TRANSCRIPTIONAL REGULATION OF MICRORNA GENES AND THE REGULATORY NETWORKS IN WHICH THEY PARTICIPATE

by

David Lee Corcoran

BS, The University of Minnesota, 2002

MS, The University of Pittsburgh, 2004

Submitted to the Graduate Faculty of

The Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

David Lee Corcoran

It was defended on

July 8, 2008

and approved by

Dissertation Advisor:
Panayiotis V. Benos, PhD
Associate Professor, Department of Computational Biology
School of Medicine, University of Pittsburgh

Eleanor Feingold, PhD
Associate Professor, Department of Human Genetics
Graduate School of Public Health, University of Pittsburgh

Naftali Kaminski, MD
Associate Professor, Division of Pulmonary, Allergy and Critical Care Medicine
School of Medicine, University of Pittsburgh

M. Michael Barmada, PhD
Associate Professor, Department of Human Genetics
Graduate School of Public Health, University of Pittsburgh

Panayiotis V. Benos, PhD

TRANSCRIPTIONAL REGULATION OF MICRORNA GENES AND THE REGULATORY NETWORKS IN WHICH THEY PARTICIPATE

David Lee Corcoran, PhD

University of Pittsburgh, 2008

MicroRNA genes are short, non-coding RNAs that function as post-transcriptional gene regulators. Although they have been implicated in organismal development as well as a variety of human diseases, there is still surprisingly little known about their transcriptional regulation. The understanding of microRNA transcription is very important for determining their regulators as well as the specific role they may play in signaling cascades. This dissertation focused on the comparison of mammalian microRNA promoters and upstream sequences to those of known protein coding genes. This dissertation is also focused on determining potential regulatory networks that microRNA genes may participate in, particularly those networks involved in the TGFβ / SMAD signaling pathway.

The comparison of intergenic microRNA upstream sequences to those of protein coding genes revealed that the former are up to twice as conserved as the latter, except in the first 500 base pairs where the conservation is similar. Further investigation of the upstream sequences by RNA Polymerase II ChIP-chip revealed the transcription start site for 35 primary-microRNA transcripts. The identification of features capable of distinguishing core promoter regions from background sequences using a support vector machine approach revealed that the transcription start site of primary-microRNA genes share the same sequence features as protein coding genes. These results suggest that in fact microRNA genes are transcribed by the same mechanism by which protein coding genes are transcribed. This information allowed us to then identify the regulatory elements of microRNA genes in the same manner in which we use for protein coding genes. Identification of a SMAD family transcription factor binding site upstream of the human let-7d microRNA revealed a feed-forward regulatory circuit involved in epithelial mesenchymal transition. This provided the first evidence of a direct link between a growth factor and the expression of a microRNA gene.

The understanding of microRNA transcriptional regulation has great public health significance. The ability to understand how these post-transcriptional gene regulators function in cellular networks may provide new molecular targets for cures or therapies to a variety of human diseases.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 MICRORNA GENES

microRNA (miRNA) genes are small, non-coding RNAs (~22nt) that function as post-transcriptional gene regulators. miRNAs were first identified in 1993 by Lee *et al.* who found that, in *C. elegans,* the non-coding gene *lin-4* contained two short transcripts that were capable of post-transcriptional regulation of the gene *lin-14* resulting in the proper cellular signal timing in larval development (1). Since the first description of miRNAs a wide variety of studies have been published analyzing everything from their biogenesis, their regulation, their function and the role that they play in a variety of cellular pathways.

miRNA genes, or miRNAs, are currently believed to be mostly transcribed by RNA Polymerase II (Pol II) (2), although a few may be transcribed by RNA Polymerase III (3). There are two different classes of miRNAs when discussing their transcription mechanism, those found within annotated genes (intronic miRNAs) and those found in intergenic regions of the genome (intergenic miRNAs). It is presently believed that all intronic miRNAs are co-transcribed along with their host gene; this has been shown in both expression correlation studies (4) as well as PCR based biochemical verification (5). Intergenic miRNAs have been postulated to come from transcripts of up to 50kb in length, allowing for the co-transcription of neighboring miRNAs (polycistronic miRNA clusters) (4).

The initial full-length miRNA transcripts are called primary-miRNAs (pri-miRNAs); these transcripts may contain one or multiple miRNA genes. The region on the transcript surrounding the miRNAs form a hairpin structure which is cleaved out by the protein *Drosha*; these resulting hairpin structures are called preliminary-miRNAs (pre-miRNAs). The pre-miRNAs then leave the nucleus for the cytoplasm where they are processed by another protein called *Dicer* resulting in their final size (~22nt). The short mature miRNA then associates with a protein complex containing, among others, the protein *Argonaut* whereupon it is then able to carry out its function (6).

Once the mature miRNAs have become associated with the *Argonaut* protein complex they become free to bind to their target sites through base complementarity on the 3' untranslated region of mRNA transcripts, typically 7 – 8nt in length (6). This binding can result in either the full degradation of the target mRNA transcript or the blocking of its translation. miRNA target prediction methods have shown that each miRNA gene may be able to target many mRNAs, while each mRNA may be the target of multiple miRNAs (7-9).

While the process of miRNA regulation of mRNAs is beginning to be well understood, the transcriptional regulation of miRNA genes, specifically intergenic miRNAs, is still being intensely studied. The finding, as stated above, that Pol II may be regulating miRNAs led us and others to the hypothesis that the same type of regulatory elements that control protein coding gene transcription may as well control miRNA gene transcription (10). Pri-miRNA transcripts have been identified, in some cases, to have 5' caps and Poly(A) tails, both properties of the Pol II transcription of protein coding genes (2). Currently only a handful of pri-miRNA transcription start sites have been identified biochemically (2, 10-12), though computational tools have been used to attempt the prediction of many more of them (11, 13).

The identification of pri-miRNA transcription start sites is important for the identification of their promoter regions, the sequence that contains the transcription factor binding sites responsible for their transcriptional regulation. The identification of these biding sites will further help in the understanding of the gene regulatory networks in which miRNAs participate.

1.2 PROMOTER REGIONS

Transcription of protein coding genes as well as some small RNAs, such as microRNAs (miRNAs), is carried out by Pol II. While Pol II binds to the DNA at the transcription initiation point, it is not capable of directly recognizing its target (14). A complex of proteins in a region known as the core promoter binds to the DNA whereupon they recruit Pol II to the transcription start site (TSS). Other proteins, called transcription factors (TFs), then bind to the proximal promoter or enhancer regions to either initiate (activators) or block (repressors) the activation of Pol II.

The core promoter region typically consists of the couple hundred base pairs surrounding the TSS of a gene. This region was once thought to contain a handful of known features able to be bound by elements of the Pol II protein complex; though it is now known that there is a wide diversity of properties that can be identified. It was initially believed that the core promoter regions consisted of a TATA box (~30bp upstream of the TSS) and an initiator sequence (Inr; overlaps the TSS). Recent studies have estimated the prevalence of these two sequences in only about 16% of human promoters (15), typically for tissue specific genes (16). It has been identified that approximately half of the human core promoters are located around CpG islands. CpG islands are regions of a high concentration of CG dinucleotides, which are very underrepresented across the genome (17).

For non-CpG island related promoters it was discovered that addition sequences such as the down stream promoter element (DPE; ~28bp downstream of the TSS) and the TFIIB recognition element (BRE; ~35bp upstream of the TSS) were also targets for proteins involved in the recruitment of RNA Pol II to the TSS besides the TATA box and the Inr. While some hypothesized that at least 2 of these 5 elements were necessary for the transcriptional initiation complex to bind (18), other researchers have identified gene specific elements capable of binding this complex such as the downstream core element (DCE) in the human *β-globin* promoter (19) and the multiple start site downstream element (MED-1) in the *pgp1* promoter (20). This suggests that the core promoter structure may be more complicated than originally hypothesized.

CpG islands are found at a low frequency in the genome because methylated CG dinucleotides can easily be mutated to TG dinucleotides, a process that is not corrected by DNA repair mechanisms (19). Genes that have CpG islands in their promoters tend to be ubiquitously expressed across most tissues and throughout development. DNA methylation is a mechanism for which the cell can block the binding of transcription factors to promoters in order to prevent transcription of certain genes; because genes with CpG islands tend to be continuously expressed, they have not had as many opportunities to be methylated relative to the remainder of the genome (17). Another feature that distinguishes genes with CpG islands is that they are more likely to have multiple TSSs that can span over 100bp whereas TATA-Inr containing promoters tend to have just one TSS (16). The transcription factor *Sp1* is capable of binding to CpG islands and recruiting Pol II (21).

1.3 TRANSCRIPTION FACTORS

The regulatory regions outside of the core promoter are the proximal promoter region and the enhancer regions. These regions, which can be located upstream of the gene, downstream of the gene, or even in the gene's introns, are bound by TFs that activate or repress the functionality of Pol II (22). TFs will typically consist of a DNA binding domain and in the case of activators, an activation domain. Each TF usually binds to a specific set of sequence motifs 6-15bp in length (23). The over 2,000 human TFs can be broken down into families based upon their structural properties that will typically correspond to their preferred binding motif (24, 25). TFs can function individually, in tandem, or in competition with each other (26).

The identification of transcription factor binding sites (TFBSs) in the genome is an important and highly researched subject. The advancement of large-scale chromatin immunoprecipitation technology (ChIP-chip) has provided biologists with the tools necessary to identify many binding sites for a specific factor (27). The limitations of these studies are that the results only provide information on one specific cell type, one cellular condition, and only a single TF. In addition to laboratory procedures that can identify TFBSs, computational biologists have taken up the task of locating their motifs given the complete sequences of genomes and some external information.

A variety of computational methods have been developed over the years for the identification of TFBSs. The short and often degenerate nature of the sequences makes them a challenge to identify over large genomic regions (28). Algorithms have been developed that search for specific strings of sequences that match known TFBSs, called library based methods. Other methods use an IUPAC alphabet in the form of a consensus sequence to match the variability in TFBSs (29). The most common method of representing the binding motifs of a TF is a position-specific scoring matrix (PSSM). A PSSM provides a mathematical model that represents all of the known binding sites for a given TF (23). The PSSM can be combined with other features to help in the efficiency of identifying TFBSs such as sequence conservation across species or looking for common TFBSs in the promoters of genes that appear to be co-regulated given the similarity in their expression profiles (28, 30-33).

The identfication of TFBSs is an essential step in the understanding of gene regulatory networks. The further analysis of miRNA promoters and the transcription factors that bind them will be an essential component to the understanding of the regulatory networks in which they participate.

