a5080g_t5083c. In order to generate pRBAKInterim, the RBAK minigene (see below) was excised using HindIII and KpnI and ligated into pIS7L-RBAK, also digested with HindIII and KpnI. Finally, to generate pIS7L-mRBAK, the mu-

tated N-terminal half of RBAK was excised from pRBAKMutate2 with HindIII and ligated into pRBAKInterim, which had also been digested with HindIII.

RB1 mutant ORF luciferase fusion. pIS7L-Rb1 was mutated using QuikChange Multi kit (Stratagene), according to manufacturers instructions, and the following oligonucleotides: RB1_a1368g_t1371c, RB1_g1620c_t1623c and RB1_g2076c_t2079c.

3'UTR Luciferase Constructs. The RBAK 3'UTR was amplified from the cDNA clone BC136676 using the oligonucleotides RBAK 3'UTR_Fw/Rev. The resulting product was digested with AgeI and NotI and ligated into digested pIS7L or pIS7L-RBAK in order to generate pIS7L-RBAK UTR or pIS7L-RBAK O+U, respectively. In order to generate pIS7L RBAK mUTR, QuikChange Multi kit (Stratagene) was used to mutate the original, wild-type plasmid with the oligonucleotides RBAK UTR site 1 and RBAK UTR site 2. The resulting mutant UTR was excised with AgeI and NotI and ligated into digest pIS7L-mRBAK in order to generate pIS7L-RBAK mO+U. The RB1 3' UTR was amplified from the cDNA clone BC039060 using the oligonucleotides RB1_UTR_For and RB1_UTR_Rev. The resulting product was digested with AgeI and NotI. In order to mutate the single miR-181 site, QuikChangeII mutagenesis kit was used and the oligonucleotides RB1_a5691g_t5694c and RB1_a5691g_t5694c_antisense.

Minigenes

ZNF20. Generated in pIDTSMART (IDT Technologies) and flanked by StuI restriction sites. Mutated sites are in lower-case.

TTACTCGTTCCACTACCCTTCCAGTACATGAAAGAACTCACACAGGAGTGAA
TGCCGAtgagtgcaAAGAATGTGGGAATGCATTCAGTTTTCCTAGTGAAATTCGT

AGACATAAAAGGTCTCACACTGGAGAAAAACCCTATGAGTGTAAGCAATGTG
GGAAAGTCTTCATTTCTTTCAGTTCCATTCAGTATCATAAGATGACTCACACT
GGAGAGAAACCCTAtgagtgcaAGCAGTGTGGGAAAGCCTTTAGATGTGGCTC
ACACCTTCAAAAGCATGGAAGGACTCACACTGGAGAGAAACCCTAtgagtgca
GGCAATGTGGTAAAGCCTTCAGATGTACCTCGGACCTTCAAAGGCATGAAA
AGACACACACTGAGGATAAACCCTATGGATGTAAGCAGTGTGGGAAAGGCT
TTAGATGTGCTTCACAACTTCAAATTCATGAAAGGACGCACAGTGGAGAGAA
ACCCCAtgagtgcaAGGAATGTGGAAAAGTATTCAAGTATTTTTCTTCCTTGCGT
ATACATGAAAGGACGCACACTGGAGAGAAGCCCCAtgagtgcaAGCAATGTGG
AAAAGCATTCAGGTATTTCTCTTCCTTGCATATACATGAAAGGACACACACTG
GAGATAAGCCATATGAGTGTAAGGTATGTGGCAAAGCCTTCACTTGTTCCAG
TTCCATTCGATATCATGAAAGGACTCACACTGGAGAGAAACCCTAtgagtgcaAG
CACTGTGGTA

RBAK. Generated in pIDTSMART (IDT Technologies) and flanked by HindIII (5' side) and KpnI (3' side). Mutated sites are in lower-case.
TATAAATGTAATGAgTGcGGGAAATCCTACTACCGAAAGTCTACTCTGATTACA
CATCAGAGAACACACACAGGAGAGAAGCCCTATCAGTGTAGCGAGTGTGGGA
AATTCTTTTCTCGGGTGTCATACCTCACTATACATTATAGAAGTCATTTAGAAGA
GAAACCCTATGAgTGcAATGAgTGcGGCAAAACCTTCAATTTAAATTCAGCCTTC
ATTAGACATCGGAAAGTACACACAGAAGAGAAATCCCATGAgTGcAGTGAgTG
cGGAAAGTTCTCTCAGTTGTATCTCACCGACCATCATACAGCTCATTTAGAAGA
GAAACCCTATGAgTGcAATGAgTGcGGGAAAACCTTCCTTGTAAATTCAGCCTTC
GATGGGCACCAGCCACTTCCAAAAGGGGAGAAATCCTATGAgTGcAATGTATG
TGGAAAGTTATTCAATGAGTTGTCATACTATACTGAACATTATAGAAGTCATTCA
GAAGAGAAACCTTATGGATGTAGCGAATGTGGGAAAACCTTTTCCCATAATTCA

TCCCTCTTCAGACATCAAAGAGTACACACAGGCGAGAAACCCTATGAgTGcTAC
GAATGTGGAAAATTCTTCTCTCAGAAATCATATCTCACTATACATCATCGAATTC
ATTCAGGAGAGAAACCCTATGAgTGcAGTAAATGTGGAAAAGTCTTCTCTCGGAT
GTCAAACCTCACTGTCCACTACAGAAGCCATTCAGGAGAGAAACCCTATGAgT
GcAATGAgTGcGGGAAAGTCTTTTCTCAGAAGTCATACCTCACTGTACACTATAG
AACTCATTCAGGAGAGAAACCCTATGAgTGcAATGAGTGTGGGAAAAAATTCCA
CCACAGATCAGCCTTCAATAGCCATCAGAGAATTCATAGAAGAGGAAATATGAA
cGTcCTTGATGTGGAAAATCTC

Appendix B

Repeat Rich Target Gene Lists

Table B.1: Genes containing at least four miR-23 8mers

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF208	NM_007153	chr19	-	22148898	3843	32	3	0
ZNF91	NM_003430	chr19	-	23540498	3576	17	5	0
ZNF107	NM_001013746	chr7	+	64126510	2352	15	4	1
ZNF43	NM_003423	chr19	-	21987752	2430	15	1	2
ZNF676	NM_001001411	chr19	-	22361902	1767	13	0	0
ZNF714	NM_182515	chr19	+	21264952	1665	12	4	1
ZNF626	NM_001076675	chr19	-	20802744	1587	12	3	1
ZNF92	NM_007139	chr7	+	64838767	1554	11	2	1
ZNF257	NM_033468	chr19	+	22235265	1692	11	2	0
ZNF737	NM_001159293	chr19	-	20720798	1611	11	1	1
ZNF431	NM_133473	chr19	+	21324839	1731	10	4	2
ZNF90	NM_007138	chr19	+	20188802	1806	10	3	1
ZNF85	NM_003429	chr19	+	21106058	1788	10	3	1
ZNF254	NM_203282	chr19	+	24269975	1980	10	2	0
ZNF708	NM_021269	chr19	-	21473962	1692	10	0	0
ZNF681	NM_138286	chr19	-	23921998	1938	9	3	3
ZNF98	NM_001098626	chr19	-	22573898	1719	9	2	1
ZNF675	NM_138330	chr19	-	23835708	1707	8	5	2
ZNF430	NM_025189	chr19	+	21203425	1713	8	3	3
ZNF492	NM_020855	chr19	+	22817125	1596	8	2	2
ZNF721	NM_133474	chr4	-	433780	2772	8	0	5

Continued. . .