1.4 PROJECT OVERVIEW

The purpose of this project is to develop an understanding of the transcriptional regulation of microRNA genes. This knowledge will then be used to determine which factors might regulate these genes and therefore which regulatory networks they may be involved in. Previous studies have suggested that microRNA genes may be transcribed by either Pol II (2) or RNA Polymerase III (3) . To identify which is most likely responsible for the transcription of microRNA genes we will compare the features of the microRNA upstream regions to those of genes transcribed by Pol II or Pol III. Once the RNA Polymerase found to be the most likely responsible for microRNA transcription is identified further analysis into which factors are regulating specific microRNA genes will be performed. With putative regulators identified, we will attempt to identify specific regulatory networks in which microRNA genes may participate.

Chapter 2 of this dissertation will begin with the analysis of the conservation across species of the upstream sequences of microRNA genes and how they compare to other known classes of genes such as protein coding genes which are transcribed by Pol II as well as non-coding RNAs known to be transcribed by Pol III. From that analysis I will show that in fact microRNA upstream sequences are similar to those of protein coding genes, which are transcribed by RNA Polymerase II. Chapter 3 will then demonstrate that one of the most common biological methods of transcription factor binding site identification, ChIP-chip, can be best modeled by the number of motifs recognized by that factor within the probed genomic region. In Chapter 4 we will analyze RNA Polymerase II chromatin immunoprecipitation data to identify where the transcription start site is for a variety of microRNA genes. With the location of the true transcription start sites; we will then be able to compare the features of microRNA core promoters to those of protein coding genes. The results of that analysis will further confirm that the same transcriptional machinery of protein coding genes in fact transcribes microRNA genes.

With the knowledge gained in Chapter 4, we will then proceed in Chapter 5 to identify any putative feed-forward loops involving microRNA genes by comparing the presence of sites that may be bound by a given transcription factor upstream of both a microRNA gene and any of its targets. The focus of Chapter 5 will be on transcription factors and microRNA genes that may be involved in the TGFβ / SMAD signaling pathway. Chapter 6 will then focus on verifying one of the putative feed-forward loops and the role that the human microRNA gene let-7d may play in epithelial mesenchymal transition, a cellular process seen in development, cancer and idiopathic pulmonary fibrosis.

1.5 PUBLIC HEALTH SIGNIFICANCE

Recent studies have demonstrated that miRNAs may play an important role in a variety of human diseases such as cancer (34), fragile X syndrome (35) and heart failure (36). As more is learned about these gene regulators we are likely to see the number of human diseases that miRNAs are involved in greatly increase. The understanding of the transcriptional regulation of miRNAs is a vital step toward complete deciphering of the cellular processes in which they are involved. The understanding of these pathways brings with them the potential for new treatments and therapies for a variety of genetic factors that affect public health.

2.0 REGULATORY CONSERVATION OF PROTEIN CODING AND MICRORNA GENES IN VERTABRATES

David L. Corcoran[1], Shaun Mahony[2], Eleanor Feingold[1], Panayiotis V. Benos[2]

[1]Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA; [2]Department of Computational Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

This chapter has been published in a modified form by BioMed Central, © Mahony *et al*. Full Citation:

Mahony S, Corcoran DL, Feingold E, and Benos PV 2007 Regulatory conservation of protein coding and microRNA genes in vertebrates: lessons from the opossum genome. Genome Biol 8(5):R84

2.1 BACKGROUND

In protein coding genes, gene regulation is primarily controlled by short DNA sequences in the vicinity of the gene's transcription start site (TSS) that are targets for transcription factor proteins. A high degree of evolutionary conservation of these promoter regions can be attributed to functional *cis*-regulatory elements. The increased conservation in the biologically more important parts of the promoter region has been explored by various phylogenetic footprinting algorithms (28, 32, 37, 38) to improve the prediction of transcription factor binding sites (TFBSs) in vertebrate genomes.

Another major mechanism for control of gene expression is provided by microRNA (miRNA) genes. miRNAs are small (22 to 61 bp long), noncoding RNAs that downregulate their target genes via base complementarity to their mRNA molecules (1, 39). Each miRNA can target multiple genes and each gene can be the target of multiple miRNAs (8, 9, 40). In vertebrates, their expression often is tissue specific (41) and has been shown to play an important role during development (42-44). Although some miRNAs are found in the introns of coding genes and therefore are probably regulated by the promoters of the genes in which they reside (5), others are located in the intergenic parts of the genome. Little is known about the transcriptional regulation of these intergenic miRNAs, although RNA polymerase II appears to be involved in the process (2). This suggests that miRNAs may have active promoter regions that contain *cis*-regulator elements similar to coding genes. In this study we will further explore this hypothesis by comparing the conservation of the upstream region of known miRNAs to that of promoters for protein coding genes which are regulated by RNA Polymerase II as well as non-coding RNAs known to be regulated by RNA Polymerase III. Identification of the method by which miRNA genes are transcribed can lead to the identification of the factors that are responsible for their regulation.

2.2 RESULTS AND DISCUSSION

2.2.1 Conservation of the Upstream Regions of Protein Coding and Intergenic miRNA Genes

We calculated the conservation of the 5 kilboase (kb) upstream regions of all RefSeq protein coding genes as well as the known intergenic pre-miRNAs using a sliding window approach, as described in *Materials and Methods*. Only intergenic miRNAs were included in this analysis because intronic miRNAs have been shown to be co-transcribed with their corresponding host gene (5). With little known about the transcriptional regulation of non-intronic miRNA genes it cannot be assessed whether the miRNA upstream regions evolved at the same, slower, or faster rate than those of protein coding genes, and whether their conservation pattern across the upstream region indicates parts of potential biological importance. The phylogenetic tree of the species included in this analysis is plotted in Figure 2.1.

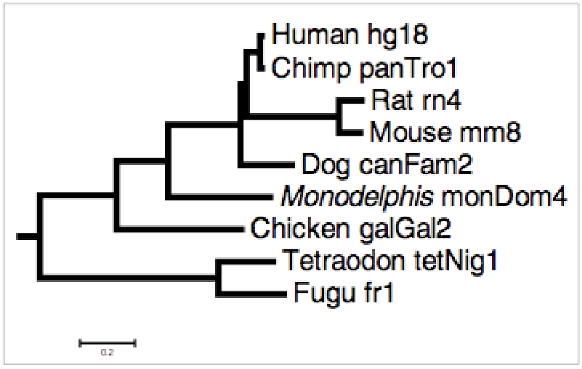
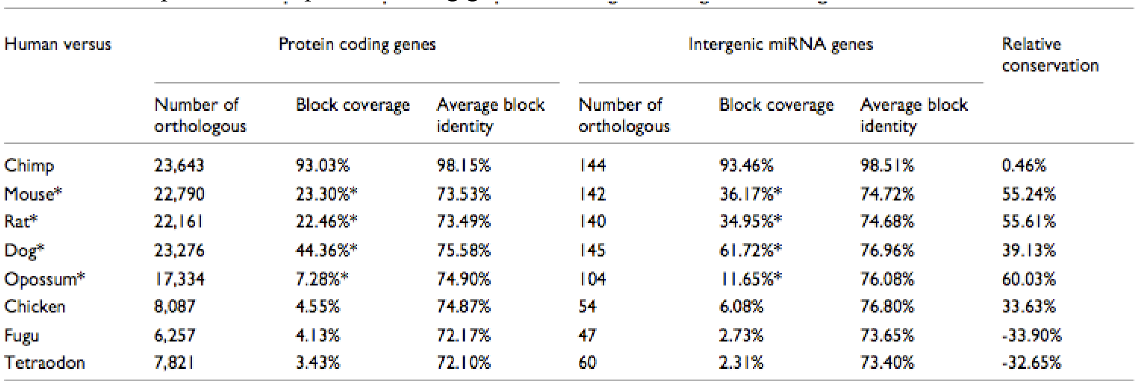


Figure 2.1 Phylogenetic tree of the species examined in this chapter

Table 2.1 presents the number of orthologous genes in each species (derived from the MULTIZ University of California, Santa Cruz [UCSC] synteny based alignments), the average block coverage of their upstream regions, and the average percentage identity within these conserved blocks. For the calculation of the average percentage identity, the conservation percentage of each block is multiplied by the total length of the block. In other words, the average block conservation corresponds to the number of bases that are identical in all conserved blocks of one promoter over the length of the blocks in this promoter. The human genes were used as reference for all pair-wise comparisons. Surprisingly it was found that, with the exception of teleosts and chimp, the conservation in the upstream regions of the miRNA genes is 34% to 60% higher on average than that in the protein coding genes. This is independent of the average block identity, which remains practically the same between two types of genes in these comparisons (Table 2.1). In all non-primate mammals the average block coverage in the miRNA upstream sequences is significantly higher than that in the promoters of the protein coding genes (Wilcoxon rank-sum test: p = 6x10^{-4} for opposum and p = 10^{-14} to 10^{-16} for rodents and dog).

In all of the pair-wise comparisons (Table 2.1), except human-chimp, the average block identity is about the same (72% to 77%), regardless of the evolutionary distance or the type of gene (protein coding or miRNA); because the block conservation threshold was 65%, this equivalency indicates that a reduction in the number of conserved blocks rather than a uniform decrease in the similarity is responsible for the observed conservation rates. Such a pattern of evolution is expected if the *cis*-regulatory sites are organized in clusters located in these upstream regions. These clusters might contain regulatory elements specific to, for instance, primates only, eitherians only, and so on.

Table 2.1 Conservation in the 5kb upstream sequences in all protein coding and intergenic miRNA genes. *Species for which the block coverage of miRNA gene upstream regions is statistically significantly higher than that of the promoters of protein coding genes



Human versus	Protein coding genes			Intergenic miRNA genes			Relative conservation
	Number of orthologous	Block coverage	Average block identity	Number of orthologous	Block coverage	Average block identity	
Chimp	23,643	93.03%	98.15%	144	93.46%	98.51%	0.46%
Mouse*	22,790	23.30%*	73.53%	142	36.17%*	74.72%	55.24%
Rat*	22,161	22.46%*	73.49%	140	34.95%*	74.68%	55.61%
Dog*	23,276	44.36%*	75.58%	145	61.72%*	76.96%	39.13%
Opossum*	17,334	7.28%*	74.90%	104	11.65%*	76.08%	60.03%
Chicken	8,087	4.55%	74.87%	54	6.08%	76.80%	33.63%
Fugu	6,257	4.13%	72.17%	47	2.73%	73.65%	-33.90%
Tetraodon	7,821	3.43%	72.10%	60	2.31%	73.40%	-32.65%

2.2.2 Distribuition of Conserved Blocks in the Upstream Sequence of Protein Coding and Intergenic miRNA Genes

We plotted sequence conservation as a function of the distance from the transcription start site to investigate further the differences in conservation of the upstream regions (Figure 2.2). We found that in the first 500bp the sequence conservation of the miRNA genes is almost identical to that of the promoters of the protein coding genes (*R* values > 0.9 and usually much higher; regression *t-test*: $p < 10^{-19}$). In protein coding genes this is typically the region with the highest concentration of known *cis*-regulatory elements. From all known human and mouse TFBSs in the TRANSFAC database (45), 69.1% and 65.1% respectively, are annotated as being located in the proximal 500bp region (data not shown). Interestingly, Lee and coworkers (2) showed that this region is sufficient to drive expression of the miR-23a~27a~24-2 intergenic miRNA cluster by RNA polymerase II. These findings were further tested by analysis of the upstream sequence conservation of the tRNA genes in the human genome. It has been long established that the *cis*-regulatory elements of the tRNA genes are located downstream of their transcription start site (46). It was found that the sequence conservation for the tRNA genes was constant throughout their 5kb upstream regions (Figure 2.2; green dashed line).