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF680	NM_178558	chr7	-	63980254	1593	7	6	0
ZNF429	NM_001001415	chr19	+	21688436	2025	7	3	1
ZNF93	NM_031218	chr19	+	20011721	1863	7	3	0
ZNF117	NM_015852	chr7	-	64434829	1452	7	3	0
ZNF273	NM_021148	chr7	+	64363619	1710	7	2	0
ZNF682	NM_001077349	chr19	-	20115228	1401	7	1	1
ZNF100	NM_173531	chr19	-	21906843	1629	6	4	1
ZNF678	NM_178549	chr1	+	227751219	1578	6	3	1
ZNF225	NM_013362	chr19	+	44617547	2121	5	6	0
ZNF845	NM_138374	chr19	+	53837001	2913	5	6	0
FBN1	NM_000138	chr15	-	48700503	8616	5	5	6
ZNF28	NM_006969	chr19	-	53300661	1998	5	4	0
ZNF253	NM_021047	chr19	+	19976713	1500	5	3	2
ZNF479	NM_033273	chr7	-	57187327	1575	5	1	2
ZNF486	NM_052852	chr19	+	20278082	1392	5	1	1
ZNF267	NM_003414	chr16	+	31885078	2232	5	1	0
TEX15	NM_031271	chr8	-	30689061	8370	5	0	2
FBN2	NM_001999	chr5	-	127593601	8739	4	5	2
ZNF235	NM_004234	chr19	-	44790501	2217	4	4	0
ZNF141	NM_003441	chr4	+	331595	1425	4	2	2
ZNF716	NM_001159279	chr7	+	57509882	1488	4	1	3
ZNF826	NM_001039884	chr19	-	20574520	534	4	0	0

Table B.2: Genes containing at least four miR-181 8mers

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF780B	NM_001005851	chr19	-	40534167	2502	16	0	0
ZNF658	NM_033160	chr9	-	40771401	3180	14	14	1
ZNF699	NM_198535	chr19	-	9405985	1929	11	3	1
ZNF709	NM_152601	chr19	-	12571998	1926	11	0	0
ZNF14	NM_021030	chr19	-	19821280	1929	10	0	0
ZNF12	NM_006956	chr7	-	6728063	1980	9	9	2
ZNF546	NM_178544	chr19	+	40502942	2511	9	3	3
ZNF573	NM_152360	chr19	-	38229202	1824	9	2	0
ZNF420	NM_144689	chr19	+	37569381	2067	9	1	3
ZNF607	NM_032689	chr19	-	38187264	2091	9	1	2
ZNF700	NM_144566	chr19	+	12035899	2229	9	0	0

Continued...

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF283	NM_181845	chr19	+	44331472	2040	9	0	0
RBAK	NM_021163	chr7	+	5085552	2145	8	9	1
ZNF470	NM_001001668	chr19	+	57078889	2154	8	7	2
ZFP37	NM_003408	chr9	-	115804173	1893	8	7	0
ZNF334	NM_018102	chr20	-	45129708	2043	8	6	1
ZNF180	NM_013256	chr19	-	44979859	2079	8	3	0
ZFP14	NM_020917	chr19	-	36827161	1602	8	1	2
ZNF345	NM_003419	chr19	+	37341266	1467	8	1	1
ZNF790	NM_206894	chr19	-	37308332	1911	8	1	0
ZNF594	NM_032530	chr17	-	5082830	2424	7	10	0
ZNF461	NM_153257	chr19	-	37128283	1692	7	3	0
ZNF383	NM_152604	chr19	+	37717365	1428	7	3	0
ZNF570	NM_144694	chr19	+	37959981	1611	7	1	3
ZNF30	NM_001099437	chr19	+	35417806	1875	7	1	0
ZFP82	NM_133466	chr19	-	36882860	1599	7	0	3
ZFP30	NM_014898	chr19	-	38123388	1560	7	0	1
ZNF844	NM_001136501	chr19	+	12175545	2001	7	0	0
ZNF44	NM_001164276	chr19	-	12382626	1992	7	0	0
ZNF571	NM_016536	chr19	-	38055155	1830	7	0	0
ZNF829	NM_001037232	chr19	-	37379026	1299	6	2	2
ZNF564	NM_144976	chr19	-	12636184	1662	6	2	0
ZNF433	NM_001080411	chr19	-	12125531	2022	6	1	0
ZNF20	NM_021143	chr19	-	12242802	1599	6	0	2
ZNF778	NM_182531	chr16	+	89284110	2190	6	0	1
ZNF33B	NM_006955	chr10	-	43084554	2337	5	13	0
ZNF569	NM_152484	chr19	-	37902061	2061	5	8	0
ZNF471	NM_020813	chr19	+	57019211	1881	5	6	2
ZNF568	NM_198539	chr19	+	37407233	1935	5	6	1
ZNF527	NM_032453	chr19	+	37862058	1830	5	5	2
ZNF260	NM_001012756	chr19	-	37001589	1239	5	5	0
ZNF583	NM_001159860	chr19	+	56915382	1710	5	3	6
ZFP90	NM_133458	chr16	+	68573660	1911	5	2	0
ZNF443	NM_005815	chr19	-	12540520	2016	5	2	0
ZNF25	NM_145011	chr10	-	38238794	1371	5	1	2
ZNF558	NM_144693	chr19	-	8920381	1209	5	1	0
ZNF560	NM_152476	chr19	-	9577030	2373	5	1	0
ZNF823	NM_001080493	chr19	-	11832079	1833	5	1	0

Continued...

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF442	NM_030824	chr19	-	12460184	1884	5	1	0
ZNF566	NM_032838	chr19	-	36936021	1257	5	0	2
ZNF791	NM_153358	chr19	+	12721731	1731	5	0	1
ZNF491	NM_152356	chr19	+	11909390	1314	5	0	0
ZNF878	NM_001080404	chr19	-	12154619	1737	5	0	0
ZNF33A	NM_006954	chr10	+	38299577	2436	4	14	0
FBN2	NM_001999	chr5	-	127593601	8739	4	14	0
ZNF157	NM_003446	chrX	+	47229998	1521	4	11	2
ZNF182	NM_001007088	chrX	-	47834250	1863	4	8	0
ZNF397OS	NM_001112734	chr18	-	32831022	1485	4	5	0
ZFP3	NM_153018	chr17	+	4981753	1509	4	4	0
ZFP28	NM_020828	chr19	+	57050316	2607	4	3	1
ZNF439	NM_152262	chr19	+	11976843	1500	4	2	0
ZKSCAN1	NM_003439	chr7	+	99613218	1692	4	2	0
ZNF540	NM_152606	chr19	+	38042307	1983	4	1	2
ZNF763	NM_001012753	chr19	+	12075868	1194	4	1	1
ZNF799	NM_001080821	chr19	-	12500828	1932	4	1	0
ZNF620	NM_175888	chr3	+	40547529	1269	4	1	0
ZNF490	NM_020714	chr19	-	12686919	1590	4	0	1
ZNF121	NM_001008727	chr19	-	9676291	1173	4	0	1
ZNF670	NM_033213	chr1	-	247199699	1170	4	0	0
ZNF440	NM_152357	chr19	+	11925106	1788	4	0	0
ZNF529	NM_001145649	chr19	-	37034517	1692	4	0	0
ZNF846	NM_001077624	chr19	-	9868150	1602	4	0	0
ZNF625	NM_145233	chr19	-	12255710	921	4	0	0
ZNF582	NM_144690	chr19	-	56894647	1554	4	0	0
ZNF621	NM_001098414	chr3	+	40566375	1320	4	0	0

Table B.3: Genes containing at least four miR-188 8mers

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF814	NM_001144989	chr19	-	58380746	2568	16	0	1
ZNF594	NM_032530	chr17	-	5082830	2424	15	1	1
ZFP62	NM_152283	chr5	-	180274610	2523	15	0	0
ZNF780B	NM_001005851	chr19	-	40534167	2502	14	1	1
ZNF546	NM_178544	chr19	+	40502942	2511	13	0	0
ZNF658	NM_033160	chr9	-	40771401	3180	12	3	0

Continued...

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF17	NM_006959	chr19	+	57922528	1989	12	1	0
ZNF699	NM_198535	chr19	-	9405985	1929	12	0	1
ZNF283	NM_181845	chr19	+	44331472	2040	12	0	1
ZNF585A	NM_199126	chr19	-	37641000	2145	12	0	0
ZNF225	NM_013362	chr19	+	44617547	2121	12	0	0
ZNF623	NM_001082480	chr8	+	144718372	1491	12	0	0
ZNF585B	NM_152279	chr19	-	37672481	2310	11	2	0
RBAK	NM_021163	chr7	+	5085552	2145	11	1	0
ZNF268	NM_001165881	chr12	+	133758059	2844	11	0	1
ZNF799	NM_001080821	chr19	-	12500828	1932	11	0	1
ZNF443	NM_005815	chr19	-	12540520	2016	11	0	1
ZNF749	NM_001023561	chr19	+	57946692	2337	11	0	1
ZNF555	NM_152791	chr19	+	2841432	1887	11	0	0
ZNF418	NM_133460	chr19	-	58433251	2031	11	0	0
ZNF16	NM_001029976	chr8	-	146155745	2049	11	0	0
ZNF573	NM_152360	chr19	-	38229202	1824	10	1	1
ZNF441	NM_152355	chr19	+	11877814	2082	10	1	0
ZNF33B	NM_006955	chr10	-	43084554	2337	10	0	1
ZNF33A	NM_006954	chr10	+	38299577	2436	10	0	1
ZNF607	NM_032689	chr19	-	38187264	2091	10	0	1
ZNF549	NM_153263	chr19	+	58038692	1884	10	0	1
ZNF778	NM_182531	chr16	+	89284110	2190	10	0	0
ZNF700	NM_144566	chr19	+	12035899	2229	10	0	0
ZNF44	NM_001164276	chr19	-	12382626	1992	10	0	0
ZNF224	NM_013398	chr19	+	44598481	2124	10	0	0
ZNF12	NM_006956	chr7	-	6728063	1980	10	0	0
ZNF84	NM_003428	chrUn_gl000223	-	42779	2217	9	1	0
ZNF433	NM_001080411	chr19	-	12125531	2022	9	1	0
ZNF442	NM_030824	chr19	-	12460184	1884	9	1	0
ZNF540	NM_152606	chr19	+	38042307	1983	9	1	0
ZNF823	NM_001080493	chr19	-	11832079	1833	9	0	1
ZNF792	NM_175872	chr19	-	35447257	1899	9	0	1
ZNF345	NM_003419	chr19	+	37341266	1467	9	0	0
ZNF491	NM_152356	chr19	+	11909390	1314	9	0	0
ZNF331	NM_018555	chr19	+	54024176	1392	9	0	0
ZNF470	NM_001001668	chr19	+	57078889	2154	9	0	0
ZNF416	NM_017879	chr19	-	58082934	1785	9	0	0

Continued. . .

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF530	NM_020880	chr19	+	58111252	1800	9	0	0
ZNF154	NM_001085384	chr19	-	58211810	1314	9	0	0
ZNF484	NM_031486	chr9	-	95608350	2559	9	0	0
ZNF157	NM_003446	chrX	+	47229998	1521	9	0	0
ZNF420	NM_144689	chr19	+	37569381	2067	8	1	1
ZNF229	NM_014518	chr19	-	44930425	2478	8	1	1
ZNF564	NM_144976	chr19	-	12636184	1662	8	1	0
ZNF45	NM_003425	chr19	-	44416776	2049	8	1	0
ZNF547	NM_173631	chr19	+	57874890	1209	8	1	0
ZNF304	NM_020657	chr19	+	57862644	1980	8	0	1
ZNF239	NM_005674	chr10	-	44051794	1377	8	0	0
ZNF26	NM_019591	chrUn_gl000223	-	93510	1602	8	0	0
ZNF287	NM_020653	chr17	-	16453630	2286	8	0	0
ZNF440	NM_152357	chr19	+	11925106	1788	8	0	0
ZNF256	NM_005773	chr19	-	58452202	1884	8	0	0
ZNF77	NM_021217	chr19	-	2933216	1638	8	0	0
ZNF560	NM_152476	chr19	-	9577030	2373	8	0	0
ZNF136	NM_003437	chr19	+	12273871	1623	8	0	0
ZNF30	NM_001099437	chr19	+	35417806	1875	8	0	0
ZNF284	NM_001037813	chr19	+	44576296	1782	8	0	0
ZNF211	NM_006385	chr19	+	58144534	1734	8	0	0
ZNF587	NM_032828	chr19	+	58361268	1728	8	0	0
ZNF35	NM_003420	chr3	+	44690232	1584	8	0	0
ZNF445	NM_181489	chr3	-	44481261	3096	8	0	0
ZNF660	NM_173658	chr3	+	44626455	996	8	0	0
ZNF197	NM_006991	chr3	+	44666510	3090	8	0	0
ZNF596	NM_173539	chr8	+	182199	1515	8	0	0
ZNF132	NM_003433	chr19	-	58944181	2121	7	2	1
ZNF624	NM_020787	chr17	-	16524047	2598	7	1	0
ZNF772	NM_001024596	chr19	-	57980954	1470	7	1	0
ZNF134	NM_003435	chr19	+	58125829	1284	7	1	0
ZNF81	NM_007137	chrX	+	47696300	1986	7	1	0
ZNF556	NM_024967	chr19	+	2867332	1371	7	0	1
ZNF563	NM_145276	chr19	-	12428305	1431	7	0	1
ZNF544	NM_014480	chr19	+	58740069	2148	7	0	1
ZNF311	NM_001010877	chr6_ssto_hap7	-	303579	2001	7	0	1
ZNF37A	NM_001007094	chr10	+	38383263	1686	7	0	0

Continued...

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF25	NM_145011	chr10	-	38238794	1371	7	0	0
ZNF426	NM_024106	chr19	-	9638682	1665	7	0	0
ZNF846	NM_001077624	chr19	-	9868150	1602	7	0	0
ZNF627	NM_145295	chr19	+	11708234	1386	7	0	0
ZNF439	NM_152262	chr19	+	11976843	1500	7	0	0
ZNF878	NM_001080404	chr19	-	12154619	1737	7	0	0
ZNF844	NM_001136501	chr19	+	12175545	2001	7	0	0
ZNF791	NM_153358	chr19	+	12721731	1731	7	0	0
ZNF181	NM_001029997	chr19	+	35225479	1716	7	0	0
ZNF382	NM_032825	chr19	+	37096220	1653	7	0	0
ZNF461	NM_153257	chr19	-	37128283	1692	7	0	0
ZNF567	NM_152603	chr19	+	37180301	1851	7	0	0
ZNF790	NM_206894	chr19	-	37308332	1911	7	0	0
ZNF180	NM_013256	chr19	-	44979859	2079	7	0	0
ZNF175	NM_007147	chr19	+	52074530	2136	7	0	0
ZNF548	NM_152909	chr19	+	57901217	1602	7	0	0
ZIK1	NM_001010879	chr19	+	58095627	1464	7	0	0
ZNF671	NM_024833	chr19	-	58231119	1605	7	0	0
ZNF250	NM_001109689	chr8	-	146102337	1668	7	0	0
ZNF79	NM_007135	chr9	+	130186652	1497	7	0	0
ZFP90	NM_133458	chr16	+	68573660	1911	6	3	0
ZNF485	NM_145312	chr10	+	44101854	1326	6	2	0
ZNF57	NM_173480	chr19	+	2900895	1668	6	2	0
ZNF667	NM_022103	chr19	-	56950693	1833	6	1	1
ZNF121	NM_001008727	chr19	-	9676291	1173	6	1	0
ZNF571	NM_016536	chr19	-	38055155	1830	6	1	0
ZNF334	NM_018102	chr20	-	45129708	2043	6	1	0
ZNF879	NM_001136116	chr5	+	178450775	1692	6	1	0
ZNF192	NM_006298	chr6	+	28109715	1737	6	1	0
ZFP37	NM_003408	chr9	-	115804173	1893	6	1	0
ZNF527	NM_032453	chr19	+	37862058	1830	6	0	1
ZNF502	NM_001134440	chr3	+	44754134	1635	6	0	1
ZNF251	NM_138367	chr8	-	145946294	2016	6	0	1
ZNF674	NM_001039891	chrX	-	46357160	1746	6	0	1
ZNF559	NM_032497	chr19	+	9434927	1617	6	0	0
ZNF565	NM_152477	chr19	-	36673187	1500	6	0	0
ZNF551	NM_138347	chr19	+	58193356	1965	6	0	0

Continued...