The conservation rates in both protein coding and miRNA genes declines after the first 500bp and becomes almost constant. We also observed that the difference between these two types of genes is that, in the case of miRNAs, the constant conservation rate is up to twofold higher than that in the protein coding genes for rodents, dog, opposum and chicken. It was found that these differences were statistically significant (data not shown). Similarly high conservation rates are observed in chimp for both types of genes, probably reflecting the generally high conservation rate throughout the genome. By contrast, similarly low conservation rates are observed for the fugu fish and tetraodon. It should be noted, however, that the higher conservation rates are statistically significant in the (nonprimate) mammals.

It is not clear whether this increased upstream sequence conservation is a general biologic feature of the miRNA upstream regions or is an artifact of the method used to discover miRNA genes. It is possible, for example, that the known intergenic miRNAs happen to fall in more conserved regions of the genome. This may be related to the way in which the miRNAs were originally identified (through high similarity to known miRNAs). However, it is also possible that because miRNAs are involved in highly regulated vital cell or organismal processes such as development (42-44), there is a much greater selective pressure on their regulatory regions. To investigate this question further, the conservation of the upstream sequences between miRNA genes and those genes identified as developmental according to GO classification were compared (Figure 2.2; light blue dashed line). It was observed that the upstream conservation of the developmental genes in all mammals is uniformly higher than the overall average and similar to the conservation of the miRNA genes, especially the first 2,000bp. This is true for all species examined, although in the non-mammalian vertebrates the overall upstream sequence conservation for all types of genes is similarly low (10% lower after the first 500bp; Figure 2.2). The fact that miRNA genes have been implicated in the regulation of various developmental processes (47) may partly explain the similar conservation rates in their upstream regions and the promoters of developmental genes, also indicating that analogous mechanisms and *cis*-elements may regulate the expression of the corresponding gene.

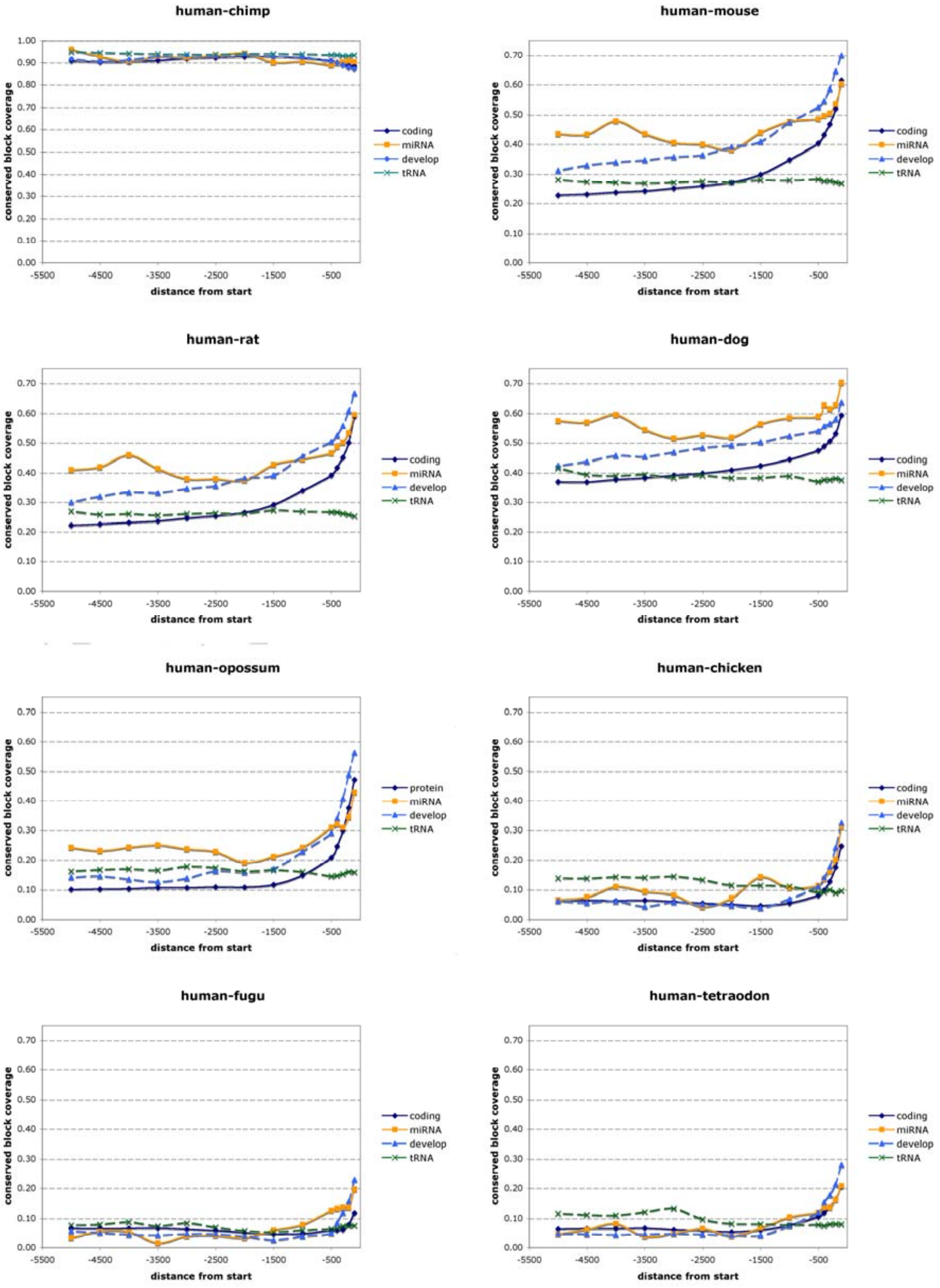


Figure 2.2 Upstream sequence conservation of protein coding, tRNA and miRNA genes

2.3 CONCLUSION

This study is the first to analyze conservation of the upstream regions of protein coding genes in relation to the upstream regions of intergenic miRNA genes. The latter was found to be about twice as conserved as the former beyond the first 500bp. The reason for this conservation is currently unknown. The first 500bp appears to be equally conserved in both types of genes, a feature that is missing from the upstream sequence of the tRNA genes. This indicates that similar mechanisms of gene regulation may be in place, which is in agreement with other studies (2, 48). The difference in conservation rates is more apparent in the mammalian lineages and may reflect similarities in mammalian gene regulation.

2.4 MATERIALS AND METHODS

2.4.1 microRNA Gene Dataset

Human miRNA genes were retrieved from the miRBase database (7) and the UCSC Genome Browser (version hg18, March 2006) (49, 50). Cross-referencing them with the miRNAMap dataset (51) identified 169 putatively intergenic miRNA genes. The sequences of these miRNAs were used in BLAST-like Alignment Tool (BLAT) (52) aligned against the latest UCSC human genome and their where there exact locations were identified. Following observations in previous studies (2, 53), two miRNA genes were considered to be co-transcribed if their starting points were less than 250 bp apart. In this way, 12 clusters containing 31 genes were identified. Only the 5'-most gene in a cluster was considered in the analysis. Five miRNA genes were found to reside within large introns of protein coding genes, and although they may have their own regulatory regions, they were excluded from further analysis. This resulted in a dataset of 145 human intergenic miRNA genes. The coordinates of the BLAT outputs were used to retrieve up to 5kb regions upstream of the gene start site.

2.4.2 Pair-wise Species Comparison

Pair-wise species alignment for both protein coding and miRNA genes were retrieved from the 17-species MULTIZ multiple alignments (54), which are available from the UCSC web server (50). The MULTIZ algorithm builds a multiple alignment from local pair-wise BLASSTZ alignments of the reference genome with each other genome of interest (54, 55). Each base in the reference genome is aligned to at most one base in the other genomes, and the alignment is guided by synteny. In this study, the results from pair-wise comparisons of human (56) with four eutherian mammals (chimpanzee (57), mouse (58), rat (59) and dog (60)), the newly sequenced opossum (61), chicken (62), fugu (63) and tetradon (64) are presented. A phylogenetic tree for those species and with branch lengths derived from the ENCODE project Multi-Species Sequence Analysis group (September 2005) is shown in Figure 2.1. This tree was generated using the phyloGif program (65) from the Threaded Blockset Aligner (TBA) alignments over 23 vertebrate species and is based on 4D sites (similar to the tree presented by Margulies and coworkers (66)).

For each pair-wise comparison, the corresponding (aligned) 5kb upstream sequence was retrieved directly from the MULTIZ alignments for greater accuracy, using the human genes as a reference. If other genes were found within this 5kb range, then the upstream sequences were shortened accordingly to exclude the additional genes. 65% was used as the conserved block threshold in this study, which is similar to that in previous studies (31, 67, 68).

2.4.3 tRNA Dataset

Human tRNA genes and pair-wise alignments were extracted from the UCSC Genome Browser Database (50, 65) using the genomic MULTIZ alignments as described previously. Genes that were found to be facing opposite directions in the genome ('head-to-head') and their start were closer than 2.5kb apart were excluded from the analysis. This rule excluded 156 genes. The final human tRNA dataset included 1,795 upstream sequences.

2.4.4 Block Conservation

In this study, sequence conservation is expressed as conserved block coverage. A sliding window of width 50bp and step size 10bp was used to find conserved regions (or blocks) of at least 65% identitiy between human and each other species. Each pair-wise alignment was extracted from the MULTIZ multiple alignments. Sauer and coworkers (69) have shown that the 65% identity threshold most effectively separates TFBSs from background sequences in human-rodent comparisons. The percentage of human 5kb upstream sequence that is located within conserved blocks is denoted the 'conserved block coverage'. The 'average block conservation' is the percentage of identifcal bases in conserved blocks over all bases in conserved blocks.

3.0 MODELING CHROMATIN IMMUNOPRECIPITATION DATA WITH TRANSCRIPTION FACTOR BINIDING SITE IDENTIFICATION METHODS

3.1 BACKGROUND

With high throughput, microarray based chromatin immunoprecipitation (ChIP-chip) quickly becoming established as the most effective means of identifying transcription factor binding sites (TFBSs) there becomes a need to determine if any genomic features can be implemented *in silico* to effectively model the ChIP-chip outcome. ChIP-chip functions by 'freezing' the cellular machinery, resulting in the fixation of the DNA to all bound proteins. The DNA is then sheared into fragments (~1kb in size) with all of the proteins still bound. Specific proteins, often transcription factors (TFs), with the DNA still attached can then be precipitated out and the DNA isolated. The DNA is then labeled with a dye and placed on an array that contains probes with complementary sequences to regions surrounding the transcription start sites of all annotated genes. An image reading system then identifies the probes with the most DNA bound, providing the location of the TFBSs (27).

While ChIP-chip is able to biochemically identify the location of TFBSs, it is limited in that it only identifies the TFs bound at the instant the cells are 'frozen.' The recent sequencing of the genomes for a variety of species provides computational biologists the opportunity to identify TFBSs *in silico* across the whole genome. TFBS prediction is a complicated task because of the short (6-18bp), often degenerate, motifs that TFs often bind (23).

One type of method others have previously developed to identify putative TFBSs is to identify clusters of genes that are co-expressed across a variety of cellular conditions or over time in response to a stimulus. The upstream sequences of the clustered genes are then collected and searched for common motifs (30, 33). If sequences for some binding sites are known, a 'consensus' sequence can be created using the IUPAC alphabet (29) to scan genomic regions for matching sequences. A more mathematically appropriate model, called a position-specific scoring matrix (PSSM), can also be generated from known sites and used to scan for matching sequences given a false discovery rate threshold cutoff (23).

The previously described approaches tend to have a low specificity given the size of the background sequence in the genome (70). To reduce the search space, studies began to incorporate the use of sequence conservation assuming that the functional elements will be conserved across species in an approach called phylogenetic footprinting (28, 31, 32). The drawback to using phylogenetic footprinting is that its ability to increase the specificity and sensitivity of the algorithm relies heavily on the evolutionary distance between the species being compared (71). Another approach to TFBS identification involves identifying clusters of low-affinity sites for a TF that may be near a true site; these low affinity sites can act either as 'backup' sites or as a means of keeping the TF in the proximity of the true site (72).

The purpose of this study is to test a variety of TFBS identification methods to determine which of these methods are able to accurately predict the results of ChIP-chip data. These models will be generated by logistic regression with the TFBS identification method being the independent variable and the outcome variable being the bound or unbound status of the TF from the ChIP-chip data.

3.2 RESULTS AND DISCUSSION

3.2.1 Analysis of Yeast ChIP-chip Data

Saccharomyces cerevisiae is an organism often used to study eukaryotic transcriptional regulation because of its relatively small number of TFs as well as its short and well-defined promoter regions (73). Due to the relative simplicity of the yeast promoter regions, a large number of biochemical studies have been carried out that allow for the verification of computational predictions, such as the study of Lee *et al.* (73) in which they have performed genome-wide ChIP-chip experiments for 88% of the 144 known TFs.

With ChIP-chip established as one of the most commonly used and reliable methods of large-scale identification of TFBSs the question arises as to which TFBS prediction method provides the most accurate model for the ChIP-chip results. To test this, we used a variety of TFBS prediction methods were used as a predictor in a logistic regression analysis with the ChIP-chip results as the outcome variable. Logistic regression is often used to determine the significant predictors of a binary outcome; in this case the binary outcome is the ChIP-chip data as it corresponds to all of the promoter regions in the yeast genome (1 = bound, 0 = unbound; see *Materials and Methods*) and the predictor variable is the value obtained from the TFBS prediction algorithms.

Table 3.1 *p*-values representing the significance of the given TFBS identification method as a predictor of ChIP-chip data in *Saccharomyces cerevisiae* for each TF produced by logistic regression.

Transcription Factor	PSSM Cutoff 1 Count		PSSM Cutoff 2 Count		Consensus String Count		Lowest PSSM Score		Lowest Window PSSM Score	
SUM1	6x10^{-18}	2x10^{-26}	2x10^{-9}	4x10^{-19}	2x10^{-19}	2x10^{-17}	4x10^{-19}	9x10^{-20}	0.001	0.036
CAD1	0.525	0.004	0.525	0.004	0.026	4x10^{-6}	0.035	0.195	0.529	0.859
ZAP1	0.001	0.018	0.037	0.391	2x10^{-5}	3x10^{-4}	2x10^{-4}	0.952	0.216	0.539
BAS1	5x10^{-11}	4x10^{-12}	5x10^{-11}	5x10^{-12}	8x10^{-15}	8x10^{-16}	0.001	0.018	0.634	0.790
ACE2	1x10^{-5}	5x10^{-7}	1x10^{-5}	5x10^{-7}	2x10^{-4}	9x10^{-4}	0.020	0.049	0.024	0.014
SFP1	5x10^{-6}	0.004	3x10^{-4}	0.007	0.005	0.004	2x10^{-7}	0.131	0.003	0.115
MCM1	2x10^{-5}	3x10^{-6}	9x10^{-4}	1x10^{-5}	2x10^{-8}	2x10^{-9}	7x10^{-6}	4x10^{-6}	0.051	0.009
STB1	0.003	2x10^{-5}	0.020	0.004	3x10^{-7}	6x10^{-10}	0.037	0.079	0.747	0.534
ROX1	0.221	0.117	0.090	0.762	0.982	0.985	0.080	0.485	0.019	0.268
ADR1	0.333	0.993	0.993	0.993	0.247	0.993	0.673	0.436	0.103	0.328
BAS1	7 x10^{-12}	5 x10^{-08}	5 x10^{-12}	2 x10^{-08}	1 x10^{-15}	5 x10^{-12}	7 x10^{-4}	0.044	0.213	0.395
CAD1	3 x10^{-11}	1 x10^{-7}	2 x10^{-13}	2 x10^{-13}	7 x10^{-10}	4 x10^{-9}	2 x10^{-16}	3 x10^{-5}	0.079	0.737
CBF1	1 x10^{-122}	2 x10^{-84}	2x10^{-126}	6 x10^{-84}	5 x10^{-111}	2 x10^{-58}	6 x10^{-125}	7 x10^{-63}	2 x10^{-14}	2 x10^{-15}
DAL81	0.148	0.074	0.997	0.997	0.998	0.998	0.648	0.244	0.128	0.123
DAL82	0.102	0.018	0.042	0.050	2 x10^{-9}	6 x10^{-6}	0.704	0.702	0.019	0.017
FHL1	2 x10^{-28}	1 x10^{-7}	1 x10^{-68}	1 x10^{-20}	6 x10^{-70}	2 x10^{-29}	9 x10^{-93}	1 x10^{-42}	0.176	0.047
GAT1	0.040	0.259	0.039	0.259	0.358	0.546	0.091	0.715	0.455	0.770
GCN4	7 x10^{-54}	3 x10^{-58}	9 x10^{-61}	3 x10^{-62}	2 x10^{-80}	1 x10^{-78}	3 x10^{-46}	4 x10^{-46}	0.678	0.957
GLN3	8 x10^{-06}	9 x10^{-13}	8 x10^{-5}	3 x10^{-9}	2 x10^{-6}	7 x10^{-8}	0.001	3 x10^{-6}	0.077	0.057
HAP4	0.477	0.086	0.260	0.048	0.254	0.032	0.273	0.116	0.739	0.799
HAP5	0.006	0.001	0.006	0.001	0.001	7 x10^{-6}	0.066	0.003	0.467	0.896
LEU3	1 x10^{-18}	4 x10^{-17}	7 x10^{-20}	3 x10^{-20}	1 x10^{-16}	6 x10^{-19}	2 x10^{-22}	8 x10^{-20}	3 x10^{-9}	3 x10^{-9}
MET31	0.961	0.085	0.004	0.006	0.051	0.008	0.230	0.271	0.934	0.598
MET32	1 x10^{-6}	8 x10^{-13}	5 x10^{-9}	2 x10^{-16}	5 x10^{-14}	6 x10^{-19}	6 x10^{-13}	1 x10^{-12}	0.594	0.626
MET4	0.006	5 x10^{-6}	1 x10^{-4}	1 x10^{-9}	0.974	0.974	3 x10^{-9}	2 x10^{-7}	0.015	0.054
MOT3	3 x10^{-10}	8 x10^{-5}	2 x10^{-10}	1 x10^{-4}	4 x10^{-9}	1 x10^{-4}	0.006	0.003	0.082	0.453
PHO2	0.075	0.749	0.032	0.247	0.989	0.989	0.052	0.686	0.761	0.798
PUT3	0.075	0.004	7 x10^{-4}	2 x10^{-5}	0.977	0.984	1 x10^{-10}	2 x10^{-4}	0.069	0.038
RCS1	0.012	0.005	5 x10^{-7}	3 x10^{-6}	0.032	0.006	3 x10^{-5}	0.002	0.893	0.490
RPH1	0.965	0.54	0.988	0.992	0.930	0.995	0.225	0.001	0.018	0.030
RTG3	0.028	0.012	0.028	0.012	0.798	0.213	0.039	0.042	0.005	0.006
SFP1	7 x10^{-30}	4 x10^{-17}	1 x10^{-31}	6 x10^{-23}	3 x10^{-37}	3 x10^{-26}	3 x10^{-29}	9 x10^{-15}	0.005	0.686
SIP4	0.005	2 x10^{-4}	5 x10^{-5}	7 x10^{-6}	0.993	0.992	0.006	0.030	0.128	0.146
STP1	0.064	0.047	0.064	0.047	0.985	0.987	0.013	0.385	0.447	0.096
UGA3	0.0389	0.987	0.537	0.990	0.993	0.995	0.012	0.013	0.205	0.076

Cutoff 1: 1 False Positive per 3000bp
Cutoff 2: 1 False Positive per 6000bp
Red text indicates statistical significance (*p*-value <= 0.005)

▨ Only sites or values within evolutionary conserved regions were evaluated, therefore taking phylogenetic footprinting into account