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF561	NM_152289	chr19	-	9718002	1461	6	0	0
ZNF568	NM_198539	chr19	+	37407233	1935	6	0	0
ZNF235	NM_004234	chr19	-	44790501	2217	6	0	0
ZNF613	NM_024840	chr19	+	52430687	1746	6	0	0
ZNF615	NM_198480	chr19	-	52494587	2196	6	0	0
ZNF264	NM_003417	chr19	+	57702867	1884	6	0	0
ZNF805	NM_001023563	chr19	+	57752052	1884	6	0	0
ZNF543	NM_213598	chr19	+	57831864	1803	6	0	0
ZNF417	NM_152475	chr19	-	58417142	1728	6	0	0
ZNF584	NM_173548	chr19	+	58920062	1266	6	0	0
ZNF662	NM_207404	chr3	+	42947401	1281	6	0	0
ZKSCAN5	NM_145102	chr7	+	99102272	2520	6	0	0
ZNF34	NM_030580	chr8	-	145998502	1683	6	0	0
ZNF510	NM_014930	chr9	-	99518146	2052	6	0	0
ZNF189	NM_003452	chr9	+	104161162	1881	6	0	0
ZNF41	NM_007130	chrX	-	47305561	2340	6	0	0
ZNF223	NM_013361	chr19	+	44556163	1449	5	2	0
ZNF419	NM_001098491	chr19	+	57999078	1536	5	2	0
ZNF630	NM_001037735	chrX	-	47917568	1974	5	2	0
ZNF155	NM_003445	chr19	+	44488354	1617	5	1	1
ZNF605	NM_183238	chr12	-	133498018	1926	5	1	0
ZNF23	NM_145911	chr16	-	71481512	1932	5	1	0
ZFP3	NM_153018	chr17	+	4981753	1509	5	1	0
ZNF221	NM_013359	chr19	+	44455396	1854	5	1	0
ZNF230	NM_006300	chr19	+	44507076	1425	5	1	0
ZNF649	NM_023074	chr19	-	52392488	1518	5	1	0
ZNF586	NM_017652	chr19	+	58281024	1209	5	1	0
ZNF8	NM_021089	chr19	+	58790317	1728	5	1	0
ZNF167	NM_018651	chr3	+	44596712	2265	5	1	0
ZNF184	NM_007149	chr6	-	27418521	2256	5	1	0
ZNF473	NM_001006656	chr19	+	50529211	2616	5	0	2
ZNF460	NM_006635	chr19	+	57791852	1689	5	0	1
ZNF14	NM_021030	chr19	-	19821280	1929	5	0	1
ZSCAN20	NM_145238	chr1	+	33938231	3132	5	0	0
ZNF436	NM_030634	chr1	-	23685941	1413	5	0	0
ZNF248	NM_021045	chr10	-	38117898	1740	5	0	0
ZNF286A	NM_001130842	chr17	+	15602890	1566	5	0	0

Continued...

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF286B	NM_001145045	chr17	-	18561741	1569	5	0	0
ZNF709	NM_152601	chr19	-	12571998	1926	5	0	0
ZNF599	NM_001007248	chr19	-	35248978	1767	5	0	0
ZNF529	NM_001145649	chr19	-	37034517	1692	5	0	0
ZNF793	NM_001013659	chr19	+	37997840	1221	5	0	0
ZFP30	NM_014898	chr19	-	38123388	1560	5	0	0
ZNF550	NM_001039654	chr19	-	58058342	1146	5	0	0
ZNF558	NM_144693	chr19	-	8920381	1209	5	0	0
ZNF20	NM_021143	chr19	-	12242802	1599	5	0	0
ZNF260	NM_001012756	chr19	-	37001589	1239	5	0	0
ZNF383	NM_152604	chr19	+	37717365	1428	5	0	0
ZNF222	NM_001129996	chr19	+	44529493	1476	5	0	0
ZNF234	NM_001144824	chr19	+	44645709	2103	5	0	0
ZNF432	NM_014650	chr19	-	52536677	1959	5	0	0
ZNF471	NM_020813	chr19	+	57019211	1881	5	0	0
ZNF776	NM_173632	chr19	+	58258163	1557	5	0	0
ZNF70	NM_021916	chr22	-	24083772	1341	5	0	0
ZNF572	NM_152412	chr8	+	125985538	1590	5	0	0
ZNF7	NM_003416	chr8	+	146052902	2061	5	0	0
ZNF266	NM_006631	chr19	-	9523271	1650	4	3	0
ZNF354A	NM_005649	chr5	-	178138530	1818	4	2	0
ZNF552	NM_024762	chr19	-	58318451	1224	4	1	1
ZNF557	NM_001044387	chr19	+	7069470	1293	4	1	0
ZNF554	NM_001102651	chr19	+	2819871	1617	4	1	0
ZNF773	NM_198542	chr19	+	58011308	1329	4	1	0
ZNF606	NM_025027	chr19	-	58488446	2379	4	1	0
ZNF354C	NM_014594	chr5	+	178487606	1665	4	1	0
ZSCAN22	NM_181846	chr19	+	58838384	1476	4	0	1
ZNF781	NM_152605	chr19	-	38158649	984	4	0	1
ZNF347	NM_032584	chr19	-	53641957	2520	4	0	1
ZNF449	NM_152695	chrX	+	134478695	1557	4	0	1
ZNF124	NM_003431	chr1	-	247319202	870	4	0	0
ZNF684	NM_152373	chr1	+	40997232	1137	4	0	0
ZNF434	NM_017810	chr16	-	3432085	1458	4	0	0
ZNF232	NM_014519	chr17	-	5009032	1335	4	0	0
VEZF1	NM_007146	chr17	-	56048909	1566	4	0	0
ZNF397	NM_001135178	chr18	+	32820993	1605	4	0	0

Continued...

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF333	NM_032433	chr19	+	14800869	1998	4	0	0
ZNF562	NM_001130031	chr19	-	9759350	1281	4	0	0
ZNF625	NM_145233	chr19	-	12255710	921	4	0	0
ZNF302	NM_001012320	chr19	+	35168566	1200	4	0	0
ZNF570	NM_144694	chr19	+	37959981	1611	4	0	0
ZNF227	NM_182490	chr19	+	44716690	2400	4	0	0
ZNF233	NM_181756	chr19	+	44764075	2013	4	0	0
ZFP112	NM_001083335	chr19	-	44830706	2742	4	0	0
ZNF160	NM_001102603	chr19	-	53569867	2457	4	0	0
ZNF582	NM_144690	chr19	-	56894647	1554	4	0	0
ZNF583	NM_001159860	chr19	+	56915382	1710	4	0	0
ZFP28	NM_020828	chr19	+	57050316	2607	4	0	0
ZNF2	NM_001017396	chr2	+	95831182	1155	4	0	0
ZNF619	NM_001145083	chr3	+	40518632	1599	4	0	0
ZNF454	NM_182594	chr5	+	178368193	1569	4	0	0
ZNF300	NM_052860	chr5	-	150273953	1815	4	0	0
ZNF323	NM_145909	chr6	-	28292516	1221	4	0	0
ZNF193	NM_006299	chr6	+	28193072	1185	4	0	0
ZSCAN12	NM_001163391	chr6	-	28356728	1836	4	0	0

Table B.4: Genes containing at least four miR-199 8mers

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF442	NM_030824	chr19	-	12460184	1884	12	1	1
ZNF709	NM_152601	chr19	-	12571998	1926	11	2	0
ZNF433	NM_001080411	chr19	-	12125531	2022	11	2	0
ZNF700	NM_144566	chr19	+	12035899	2229	11	1	0
ZNF844	NM_001136501	chr19	+	12175545	2001	11	0	0
ZNF44	NM_001164276	chr19	-	12382626	1992	10	1	0
ZNF14	NM_021030	chr19	-	19821280	1929	10	1	0
ZNF441	NM_152355	chr19	+	11877814	2082	9	2	0
ZNF551	NM_138347	chr19	+	58193356	1965	9	1	0
ZNF791	NM_153358	chr19	+	12721731	1731	9	1	0
ZNF823	NM_001080493	chr19	-	11832079	1833	9	0	1
ZNF530	NM_020880	chr19	+	58111252	1800	9	0	1
ZNF814	NM_001144989	chr19	-	58380746	2568	8	4	0
ZNF544	NM_014480	chr19	+	58740069	2148	8	2	0

Continued...