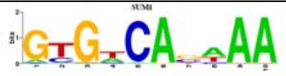
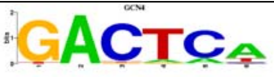
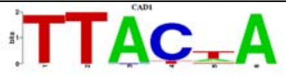
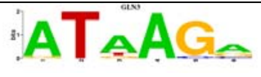
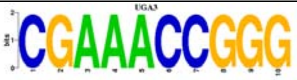
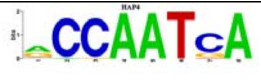
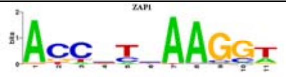
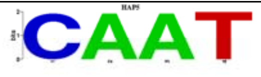
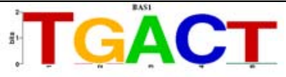
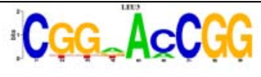
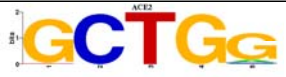
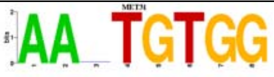
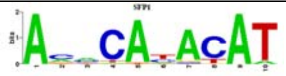
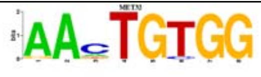
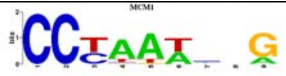
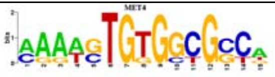
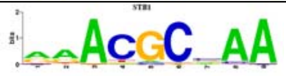
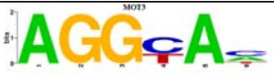
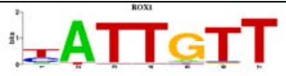
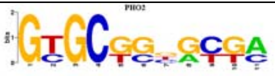
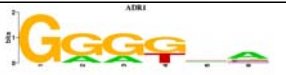
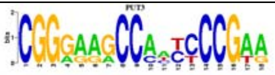
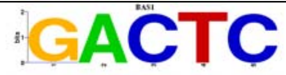
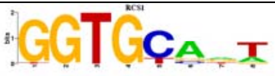
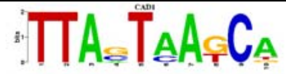
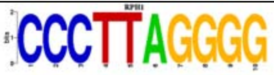
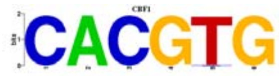
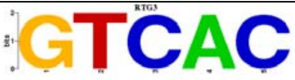
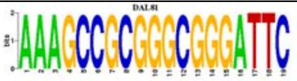
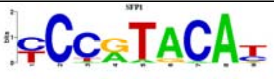
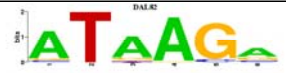
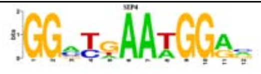
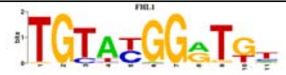
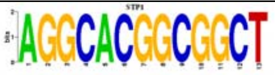
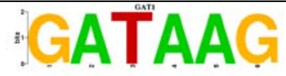
The first type of TFBS identification method we tested was to count the number of instances of the factor motif being present in the promoter region of each gene (see *Materials and Methods).* This approach was accomplished by using the consensus sequence and a PSSM with two different score cutoffs for each TF, which allow for different ‘noise-to-signal’ ratios. The next approach analyzed was to take the lowest scoring, and therefore highest affinity, site in each promoter and use that as the independent variable in the logistic regression. The final method compared with the ChIP-chip data was based upon Zhang *et al.* (72) study that suggests clusters of low-affinity sites may be found around the true binding site in an attempt to assist in recruiting the TFs to the correct site. To test this, the cumulative score for all sites within a sliding window of 50bp (step size 5bp) was used in the regression analysis. If a window contained a number of semi-optimal sites, the window should have a lower cumulative score than a random background window. The window with the lowest score in each promoter region was then used in the analysis.

A tool commonly used to filter out background ‘noise’ in TFBS search algorithms is phylogenetic footprinting, which operates on the assumption that the biologically important features of the genome will be conserved across species. To incorporate this methodology into our analysis, each of the previously described approaches was repeated though only allowing putative sites that were found to be evolutionarily conserved across yeast species (Table 3.1, shaded boxes; defined by PhastCons average score of 0.5 in the 7 species alignment for single sites; average PhastCons conservation score of 0.25 in the 50bp window for the ‘lowest window PSSM score’).

The results of our analysis show (Table 4.1) that the number of high-affinity sites within a promoter is much more predictive of having a site detected positive by ChIP-chip than either having a single high-affinity (low-scoring) site or having a cluster sites (*p*-value cutoff < 0.005). With only one exception, all of the factors for which the cutoff count value did not function as a significant predictor for ChIP-chip confirmed binding sites were not significant for any of the methods; the lone exception being RTG3 which was modeled accurately by the window score approach. The use of phylogenetic footprinting did assist in the predictive ability of the count method in 1 factor, CAD1. The most likely reason for the lack of additional significance of the phylogenetic footprinting approach, or taking evolutionary conservation into account, is the evolutionary distance of the species of yeast being used. It is estimated that the difference between the species may be as much as 300 million years (74); a study of TFBSs between mammals of that distance suggests that only 12-22% of the binding sites are retained (71).

Next, a qualitative assessment of TF binding motifs was compared to the ability of the motif counting approaches to model ChIP-chip data (Table 3.2). The TFs that we found for which the method was not able to model the ChIP-chip data included all of the relatively longer (10-20bp) binding motifs with very little degeneracy. Table 3.2 shows the binding motifs for the different TFs we used in this study. The inability of the longer, highly constrained motifs to be found as significant predictors may be a product from the creation of the motifs in that not enough binding sites were previously identified for truly representative PSSMs. The inability of the presence of other motifs to be a significant predictor of the ChIP-chip data may be caused by the requirement of a cofactor to be located near the binding site. In this case, the presence or absence of motifs for both cofactors may be a much better predictor.

Table 3.2 Binding motif of all yeast transcription factors modeled in this study
Factors whose name is in red are those we found to be statistically significant predictors of the ChIP-chip data using at least one of the motif counting methods (string count, cutoff A count or cutoff B count).

Transcription Factor	Motif	Transcription Factor	Motif
SUM1		GCN4	
CAD1		GLN3	
UGA3		HAP4	
ZAP1		HAP5	
BAS1		LEU3	
ACE2		MET31	
SFP1		MET32	
MCM1		MET4	
STB1		MOT3	
ROX1		PHO2	
ADR1		PUT3	
BAS1		RCS1	
CAD1		RPH1	
CBF1		RTG3	
DAL81		SFP1	
DAL82		SIP4	
FHL1		STP1	
GAT1			

3.2.2 Analysis of Human ChIP-chip Data

To confirm our findings with the yeast data that the count of putative sites is the best computational method for modeling ChIP-chip data, a series of methods was also tested on human ChIP-chip data (Table 4.3). In addition to all of the methods tested with the yeast ChIP-chip data, we also used 2 different pairwise alignment methods (DBA (75) and BlastZ (76); conservation threshold of 65%) as well as Footer, a phylogenetic footprinting algorithm optimized for human – mouse sequence comparison (28) (Table 4.3). The human dataset had no features that were statistically significant predictors of the ChIP-chip data (p-value < 0.001), though the lowest p-values were seen in the PSSM cutoff count method for 4 of the TFs, suggesting a similarity to the yeast data. The human – mouse pairwise comparison was chosen because it has previously been shown that the use of these two species in identifying TFBSs is very efficient (28, 71).

The inability of the human ChIP-chip to be modeled is likely due to the much larger size of the promoter region being searched. The raw data of a promoter-tiling array may narrow down the location of the significantly bound region and therefore provide a smaller search space to model, resulting in more accurate results. Unlike the yeast ChIP-chip modeling results, there wasn't any correlation observed between the size and degeneracy of the pattern and the p-values recorded for its predictor (data not shown).

Table 3.3 *p*-values representing the significance of the given TFBS identification method as a predictor for ChIP-chip data in humans for each TF.

Transcription Factor	Alignment Method	CREB	HNF-1	NF-Kb (p65)	HNF6	SRF	c-Jun	c-Myc	OCT.4
Footer Score		0.652	0.09	0.164	0.104	0.026	0.274	0.986	0.302
PSSM Cutoff Count 1		0.005	0.037	0.046	0.699	0.004	0.552	0.265	0.098
PSSM Cutoff Count 2		0.002	0.004	0.018	0.386	0.009	0.552	0.281	0.112
PSSM Cutoff Count 1	PhastCons	0.995	0.996	*NA*	0.995	0.995	*NA*	*NA*	0.995
PSSM Cutoff Count 1	DBA	0.31	0.076	0.287	0.582	0.015	0.203	0.49	0.168
PSSM Cutoff Count 1	BlastZ	0.037	0.5	0.101	0.931	0.172	0.614	0.367	0.374
PSSM Cutoff Count 2	PhastCons	*NA*	0.996	*NA*	0.994	0.995	*NA*	*NA*	0.995
PSSM Cutoff Count 2	DBA	0.019	0.037	0.131	0.172	0.031	0.203	0.791	0.136
PSSM Cutoff Count 2	Blastz	0.061	0.168	0.126	0.939	0.065	0.614	0.527	0.341
Lowest PSSM Score		0.111	0.005	0.047	0.507	0.08	0.025	0.644	0.007
Lowest PSSM Score	PhastConst	0.186	0.451	0.992	0.266	0.975	0.308	0.27	0.79
Lowest PSSM Score	DBA	0.101	0.086	0.363	0.873	0.089	0.488	0.444	0.151
Lowest PSSM Score	BlastZ	0.34	0.122	0.273	0.669	0.455	0.675	0.834	0.652
Lowest Window PSSM Score		0.004	0.373	0.559	0.194	0.963	0.012	0.494	0.132
Lowest Window PSSM Score	PhastCons	0.352	0.845	0.669	0.521	0.637	0.253	0.338	0.81
Lowest Window PSSM Score	DBA	0.007	0.591	0.544	0.246	0.237	0.006	0.377	0.046
Lowest Window PSSM Score	BlastZ	0.021	0.777	0.388	0.967	0.724	0.033	0.91	0.222
Consensus String Count		0.261	0.003	0.996	*NA*	*NA*	0.119	0.281	0.002
Consensus String Count	PhastCons	*NA*	*NA*	0.997	*NA*	*NA*	*NA*	*NA*	*NA*
Consensus String Count	DBA	0.273	0.011	0.795	*NA*	*NA*	0.992	0.791	0.992
Consensus String Count	BlastZ	0.667	0.254	0.305	*NA*	*NA*	0.614	0.527	0.993

Cutoff 1: 1 False Positive per 3000bp
Cutoff 2: 1 False Positive per 6000bp
Blue text indicates *p*-value <= 0.05; not statistical significance

▨ Only sites or values within conserved regions were evaluated

3.3 CONCLUSION

This study has demonstrated that the number of instances a TF binding motif is found in a genomic region can be very predictive of the region considered bound by the TF according to the ChIP-chip data. Our finding fits well with the biochemical methodology of ChIP-chip, in that a region of DNA that is bound by the factor in multiple locations is more likely to be collected and therefore have it's probe detect a stronger signal. One important fact to keep in mind with our results is that the ChIP-chip data was only collected under a certain cellular condition and therefore genes that had binding motifs in the promoter and classified as a non-true site may in fact be bound by the factor under different conditions or in different cell types. Overall, it is likely that ChIP-chip will show a more significant signal for a region containing multiple binding sites even though those sites may exhibit the same amount of function when regulating transcription as a promoter containing just a single site

3.4 MATERIALS AND METHODS

3.4.1 ChIP-chip Datasets

Yeast ChIP-chip data was collected from the Lee *et al.* study (73). A TF was considered bound if the reported *p*-value for a promoter region was less than 0.001. Lee and colleagues established this cutoff as it incorporated the maximal number of previously known interactions while reducing the number of false positives. Human ChIP-chip data was collected from Zhang *et al.* for CREB (77); Odom *et al.* for HNF-1 and HNF-6 (78); Hong *et al.* for NF-kB (79, 80); Cooper *et al.* for SRF *(81)*; Bruce *et al.* for c-JUN (80, 82); Kim *et al.* for c-Myc (80, 83); and Jin *et al.* for OCT4 (84). To identify binding a *p*-value cutoff of 0.005 was used for factors NF-kB, CREB, HNF-1, HNF-6 and SRF. The data for c-Myc and c-Jun were only provided as $\log_2$ ratios of bound to mock control, so the genes with the top 0.5% of $\log_2$ ratios were considered bound.

3.4.2 Transcription Factor Binding Motifs

Yeast transcription factor consensus sequences were collected from the *Saccharomyces* Genome Database (85) and their PSSMs were obtained from Harbison *et al.* (86). Human consensus sequences and PSSMs were obtained from the Transfac Database (45). Threshold cutoffs for the PSSMs were calculated by the generation of 100,000 random sequences based upon a 3rd order hidden markov model trained with the intergenic regions of the entire human or yeast genome (87). The PSSM score that allowed only 1 false positive site per 3,000bp or 6,000bp was then established and used in the scanning of the promoter regions.

3.4.3 Promoter Sequence Collection

Promoter sequences for human, mouse and yeast as well as the PhastCons multi-alignment scores were collected from the UCSC genome browser (49, 50). Yeast promoter regions were defined as the region between the 5' region of a gene and either the end of the nearest upstream gene or 1kb, whichever was smallest. Gene's that share a promoter region by the 5' end of one gene being within 2kb of the 5' end of a gene transcribed on the opposite strand were removed for this analysis. Human and mouse promoter regions were defined as the 3kb immediately upstream of the 5' end of the gene. For the yeast analysis, all known genes were used in the analysis whereas in the humans only those genes bound by the factor as well as an equal number of randomly selected non-bound genes were used.

3.4.4 Logistic Regression

We performed logistic regression using the R statistical package (88), `glm` function with `family` = "`binomial`". The independent variable was the value obtained from the given TFBS identification method and the ChIP-chip result was used as the binary outcome variable (1 = bound, 0 = unbound). For the Footer method (28) the score from the top-scoring site was used as the independent variable. The *p*-value for the independent variable was then presented in tables 4.1 and 4.3. Significance for the *p*-value was set at 0.05 and then adjusted with Bonferroni correction for multiple testing resulting in cutoffs of 0.005 for the yeast analysis and 0.001 for the human analysis.

4.0 FEATURES OF MAMMALIAN MICRORNA PROMOTERS EMERGE FROM POLYMERASE II CHROMATIN IMMUNOPRECIPITATION DATA

David L. Corcoran[1*], Kusum V. Pundit[1,2*], Arindam Bhattacharjee[3], Naftali Kaminski[2], Panayiotis V. Benos[4]

[1]Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA; [2]Division of Pulmonary, Allergy and Critical Care Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA; [3]Agilent Technologies Inc., Santa Clara, CA, USA; [4]Department of Computational Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA; [*]These authors contributed equally.

4.1 BACKGROUND

MicroRNAs (miRNAs) are short, ~22nt, single-stranded RNAs that act as regulators of genes' expression. By virtue of base complementarity, they bind to their target gene mRNAs and can block translation or accelerate their degradation (6). miRNAs have recently been implicated in a variety of human diseases (34, 36, 89) as well as their role discovered in particular cellular pathways (90). Mature miRNAs are generated by a cellular process in which the primary-miRNA (pri-miRNA) transcripts are processed by the enzyme *Drosha* after which the preliminary-miRNA (pre-miRNA) stem-loop structures are cleaved by the enzyme *Dicer* whereupon they become associated with a protein complex containing *Argonaute* that allows them to bind to their target (6).

Although miRNA genes play an important role in many biological processes, we still know surprisingly little about their transcription regulation. Understanding the miRNA transcription process is essential for determining which factors regulate them and subsequently, the specific role they play in signaling cascades. Accumulated evidence indicates that most miRNA genes are transcribed by RNA polymerase II (Pol II) (2, 5), although some exceptions have also been reported (3). While miRNA genes found within introns of other genes are thought to be co-transcribed with their host gene and therefore share the same promoter region (5), almost nothing is known about the promoter region of intergenic miRNA genes. A first step toward understanding intergenic miRNA regulation is to identify their transcription start sites (TSSs). Currently, only a small number of intergenic pri-miRNA transcripts have confirmed TSSs (2, 10, 91), which is insufficient for studying the promoter sequence features and comparing them to those of protein coding genes. Due to this lack of information, all studies attempting to analyze the miRNA core promoters have focused on the area immediately upstream of the computational prediction of the pri-miRNA (13, 92, 93).

To create the number of miRNA TSSs needed for feature comparison, we performed a high-throughput chromatin immunoprecitiation (ChIP)-chip experiment that can identify the areas upstream of the pri-miRNAs bound by Pol II. A model is then needed that is capable of distinguishing random sequences from core promoters and able to identify which sequence features are the most important for that classification. Existing algorithms for modeling Pol II core promoters vary, both methodologically and in terms of performance. One of the first approaches developed made use of the comparison of transcription factor binding site frequency between true core promoters and random intergenic sequences (94). Another method for modeling core promoters took into account the size and location of CpG islands as well as their distance to known TSSs (95). A variety of other methods have also been developed that take into account physical properties of the DNA and analyze them with neural networks (96), relevance vector machines (97), and additive logistic regression with boosting (98).

In this chapter, we develop a new highly efficient model for Pol II core promoter identification based upon SVM literature that when used with the results from the analysis of the high-throughput Pol II ChIP-chip data sheds light on how intergenic and intronic miRNA genes are transcribed and how the features in miRNA core promoters compare to those of protein coding genes.

4.2 RESULTS AND DISCUSSION

4.2.1 Identification of pri-miRNA TSSs From Pol II ChIP-chip Data

To identify the true TSS for pri-miRNAs we performed a Pol II ChIP-chip on A549 lung cells, as described in *Materials and Methods*. Statistical analysis (99) was used to identify 1000bp regions that exhibit high Pol II signals. The window nearest to the 5' end of each of the 531 known pre-miRNAs was recorded as containing the putative TSS, if it was closer than 50kb. The array that was designed for these experiments only included up to 50kb upstream of known miRNA genes (see *Materials and Methods*); this threshold is also based on previous studies that showed high correlation of expression between miRNA genes located within 50kb of each other (4). This method resulted in 35 intergenic pre-miRNAs or polycistronic pri-miRNAs having a statistically significant Pol

II signal associated with them (Table 4.1). Regions with significant Pol II signals that overlapped the 5' end of a gene (as identified by the UCSC table browser (49, 50)) were excluded from subsequent analysis. This was necessary because the ChIP-chip data cannot distinguish the shared core promoter regions. An example of the distribution of the Pol II binding signals is presented for the identification of the pri-miR-10a TSS (Figure 4.1)

The miR-23a~miR-27a~miR24-2 cluster is probably the best-studied human pri-miRNA transcript. Lee *et al.* (2) have shown that its TSS is located 124 nucleotides upstream of miR-23a, which our ChIP-chip data confirmed. Our ChIP-chip data also confirmed the previously reported pri-miRNA TSSs listed in Fujita and Iba (11) for miR-21 and miR-10b genes (Table 4.1). The distance between the Pol II peak and the beginning of the pre-miRNA varies significantly between genes, from a minimum of zero to a maximum of 40 kb. The average and median values are 10.5 kb and 6.8 kb, respectively.

The analysis of our Pol II ChIP-chip data also provided the location of TSSs for intronic miRNAs, currently thought to be transcribed along with their host gene (5). The nearest upstream Pol II ChIP-chip significant regions for many of the intronic miRNAs overlapped the 5' region of their host gene (Table 4.3) confirming the previous findings for miR-146a (12) and the miR-17~miR-18a~miR-19a~miR-20a~miR-19b-1~miR-92a-1 cluster (10). Interestingly, our analysis found that some of the intronic miRNA genes may be transcribed by their own promoter (Table 4.2). The distance between the Pol II peak and the beginning of the (intronic) pre-miRNA gene varies between 200bp and 41 kb, but with more peaks observed at longer distances (avg=18 kb; median=19 kb). As will be described in section 4.2.6, some of these ChIP-chip peaks were found to contain core promoter features as identified by our newly developed model for TSS identification, described in the next few sections.

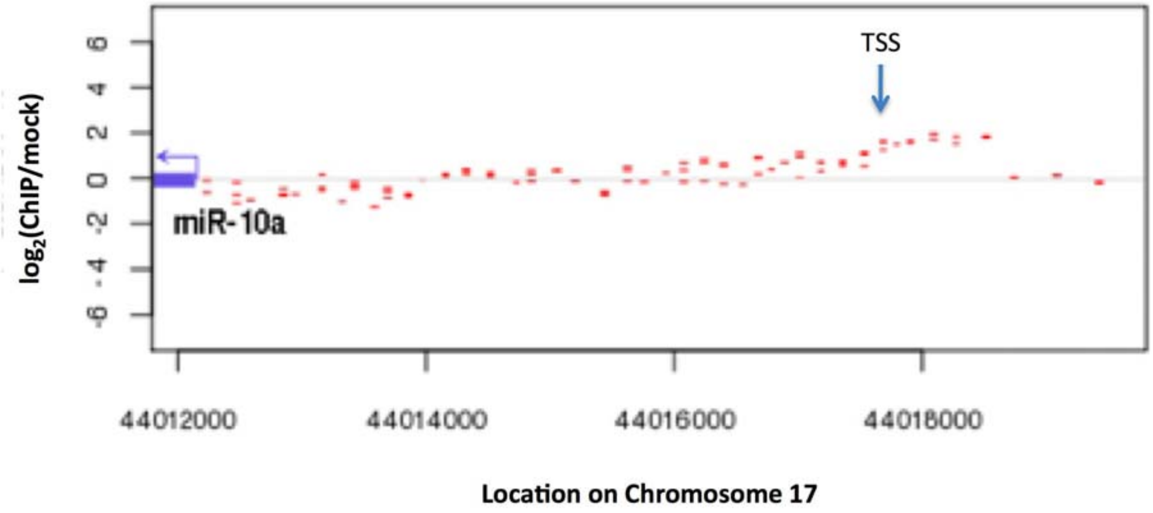


Figure 4.1 Pol II ChIP-chip results for miR-10a
The blue arrow represents the location and transcriptional direction of hsa-miR-10a. The red dashes represent the location and value of the ChIP-chip probes. The labeled TSS mark is the 500bp region identified by our model as the core promoter and transcription start site of hsa-miR-10a.