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF20	NM_021143	chr19	-	12242802	1599	8	1	0
ZNF773	NM_198542	chr19	+	58011308	1329	8	1	0
ZNF443	NM_005815	chr19	-	12540520	2016	8	0	1
ZNF304	NM_020657	chr19	+	57862644	1980	8	0	1
ZNF419	NM_001098491	chr19	+	57999078	1536	8	0	0
ZNF418	NM_133460	chr19	-	58433251	2031	7	3	0
ZNF136	NM_003437	chr19	+	12273871	1623	7	2	0
ZNF564	NM_144976	chr19	-	12636184	1662	7	2	0
ZNF878	NM_001080404	chr19	-	12154619	1737	7	1	0
ZNF587	NM_032828	chr19	+	58361268	1728	7	1	0
ZNF417	NM_152475	chr19	-	58417142	1728	7	1	0
ZNF440	NM_152357	chr19	+	11925106	1788	7	0	0
ZNF266	NM_006631	chr19	-	9523271	1650	7	0	0
ZNF91	NM_003430	chr19	-	23540498	3576	7	0	0
ZNF491	NM_152356	chr19	+	11909390	1314	6	2	0
ZNF625	NM_145233	chr19	-	12255710	921	6	2	0
ZNF132	NM_003433	chr19	-	58944181	2121	6	1	3
ZNF563	NM_145276	chr19	-	12428305	1431	6	1	1
ZNF555	NM_152791	chr19	+	2841432	1887	6	1	0
ZNF699	NM_198535	chr19	-	9405985	1929	6	1	0
ZNF627	NM_145295	chr19	+	11708234	1386	6	1	0
ZNF235	NM_004234	chr19	-	44790501	2217	6	1	0
ZNF799	NM_001080821	chr19	-	12500828	1932	6	0	1
ZNF776	NM_173632	chr19	+	58258163	1557	6	0	1
ZNF439	NM_152262	chr19	+	11976843	1500	6	0	0
ZNF547	NM_173631	chr19	+	57874890	1209	6	0	0
ZIK1	NM_001010879	chr19	+	58095627	1464	6	0	0
ZNF211	NM_006385	chr19	+	58144534	1734	6	0	0
ZNF596	NM_173539	chr8	+	182199	1515	6	0	0
ZNF251	NM_138367	chr8	-	145946294	2016	5	4	1
ZNF543	NM_213598	chr19	+	57831864	1803	5	3	0
ZNF45	NM_003425	chr19	-	44416776	2049	5	2	1
ZNF264	NM_003417	chr19	+	57702867	1884	5	2	0
ZNF805	NM_001023563	chr19	+	57752052	1884	5	2	0
ZNF57	NM_173480	chr19	+	2900895	1668	5	1	0
ZNF227	NM_182490	chr19	+	44716690	2400	5	1	0
ZNF17	NM_006959	chr19	+	57922528	1989	5	1	0

Continued...

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF772	NM_001024596	chr19	-	57980954	1470	5	0	1
ZNF416	NM_017879	chr19	-	58082934	1785	5	0	1
ZNF77	NM_021217	chr19	-	2933216	1638	5	0	0
ZNF502	NM_001134440	chr3	+	44754134	1635	4	4	0
ZNF550	NM_001039654	chr19	-	58058342	1146	4	2	0
ZNF7	NM_003416	chr8	+	146052902	2061	4	2	0
ZNF16	NM_001029976	chr8	-	146155745	2049	4	2	0
ZNF778	NM_182531	chr16	+	89284110	2190	4	1	1
ZNF599	NM_001007248	chr19	-	35248978	1767	4	1	1
ZNF671	NM_024833	chr19	-	58231119	1605	4	1	1
ZNF10	NM_015394	chr12	+	133707213	1722	4	1	0
ZNF490	NM_020714	chr19	-	12686919	1590	4	1	0
ZNF177	NM_003451	chr19	+	9473695	966	4	1	0
ZNF560	NM_152476	chr19	-	9577030	2373	4	1	0
ZNF426	NM_024106	chr19	-	9638682	1665	4	1	0
ZNF749	NM_001023561	chr19	+	57946692	2337	4	1	0
ZNF549	NM_153263	chr19	+	58038692	1884	4	1	0
ZNF80	NM_007136	chr3	-	113953480	822	4	1	0
ZNF83	NM_001105549	chr19	-	53115619	1551	4	0	2
PIKFYVE	NM_015040	chr2	+	209130990	6297	4	0	2
ZSCAN20	NM_145238	chr1	+	33938231	3132	4	0	1
ZNF436	NM_030634	chr1	-	23685941	1413	4	0	0
ZNF397OS	NM_001112734	chr18	-	32831022	1485	4	0	0
ZNF121	NM_001008727	chr19	-	9676291	1173	4	0	0
ZNF846	NM_001077624	chr19	-	9868150	1602	4	0	0
ZNF101	NM_033204	chr19	+	19779662	1311	4	0	0
ZNF585A	NM_199126	chr19	-	37641000	2145	4	0	0
ZNF585B	NM_152279	chr19	-	37672481	2310	4	0	0
ZFP112	NM_001083335	chr19	-	44830706	2742	4	0	0
ZNF548	NM_152909	chr19	+	57901217	1602	4	0	0
ZNF623	NM_001082480	chr8	+	144718372	1491	4	0	0

Table B.5: Genes containing at least four miR-370 8mers

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
IVL	NM_005547	chr1	+	152881038	1758	22	6	7
TCHH	NM_007113	chr1	-	152078792	5832	14	1	18

Continued...

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF853	NM_017560	chr7	+	6655526	1980	11	3	1
MKI67	NM_001145966	chr10	-	129894926	8691	7	8	3
CEP250	NM_007186	chr20	+	34043222	7329	6	4	7
NUMA1	NM_006185	chr11	-	71713910	6348	6	3	2
HDAC5	NM_001015053	chr17	-	42154120	3372	6	1	1
LRRC37A3	NM_199340	chr17	-	62850487	4905	6	1	0
LRRC37A	NM_014834	chr17_ctg5_hap1	-	415626	5103	6	1	0
LRRC37A2	NM_001006607	chr17_ctg5_hap1	-	197705	5103	6	1	0
TRIOBP	NM_001039141	chr22	+	38092994	7098	5	5	4
CSPG4	NM_001897	chr15	-	75966663	6969	5	2	0
SYNE1	NM_015293	chr6	-	152442822	9966	5	1	0
EVPL	NM_001988	chr17	-	74002926	6102	4	3	3
CROCC	NM_014675	chr1	+	17248444	6054	4	2	7
MYH9	NM_002473	chr22	-	36677323	5883	4	2	3
AMBRA1	NM_017749	chr11	-	46417963	3627	4	2	1
GOLGA3	NM_005895	chr12	-	133345494	4497	4	2	1
CCHCR1	NM_001105563	chr6_mann_hap4	-	2458574	2508	4	2	1
GRIPAP1	NM_020137	chrX	-	48830133	2526	4	2	1
BTBD12	NM_032444	chr16	-	3631183	5505	4	1	2
MYO10	NM_012334	chr5	-	16662016	6177	4	1	1
GIGYF1	NM_022574	chr7	-	100277129	3108	4	1	1
NCOR1	NM_006311	chr17	-	15933409	7323	4	0	3
DMBT1	NM_004406	chr10	+	124320180	5358	4	0	1

Table B.6: Genes containing at least four miR-766 8mers

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
ZNF208	NM_007153	chr19	-	22148898	3843	9	0	23
CCDC88C	NM_001080414	chr14	-	91737668	6087	7	14	5
SPTBN2	NM_006946	chr11	-	66452720	7173	7	4	3
MYH9	NM_002473	chr22	-	36677323	5883	6	15	3
SPTBN4	NM_020971	chr19	+	40973125	7695	6	12	2
MYO18B	NM_032608	chr22	+	26138119	7704	6	7	5
MYH14	NM_001077186	chr19	+	50706884	6012	5	19	3
CCDC88B	NM_032251	chr11	+	64107689	4431	5	19	1
CEP250	NM_007186	chr20	+	34043222	7329	5	10	8
MYO9B	NM_001130065	chr19	+	17186590	6069	5	9	2

Continued...

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
LMTK3	NM_001080434	chr19	-	48988529	4470	5	1	2
GOLGA6L10	NM_001164465	chr15	-	83009408	1422	5	1	1
SHROOM3	NM_020859	chr4	+	77356252	5991	5	0	0
MYH7	NM_000257	chr14	-	23881947	5808	4	19	4
MYH6	NM_002471	chr14	-	23851198	5820	4	19	3
EVPL	NM_001988	chr17	-	74002926	6102	4	10	3
CGN	NM_020770	chr1	+	151483861	3612	4	10	1
MYH3	NM_002470	chr17	-	10531842	5823	4	9	4
OBSL1	NM_015311	chr2	-	220415450	5691	4	6	0
BZRAP1	NM_004758	chr17	-	56378595	5574	4	5	2
MAD1L1	NM_001013836	chr7	-	1855427	2157	4	5	1
KIF1A	NM_004321	chr2	-	241653184	5073	4	4	0
PREX1	NM_020820	chr20	-	47240792	4980	4	3	2
WWC1	NM_001161661	chr5	+	167719064	3360	4	2	1
ABCF3	NM_018358	chr3	+	183903862	2130	4	2	0
GOLGA6L9	NM_198181	chr15	+	82722184	1299	4	1	1
COL12A1	NM_004370	chr6	-	75794042	9192	4	0	4

Table B.7: Genes containing at least four miR-1248 8mers

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
CYLC2	NM_001340	chr9	+	105757592	1047	7	2	1
BOD1L	NM_148894	chr4	-	13570365	9156	6	7	2
NEFH	NM_021076	chr22	+	29876180	3063	6	3	8
CYLC1	NM_021118	chrX	+	83116169	1956	6	2	0
MAP1B	NM_005909	chr5	+	71403117	7407	6	1	2
KNTC1	NM_014708	chr12	+	123011808	6630	6	0	3
BRD3	NM_007371	chr9	-	136895453	2181	5	4	3
MYH8	NM_002472	chr17	-	10293641	5814	5	3	3
MAP1A	NM_002373	chr15	+	43809805	8412	5	2	11
MYH6	NM_002471	chr14	-	23851198	5820	5	2	7
RANBP2	NM_006267	chr2	+	109335936	9675	5	2	1
ANKRD11	NM_013275	chr16	-	89334034	7992	5	1	11
C1orf173	NM_001002912	chr1	-	75033795	4593	4	4	2
MYH13	NM_003802	chr17	-	10204182	5817	4	3	6
SRRM1	NM_005839	chr1	+	24969593	2715	4	3	2
SH3PXD2A	NM_014631	chr10	-	105353783	3318	4	3	2

Continued...

Symbol	Refseq ID	Chrom	Strand	Location	Length	8mer	7mer(m8)	7mer(1A)
MYH7	NM_000257	chr14	-	23881947	5808	4	2	7
TOP1	NM_003286	chr20	+	39657461	2298	4	2	4
MYO18B	NM_032608	chr22	+	26138119	7704	4	2	4
BDP1	NM_018429	chr5	+	70751441	7875	4	2	1
TRDN	NM_006073	chr6	-	123537482	2190	4	2	1
AIM1	NM_001624	chr6	+	106959729	5172	4	2	1
SMC1A	NM_006306	chrX	-	53401071	3702	4	1	5
CENPF	NM_016343	chr1	+	214776531	9345	4	1	3
FAM9A	NM_174951	chrX	-	8758838	999	4	1	2
SAFB2	NM_014649	chr19	-	5587010	2862	4	1	1
MICAL3	NM_015241	chr22	-	18270417	6009	4	1	1
LOXHD1	NM_144612	chr18	-	44057216	6636	4	0	2
LONP1	NM_004793	chr19	-	5691845	2880	4	0	2
RAB11FIP1	NM_001002814	chr8	-	37716465	3852	4	0	2
JAK1	NM_002227	chr1	-	65298905	3465	4	0	0
PDHX	NM_001135024	chr11	+	34937676	1461	4	0	0
PPP1R12A	NM_001143886	chr12	-	80167343	2832	4	0	0
DNAJC2	NM_001129887	chr7	-	102952921	1707	4	0	0

Bibliography

[1] I Alvarez-Garcia and EA Miska. MicroRNA functions in animal development and human disease. *Development*, 132:4653–4662, 2005.

[2] V Ambros. MicroRNAs: Tiny regulators with great potential. *Cell*, 107:823–826, 2002.

[3] M Andronescu, AP Fejes, F Hutter, HH Hoos, and A Condon. A new algorithm for RNA secondary structure design. *Journal of Molecular Biology*, 336:607–624, 2004.

[4] A Arvey, E Larsson, C Sander, CS Leslie, and DS Marks. Target mRNA abundance dilutes microRNA and siRNA activity. *Molecular Systems Biology*, 6:363, 2010.

[5] D Baek, J Villen, C Shin, FD Camargo, SP Gygi, and DP Bartel. The impact of microRNAs on protein output. *Nature*, 455:64–71, 2008.

[6] DP Bartel. MicroRNAs: Target recognition and regulatory functions. *Cell*, 136:215–233, 2009.

[7] DP Bartel and CZ Chen. Micromanagers of gene expression: The potentially widespread influence of metazoan microRNAs. *Nature Reviews Genetics*, 5:396–400, 2004.

[8] RR Breaker. Are engineered proteins getting competition from RNA? *Current Opinion in Biotechnology*, 7:442–448, 1996.

[9] J Brennecke, A Stark, RB Russell, and SM Cohen. Principles of microRNA-target recognition. *PLoS Biology*, 3:e85, 2005.

[10] P Brest, P Lapaquette, M Souidi, K Lebrigand, A Cesaro, V Vouret-Craviari, B Mari, P Barbry, JF Mosnier, X Hébuterne, A Harel-Bellan, B Mograbi,

A Darfeuille-Michaud, and P Hofman. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nature Genetics*, 43:242–245, 2011.

[11] A Busch and R Backofen. INFO-RNA- a fast approach to inverse RNA folding. *Bioinformatics*, 22:1823–1831, 2006.

[12] N Bushati and SM Cohen. microRNA functions. *Annual Review of Cellular and Developmental Biology*, 23:175–205, 2007.

[13] G Butterfoss and B Kuhlman. Computer-based design of novel protein structures. *Annual Review of Biophysics and Biomolecular Structure*, 35:49–65, 2005.

[14] S Carbon, A Ireland, CJ Mungall, SQ Shu, B Marshall, S Lewis, the AmiGO Hub, and the Web Presence Working Group. AmiGO: Online access to ontology and annotation data. *Bioinformatics*, 25:288–289, 2009.