Table 4.1 Identification of promoters of *intergenic* miRNA genes

Cells in light green designate the previously verified TSSs. miRNA: miRNA gene symbol, multiple symbols designate cluster of co-expressed miRNAs; Chromosomal location: the chromosomal position and orientation of the miRNA gene; ChIP-chip region: the region with a statistically significant peak; CPPP model: CpG (CpG+) or non-CpG (CpG-) model used for the TSS prediction; Predicted TSS: TSS predicted by our SVM model; Distance: the distance of the predicted TSS from the most 5' pre-miRNA transcript.

miRNA	Chromosomal location	ChIP-chip region	CPPP Model	Predicted TSS (CPPP)	Distance
miR-200b~miR-200a~miR-429	Chr1: 1092346 (+)	[1082033, 1083782]	CpG+	1083333	8763
miR-34a	Chr1: 9134423 (-)	[9162283, 9166532]	CpG+	9165233	30560
miR-101-1	Chr1: 65296779 (-)	[65304283, 65307532]	CpG+	65306008	8979
miR-181a-1~miR-181b-1	Chr1: 197094905 (-)	[197125783, 197127032]	**CpG-**	***NA***	***NA***
miR-202	Chr10: 134911115 (-)	[134919994, 134925743]	CpG+	134925294	13929
miR-210	Chr11: 558198 (-)	[559355, 560354]	CpG+	558988	540
miR-194-2~miR-192	Chr11: 64415487 (-)	[64416605, 64418104]	CpG-	64416930	1193
miR-200c~miR-141	Chr12: 6943122 (+)	[6940546, 6942545]	CpG+	6941071	1801
let-7i	Chr12: 61283732 (+)	[61279796, 61291045]	CpG+	33705771	461
miR-379~miR411~…~miR-410~miR-656	Chr14: 100558155 (+)	[100524119, 100525868]	**CpG-**	***NA***	***NA***
miR-193b	Chr16: 14305324 (+)	[14302031, 14310280]	CpG+	14304281	793
miR-138-2	Chr16: 55449930 (+)	[55439531, 55441030]	CpG-	55439856	9824
miR-497~miR-195	Chr17: 6862065 (-)	[6863309, 6865058]	CpG-	6864759	2444
miR-10a	Chr17: 44012308 (-)	[44017059, 44018808]	CpG+	44017609	5051
miR-196a-1	Chr17: 44064920 (-)	[44078809, 44080558]	CpG+	44079134	13964
miR-21	Chr17: 55273408 (+)	[55267309, 55276558]	CpG-	55271984	1174
miR-122	Chr18: 54269285 (+)	[54235566, 54236565]	CpG-	54235891	33144
miR-23a~miR-27a~miR-24-2	Chr19: 13808473 (-)	[13807348, 13809097]	CpG-	13808448	0
miR-181c~miR-181d	Chr19: 13846512 (+)	[13832848, 13834847]	**CpG-**	***NA***	***NA***
miR-99b~let-7e~miR125a	Chr19: 56887676 (+)	[56882098, 56886347]	CpG+	56884421	3005
miR-216a~miR-217	Chr2: 56069698 (-)	[56072783, 56074282]	CpG-	56073933	3985
miR-10b	Chr2: 176723276 (+)	[176705033, 176707032]	CpG+	176705608	17418
miR-301b~miR-130b	Chr22: 20337269 (+)	[20335283, 20337282]	CpG+	20336383	636
let-7a-3~let-7b	Chr22: 44887292 (+)	[44879283, 44883032]	CpG+	44881933	5109
miR-206~miR-133b	Chr6: 52117105 (+)	[52096878, 52098877]	CpG-	52098453	18402
miR-30a	Chr6: 72170045 (-)	[72164628, 72176377]	CpG-	72174203	3908
miR-129-1	Chr7: 127635160 (+)	[127593752, 127595501]	CpG+	70094852	40058
miR-183~miR-96~miR-182	Chr7: 129202090 (-)	[129206752, 129207751]	CpG+	129207227	4887

Table 4.1 (*continued*)

miR-29b-1~miR-29a	Chr7: 130212838 (-)	[130219002, 130223501]	CpG-	130223027	9939
miR-30d~miR-30b	Chr8: 135886370 (-)	[135913283, 135915782]	CpG+	135913608	26988
let-7a-1~let-7f-1~let-7d	Chr9: 95978059 (+)	[95966631, 95971380]	CpG+	95968506	9303
miR-181a-2~miR-181b-2	Chr9: 126494541 (+)	[126459631, 126464380]	CpG-	126460831	33460
miR-222~miR-221	ChrX: 45491474 (-)	[45504862, 45507861]	CpG-	89506287	14563
miR-542~miR-450a-2~miR-450a-1~miR-450b	ChrX: 133503133 (-)	[133502362, 133506611]	CpG+	133506287	2904
miR-505	ChrX: 138834056 (-)	[138842362, 138844111]	CpG+	76342712	8406

Table 4.2 Identification of promoters for *intronic* miRNA genes

Cells in light green designate genes whose expression was found to be anti-correlated with their host genes. Column names as in Table 1. Host gene: the gene whose intron the miRNA was found.

miRNA	Host Gene	Chromosomal location	ChIP-chip region	CPPP Model	Predicted TSS (CPPP)	Distance
miR-107	PANK1	Chr10: 91342564 (-)	[91382494, 91383493]	CpG-	91382844	40030
let-7a-2~ miR-100	AK091713	Chr11: 121522511 (-)	[121521855, 121523854]	CpG-	***NA***	***NA***
miR-190	TLN2	Chr15: 60903208 (+)	[60860703, 60861952]	CpG-	60861428	41530
miR-99a~ let-7c	C21orf34	Chr21: 16833279 (+)	[16826951, 16832700]	CpG-	16827826	5203
miR-125b-2	C21orf34	Chr21: 16884427 (+)	[16880451, 16883950]	CpG-	16880951	3226
miR-26a-1	CTDSPL	Chr3: 37985898 (+)	[37961854, 37963353]	CpG-	37962529	23119
miR-196b	HOXA9	Chr7: 27175707 (-)	[27178752, 27180251]	CpG+	27180227	3770
miR-489~ miR-653	CALCR	Chr7: 92951267 (-)	[92953002, 92954251]	CpG-	***NA***	***NA***
miR-101-2	RCL1	Chr9: 4840296 (+)	[4827381, 4828630]	CpG-	4828281	11765
miR-491	KIAA1797	Chr9: 20706103 (+)	[20673131, 20677880]	CpG+	20674256	31597
miR-204	TRPM3	Chr9: 72614820 (-)	[72633881, 72634880]	CpG-	***NA***	***NA***
miR-7-1	HNRPK	Chr9: 85774592 (-)	[85774131, 85775630]	CpG-	85775081	239
mir-23b~ miR-27b~ miR-24-1	C9orf3	Chr9: 96887310 (+)	[96846381, 96860880]	CpG+	96850731	36329
miR-32	C9orf5	Chr9: 1108483999 (-)	[110866881, 110868380]	CpG-	110867881	19232
miR-448	HTR2C	ChrX: 113964272 (+)	[113955612, 113956861]	CpG-	***NA***	***NA***

Table 4.3 Intronic miRNAs whose nearest ChIP-chip peak overlaps host gene TSS
Cells in light green designate genes that have been found to be co-transcribed with their host genes. Column names as in Table 2.

miRNA	Host Gene	Chromosomal location	ChIP-chip region
miR-30e~miR30c-1	NFYC	Chr1: 40992613 (+)	[40946783, 40950532]
miR-186	ZRANB2	Chr1: 71305987 (-)	[71316783, 71320532]
miR-130a	AK096335	Chr11: 57165246 (+)	[57161605, 57163604]
miR-148b	COPZ1	Chr12 53017266 (+)	[53004046, 53006295]
miR-26a-2	CTDSP2	Chr12: 56504742 (-)	[56524546, 56528295]
miR-15a~miR-16-1	DLEU2	Chr13: 49521338 (-)	[49551648, 49555397]
miR-17~miR-18a~miR-19a~miR-20a~miR-19b-1~miR-92a-1	C13orf25 v_1	Chr13: 90800859 (+)	[90798648, 90800647]
miR-423	CCDC55	Chr17: 25468222 (+)	[25467059, 25470058]
miR-301a~miR-454	FAM33A	Chr17: 54583364 (-)	[54583809, 54589308]
miR-330	EML2	Chr19: 50834185 (-)	[50833598, 50834597]
miR-26b	CTDSP1	Chr2: 218975612 (+)	[218968033, 218974282]
miR-103-2	PANK2	Chr20: 3846140 (+)	[3816001, 3820000]
miR-185	C22orf25	Chr22: 18400661 (+)	[18387533, 18389782]
miR-191~miR-425	DALRD3	Chr3: 49033146 (-)	[49026104, 49038353]
miR-15b~miR-16-2	SMC4	Chr3: 161605069 (+)	[161598354, 161603353]
miR-378	PPARGC1B	Chr5: 149092580 (+)	[149089935, 149091684]
miR-103-1	PANK3	Chr5: 167920556 (-)	[167938685, 167940184]
miR-335	MEST	Chr7: 129923187 (+)	[129912502, 129914001]
miR-31	LOC554202	Chr9: 21502184 (-)	[21539381, 21557130]
miR-421	AK125301	ChrX: 73355021 (-)	[73377862, 73379611]
miR-374b~miR-374a~miR-545	AK057701	ChrX: 73355178 (-)	[73421362, 73431611]
miR-361	CHM	ChrX: 85045368 (-)	[85188362, 85189861]
miR-503	MGC16121	ChrX: 133508094 (-)	[133506612, 133515611]
miR-452~miR-224	GABRE	ChrX: 150878840 (-)	[150889112, 150894611]
miR-22	MGC14376	Chr17: 1564031 (-)	[1563059, 1569558]
miR-636	SFRS2	Chr17: 72244225 (-)	[72244059, 72246308]
miR-146a	DQ658414	Chr5: 159844936 (+)	[159826435, 159828934]

4.2.2 Modeling Pol II Core Promoter Features With *n*-mers and Weight Matrices

In the following, we describe the development of a novel SVM-based method for prediction of Pol II TSSs. This model was used for the identification of the miRNA TSSs from the ChIP-chip data and most importantly for comparing the features of the miRNA core promoters to those of protein coding gene promoters

It is known that the genomic regions immediately upstream of the TSS of protein coding genes exhibit high levels of sequence conservation (24, 100-102), which is probably related to the high concentration of *cis*-regulatory sites in this region (103). All of the existing algorithms for modeling Pol II core promoters have used this property to different extent. Generally one can model DNA target sites using either *n*-mer frequencies or weight matrices, commonly known as position-specific weight matrices (PSSMs) (23). The first class of methods (also termed enumerating or dictionary-based methods; e.g., (104-107)) is better suited for representation of the binding preferences of those transcription factors that have a restricted set of DNA targets. *n*-mer frequencies have been used in the past to model Pol II core promoters either alone (13) or in conjunction with some promoter entropy measure (108). The DNA targets of most transcription factors are not highly conserved, which is the reason why PSSM models are widely used for representing DNA motifs. However, there are also problems in using PSSMs for Pol II core promoter recognition. First, the currently known DNA motifs are redundant. As a matter of fact, it is known that structurally similar transcription factors recognize similar "core" motifs (24, 37). Second, only a small percent of all transcription factors have known binding preferences; TRANSFAC database (45) currently contains 601 models for 2,113 known mammalian transcription factors. Third, even if the binding preference of a given transcription factor is known, the task of determining whether it binds to a given promoter is not trivial, mainly due to the high false positive prediction rate (28, 109). Regardless, in the past PSSM models have been used for Pol II core promoter identification (92).