[15] NK Cervigne, PP Reis, J Machado, B Sadikovic, G Bradley, NN Galloni, M Pintilie, I Jurisica, B Perez-Ordonez, R Gilbert, J Irish, and S Kamel-Reid. Identification of a microRNA signature associated with progression of leukoplakia to oral carcinoma. *Human Molecular Genetics*, 18:4818–4829, 2009.

[16] JV Chamary, JL Parmley, and LD Hurst. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*, 7:98–108, 2006.

[17] B Charlesworth, MT Morgan, and D Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134:1289–1303, 1993.

[18] H Chen and M Blanchette. Detecting non-coding selective pressure in coding regions. *BMC Evolutionary Biology*, 7:S9, 2007.

[19] K Chen and N Rajewsky. Natural selection on human microRNA binding sites inferred from SNP data. *Nature Genetics*, 38:1452–1456, 2006.

[20] SW Chi, JB Zang, A Mele, and RB Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460:479–486, 2009.

[21] D Cifuentes, H Xue, DW Taylor, H Patnode, Y Mishima, S Cheloufi, E Ma, S Mane, GJ Hannon, ND Lawson, SA Wolfe, and AJ Giraldez. A novel miRNA processing pathway independent of dicer requires Argonaute2 catalytic activity. *Science*, 328:1694–1698, 2010.

[22] M Classon and E Harlow. The retinoblastoma tumour suppressor in development and cancer. *Nature Reviews Cancer*, 2:910–917, 2002.

[23] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[24] G Ding, P Lorenz, M Kreutzer, Y Li, and HJ Thiesen. SysZNF: the C2H2 zinc finger gene database. *Nucleic Acids Research*, 37:D267–D273, 2009.

[25] R Dowell and S Eddy. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioninformatics*, 5:71, 2004.

[26] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probablistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[27] AM Duursma, M Kedde, M Schrier, C le Sage, and R Agami. miR-148 targets human DNMT3b protein coding region. *RNA*, 14:872–877, 2008.

[28] G Easow, AA Teleman, and Cohen SM. Isolation of microRNA targets by miRNP immunopurification. *RNA*, 13:1198–1204, 2007.

[29] I Elcheva, S Goswami, FK Noubissi, and VS Spiegelman. CRD-BP protects the coding region of βTrCP1 mRNA from miR-183-mediated degradation. *Molecular Cell*, 35:240–246, 2009.

[30] S Elizalde and K Woods. Bounds on the number of inference functions of a graphical model. *ArXiv Mathematics e-prints*, math/0610233, 2006.

[31] RO Emerson and JH Thomas. Adaptive evolution in zinc finger transcription factors. *PLoS Genetics*, 5:e1000325, 2009.

[32] WG Fairbrother, RF Yeh, PA Sharp, and CB Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297:1007–1013, 2002.

[33] JJ Forman, A Legesse-Miller, and HA Coller. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets dicer within its coding sequence. *PNAS*, 105:14879–14884, 2008.

[34] RC Friedman, KKH Farh, CB Burge, and DP Bartel. Most mammalian mRNAs are most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19:92–105, 2009.

[35] NL Garneau, J Wilusz, and CJ Wilusz. The highways and byways of mRNA decay. *Nature Reviews Molecular Cellular Biology*, 8:113–126, 2007.

[36] J Gillespie. *Population Genetics: A Concise Guide*. The Johns Hopkins University Press, 2004.

[37] BR Graveley, AN Brooks, JW Carlson, MO Duff, JM Landolin, L Yang, CG Artieri, MJ van Baren, N Boley, BW Booth, JB Brown, L Cherbas, CA Davis, A Dobin, R Li, W Lin, JH Malone, NR Mattiuzzo, D Miller, D Sturgill, BB Tuch, C Zaleski, D Zhang, M Blanchette, S Dudoit, B Eads, RE Green, A Hammonds, L Jiang, P Kapranov, L Langton, N Perrimon, JE Sandler, KH Wan, A Willingham, Y Zhang, Y Zou, J Andrews, PJ Bickel, SE Brenner, MR Brent, P Cherbas, TR Gingeras, RA Hoskins, TC Kaufman, B Oliver, and SE Celniker. The developmental transcriptome of Drosophila melanogaster. *Nature*, 471:473–479, 2010.

[38] AG Grimson, KKH Farh, WK Johnston, P Garrett-Engele, Lim LP, and DP Bartel. MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. *Molecular Cell*, 27:91–105, 2007.

[39] S Gu, L Jin, F Zhang, P Sarnow, and MA Kay. Biological basis for restriction of microRNA targets to the 3' unstranslated region in mammalian mRNAs. *Nature Structural and Molecular Biology*, 16:144–150, 2009.

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

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

[42] AT Hamilton, S Huntley, M Tran-Gyamfi, DM Baggott, L Gordon, and L Sutbbs. Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes. *Genome Research*, 16:584–594, 2006.

[43] D Hanahan and RA Weinberg. The hallmarks of cancer. *Cell*, 100:57–70, 2000.

[44] S Hannenhalli. Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics*, 24:1325–1331, 2008.

[45] L He and Hannon GJ. MicroRNAs: Small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, 5:522–531, 2004.

[46] DG Hendrickson, DJ Hogan, D Herschlag, JE Ferrell, and PO Brown. Systematic identification of mRNAs recruited to argonaute 2 by specific microRNAs and corresponding changes in transcript abundance. *PLoS ONE*, 3:e2126, 2008.

[47] R Hershberg and DA Petrov. Selection on codon bias. *Annual Review of Genetics*, 42:287–299, 2008.

[48] IL Hofacker, W Fontana, PF Stadler, LS Bonhoeffer, M Tacker, and P Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie*, 125:167–188, 1994.

[49] http://amigo.geneontology.org/cgi bin/amigo/term_enrichment.

[50] http://genome.ucsc.edu/.

[51] http://www.genenames.org.

[52] http://www.itol.embl.de.

[53] http://www.minotar.csail.mit.edu.

[54] http://www.pfam.sanger.ac.uk.

[55] http://www.targetscan.org.

[56] S Huang, S Wu, J Ding, J Lin, L Wei, J Gu, and X He. MicroRNA-181a modulates gene expression of zinc finger family members by directly targeting their coding regions. *Nucleic Acids Research*, 38:7211–7218, 2010.

[57] S Huntley, DM Baggott, AT Hamilton, M Tran-Gyamfi, S Yang, J Kim, L Gordon, E Branscomb, and L Stubbs. A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. *Genome Research*, 16:669–677, 2006.

[58] LD Hurst. Molecular genetics: The sound of silence. *Nature*, 471:582–583, 2011.

[59] D Iliopoulos, SA Jaeger, HA Hirsch, ML Bulyk, and K Struhl. STAT3 activation of miR-21 and miR-181b-1 via PTEN and CYLD are part of the epigenetic switch linking inflammation to cancer. *Molecular Cell*, 39:493–506, 2010.

[60] S Itzkovit and U Alon. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Research*, 17:405–412, 2007.

[61] J Ji, T Yamashita, A Budhu, M Forgues, HL Jia, C Li, C Deng, E Wauthier, LM Reid, QH Ye, LX Qin, W Yang, HY Wang, ZY Tang, CM Croce, and XW Wang. Identification of microRNA-181 by genome-wide screening as a critical player in epcam-positive hepatic cancer stem cells. *Hepatology*, 50:472–480, 2009.

[62] MA Jobling and P Gill. Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics*, 5:739–752, 2004.

[63] Y Kashi and DG King. Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics*, 22:253–259, 2006.

[64] M Kertesz, N Iovino, U Unnerstall, U Gaul, and E Segal. The role of site accessibility in microRNA target recognition. *Nature Genetics*, 39:1278–1284, 2007.

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

[66] WP Kloosterman and RH Plasterk. The diverse functions of microRNAs in animal development and disease. *Developmental Cell*, 11:441–450, 2006.

[67] A Krek, D Gruen, MN Poy, R Wolf, L Rosenberg, EJ Epstein, P MacMenamin, I Piedade, KC Gunsalus, M Stoffel, and N Rajewsky. Combinatorial microRNA target predictions. *Nature Genetics*, 37:495–500, 2005.

[68] D Kural, Y Ding, J Wu, AM Korpi, and JH Chuang. COMIT: Identification of noncoding motifs under selection in coding sequences. *Genome Biology*, 10:R133, 2009.

[69] A Lai, F Navarro, CA Maher, LE Maliszewski, N Yan, E O'Day, D Chowdhury, DM Dykxhoorn, P Tsai, O Hofmann, KG Becker, M Gorospe, W Hide, and J Lieberman. miR-24 inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to "seedless" 3'UTR microRNA recognition elements. *Molecular Cell*, 35:610–625, 2009.

[70] X Lampe, OA Samad, A Guigen, C Matis, S Remacle, JJ Picard, FM Rijli, and R Rezsohazy. An ultraconserved Hox-Pbx responsive element resides in the coding sequence of Hoxa2 and is active in rhombomere 4. *Nucleic Acids Research*, 36:3214–3225, 2008.

[71] MA Larkin, G Blackshields, NP Brown, R Chenna, PA McGettigan, H McWilliam, F Valentin, IM Wallace, A Wilm, R Lopez, JD Thompson,

TJ Gibson, and DG Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23:2947–2948, 2007.

[72] RC Lee, RL Feinbaum, and V Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75:843–854, 1993.

[73] BP Lewis, CB Burge, and DP Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120:15–20, 2005.

[74] LP Lim, NC Lau, P Garrett-Engele, A Grimson, JM Schelter, J Castle, DP Bartel, PS Linsley, and JM Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433:769–773, 2005.

[75] L Loewe and B Charlesworth. Background selection in single genes may explain patterns of codon bias. *Genetics*, 175:1381–1393, 2007.

[76] CM Loya, D Van Vactor, and TA Fulga. Understanding neuronal connectivity through the post-transcriptional toolkit. *Genes and Development*, 24:625–635, 2010.

[77] JR Lytle, TA Yario, and JA Steitz. Target mRNAs are repressed as efficiently by microRNA binding sites in the 5'UTR as in the 3'UTR. *PNAS*, 104:9667–9672, 2007.

[78] GAT McVean and B Charlesworth. The effects of hill-robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics*, 155:929–944, 2000.

[79] SM Mirkin. Expandable DNA repeats and human disease. *Nature*, 447:932–940, 2007.

[80] UA Orom, FC Nielsen, and AH Lund. MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. *Molecular Cell*, 30:460–471, 2008.

[81] S Park, X Yang, and Jeffery GS. Advances in computational protein design. *Current Opinion in Structural Biology*, 14:487–494, 2004.

[82] R Parker and U Sheth. P bodies and the control of mRNA translation and degradation. *Molecular Cell*, 25:635–646, 2007.

[83] N Pokala and TM Handel. Review: Protein design- where we were, where we are, where we're going. *Journal of Structural Biology*, 134:269–281, 2001.

[84] BJ Reinhart, FJ Slack, M Basson, AE Pasquinelli, JC Bettinger, AE Rougvie, HR Horvitz, and G Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403:901–906, 2000.

[85] JG Ruby, CH Jan, and DP Bartel. Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448:83–86, 2007.

[86] PC Sabeti, SF Schaffner, B Fry, J Lohmueller, P Varilly, O Shamovsky, A Palma, TS Mikkelsen, D Altshuler, and ES Lander. Positive natural selection in the human lineage. *Science*, 312:1614–1620, 2006.

[87] AJ Schetter, SY Leung, JJ Sohn, KA Zanetti, ED Bowman, N Yanaihara, ST Yuen, TI Chan, DL Kwong, GK Au, CG Liu, GA Calin, CM Croce, and CC Harris. MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma. *JAMA*, 299:425–436, 2008.

[88] M Schnall-Levin, L Chindelevitch, and B Berger. Inverting the viterbi algorithm: An abstract framework for structure design. *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[89] M Schnall-Levin, Y Zhao, N Perrimon, and B Berger. Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3'UTRs. *PNAS*, 107:15751–15756, 2010.

[90] M Selbach, B Schwanhäusser, N Thierfelder, Z Fang, R Khanin, and N Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455:58–63, 2008.

[91] WF Shen, YL Hu, L Uttarwar, E Passegue, and C Largman. MicroRNA-126 regulates HOXA9 by binding to the homeobox. *Molecular and Cellular Biology*, 14:4609–4619, 2008.

[92] DC Shields, DG Sharp, PM abd Higgins, and F Wright. "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. *Molecular Biology and Evolution*, 5:704–716, 1988.

[93] A Siepel, G Bejerano, JS Pedersen, AS Hinrichs, M Hou, K Rosenbloom, H Clawson, J Spieth, LW Hillier, S Richards, GM Weinstock, RK Wilson, RA Gibbs, WJ Kent, W Miller, and D Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15:1034–1050, 2005.

[94] SX Skapek, D Jansen, TF Wei, T McDermott, W Huang, EN Olson, and EY Lee. Cloning and characterization of a novel kruppel-associated box family transcriptional repressor that interacts with the retinoblastoma gene product, RB. *Journal of Biological Chemistry*, 275, 7212-7223 2000.

[95] A Stark, MF Lin, P Kheradpour, JS Pedersen, L Parts, JW Carlson, MA Crosby, MD Rasmussen, S Roy, AN Deoras, JG Ruby, J Brennecke, E Hodges, AS Hinrichs, A Caspi, B Paten, SW Park, MV Han, ML Maeder, BJ Polansky, BE Robson, S Aerts, J van Helden, B Hassan, DG Gilbert, DA Eastman, M Rice, M Weir, MW Hahn, Y Park, CN Dewey, L Pachter, WJ Kent, D Haussler, EC Lai, DP Bartel, GJ Hannon, TC Kaufman, MB Eisen, AG Clark, D Smith, SE Celniker, WM Gelbart, and M Kellis. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, 450:219–232, 2007.

[96] W Stephan, B Charlesworth, and McVean G. The effect of background selection at a single locus on weakly selected, partially linked variants. *Genetic Research*, 73:133–146, 1999.

[97] Y Tay, J Zhang, AM Thomson, B Lim, and I Rigoutsos. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, 455:1124–1128, 2008.

[98] S Tumpel, F Cambronero, C Sims, R Krumlauf, and LM Wiedemann. A regulatory module embedded in the coding region of Hoxa2 controls expression in rhombomere 2. *PNAS*, 105:20077–20082, 2008.

[99] S Tweedie, M Ashburner, K Falls, P Leyland, P McQuilton, S Marygold, G Millburn, and D Osumi-Sutherland. FlyBase: enhancing Drosophila gene ontology annotations. *Nucleic Acids Research*, 37:D555–D559, 2009.

[100] K Usdin. The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. *Genome Research*, 18:1011–1019, 2008.

[101] JM Vaquerizas, SK Kummerfeld, SA Teichmann, and NM Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10, 252-263 2009.

[102] VV Vazirani. *Approximation Algorithms*. Springer, 2004.

[103] AJ Viterbi. Error bounds for convolution codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269, 1967.

[104] T Warnecke and LD Hurst. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster. *Molecular Biology and Evolution*, 24:2755–2762, 2007.

[105] SA Wolfe, L Nekludova, and CO Pabo. DNA recognition by Cys2His2 zinc finger proteins. *Annual Review of Biophysics and Biomolecular Structure*, 29:183–212, 2000.

[106] DG Zisoulis, MT Lovci, ML Wilbert, KR Hutt, TY Liang, AE Pasquinelli, and GW Yeo. Comprehensive discovery of endogenous argonaute binding sites in Caenorhabditis elegans. *Nature Structural and Molecular Biology*, 17:173–179, 2010.

[107] M Zuker and P Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.