The problems of PSSM redundancy and the high number of transcription factors with yet unknown binding preferences can be diminished if one uses generalized profiles or familial binding profiles (FBPs) (37). FBPs represent an "average" of the binding affinities of transcription factors with similar DNA binding preferences. They are based on the fact that transcription factors of the same structural group usually bind to similar sets of sequences. This not only reduces the PSSM redundancy, but since the currently unknown transcription factors will likely belong to one of the known structural groups, it is very likely that their binding preferences will be represented in one of the FBPs. Sandelin and Wasserman initially built a set of 11 FBPs using a semi-manual method (25). The zinc finger proteins were excluded from these FBPs due to their high degree of target promiscuity, which in turn makes them difficult to cluster correctly. More recently, Mahony *et al.* (24) used an automatic method to construct 17 FBPs. This set of FBPs includes all but the C2H2 the zinc finger (sub)family, though for the purposes of this study the same method was applied to the C2H2 factors resulting in 31 additional FBPs.

4.2.3 Evaluating Core Promoter Features Using Support Vector Machines

In order to better understand how various features contribute in the characterization of Pol II core promoters we compared them under the support vector machine (SVM) framework (110, 111). The SVM methodology was chosen because it can combine multiple types of evidence (features) under the same general framework. In particular we tested both the n-mer frequencies ($n=3,4$) and matches to a set of generalized DNA binding profiles alone or in conjunction with GC content, which seems to be a prominent feature in a subset of eukaryotic promoters (112). A Previous study has shown that SVMs can model core promoters with hither sensitivity than other methods (108). While our approach is based upon that of Gangal and Sharma (108), there are a few distinct differences including the background training data set, length of the core promoter region and the use of the presence of absence of CpG islands (see below).

In the course of this study, five SVM models were constructed and compared: **(1)** FBPs only (49 features), **(2)** *n*-mers only (*n=3,4*) (320 features), **(3)** FBPs+GC content, **(4)** *n*-mers+GC content, and **(5)** FBPs+*n*-mers+GC content. All models were trained on the same set of 3,015 verified core promoters of protein coding genes and 3,015 randomly chosen intergenic sequences (positive and negative examples, respectively; see *Materials and Methods*). Performance was measured by a 20X cross-validation in which 75% of the examples in each dataset were used for training and the remaining 25% for testing. Our results, presented in Figure 4.2, indicate that the *n*-mer-based models perform generally better than FBP-based models, both in terms of sensitivity (percent of correctly predicted positive examples) and specificity (percent of true positive examples among all predictions). For example, the "*n*-mer only" SVM model (*n=3, 4)* exhibited $S_N=74.3\%$ and $S_P=86.1\%$ compared to $S_N=70.8\%$ and $S_P=82.2\%$ of the "FBP only" model. However, none of these differences is statistically different, so one may choose to use FBPs for such type of modeling, especially when the training examples are limited.

All SVM models above were based on the dot plot kernel function (linear discriminator). Tests with polynomial (3rd order) and radial kernels gave the same or slightly worse results (*data not shown*). Also, all SVM models were constructed using random intergenic regions as background (*see Materials and Methods*) instead of the intronic regions previous studies have used (108). Therefore, the same evaluation with intronic background was performed but found to be slightly worse (*data not shown*).

We note that some previous studies have reported better performance in some cases of predicting Pol II promoters (13, 108). We believe this is due to the smaller size of the datasets they used and the type of promoters they contain. For example, Gangal and

Sharma (108) reported S_N*>87%* and S_P*>86%* but their dataset consisted of 800 promoter sequences all taken from EPD (113) in which about 83% of the promoters contain CpG islands. A very powerful separation hyperplane can be created by using these GC-rich promoters as positive set and intronic sequences, generally AT-rich, as negative set; however this model is expected to perform poorly on non-CpG island promoters. In our case, only half of the promoters in the training/testing dataset used contain CpG islands. When the EPD dataset is used for training/testing in this study, the results were similar (intronic background) or slightly better (intergenic background) to those reported in Gangal and Sharma (108). Nevertheless, as will be shown next, partition of the promoters to those containing CpG islands and those lacking them improve the results substantially.

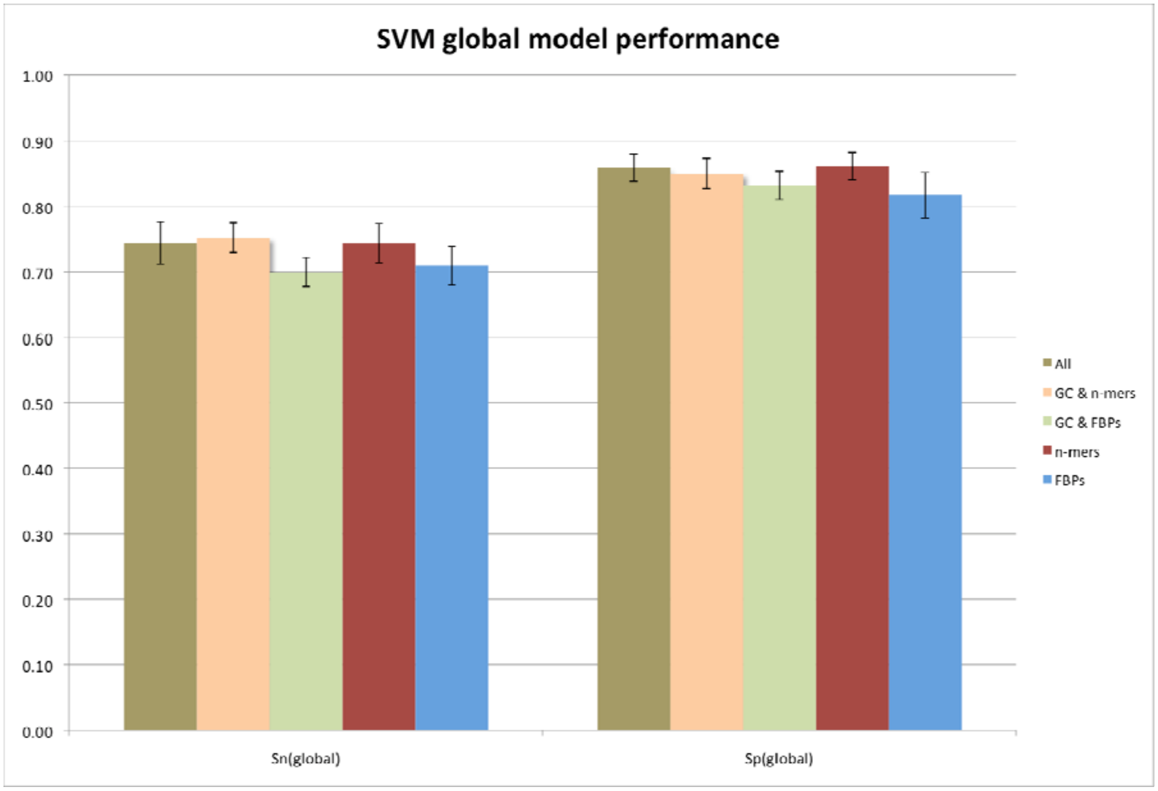


Figure 4.2 Performance of the *n*-mers and FBPs (alone and in combination) in predicting Pol II core promoter regions.

4.2.4 The Effect of the Presence or Absence of CpG Islands in the Prediction Efficiency of Pol II Core Promoters

In general, the frequency of CG dinucleotides in vertebrate genomes is lower than expected by chance (114). This is probably due to the frequent conversion of methylated-CG into TG (115). Methylation of genomic regions is a method by which the cell can repress transcription activity (116, 117). Typically, promoters of genes containing a large frequency of CG dinucleotides in long stretches (CpG islands) are found to be expressed across a variety of cell types thus reducing the likelihood of that genomic region being methylated (118). Another feature of CpG island containing promoters is that they are strongly associated with having a TSS that spans a short window instead of having 1 specific base in which transcription always initiates (118). The functional difference in CpG island containing promoters suggests that they should be modeled separately as been seen in the study by Ioshikhes and Zhang (95). The prediction efficiency of the "*n*-mer only" and "FBP only" SVM models in predicting mammalian core promoters in the presence of absence of CpG islands was then tested. Focus was placed on these two models because they are simpler (fewer features) and they perform the same or better than the other models.

The training set was then partitioned into CpG containing promoters (CpG+) and non-CpG promoters (CpG-) whereupon the two *n*-mer-based and two FBP-based SVM models were calculated. The negative dataset in all cases contained equal number of randomly selected sequences from the intergenic parts of the genome (see *Materials and Methods*). The results demonstrate that the prediction efficiency differs significantly between the two types of promoters. In particular, the "*n*-mer only" model trained on

CpG island containing promoters exhibits S_N=94.8% (sd=1.1%) and S_P=97.6% (sd=1%) in the cross-validation tests. By contrast, the "*n*-mer only" model trained on non-CpG island promoters performs worse on the self-validation (S_N=73.4%, sd=2.6% and S_P=73.2%, sd=2.9%) (Figure 4.3). The "FBP only" model gave similar results (Figure 4.3). Also, it was found that both *n*-mer-based models outperform the corresponding FBP-based models (Figure 4.3). Furthermore, the results show that *n*-mer-based models trained on CpG+ promoters tend to predict extremely well the CpG+ promoters (S_N=*94.8%,* S_P=*97.6%*), which agrees with previous reports (95). Interestingly, prediction of non-CpG promoters is significantly worse, even with a model trained on non-CpG promoters (S_N=*73.4%,* S_P=*73.2%*). This finding likely reflects on the lack of strong general features in non-CpG promoters. Still, using the two separately trained models to predict the CpG and non-CpG promoters individually yields significantly better results than a single model trained on all promoters.

The program '`gist-fselect`' from the *Gist* package (110) was used to evaluate the significance of each of the *n*-mer features of core promoter regions (*t*-test metric *p*-value was used to determine significance, Fisher score metric was used to rank features) in CpG and non-CpG promoters (Table 4.4). As expected, the most significant features for the CpG model were CG containing *n*-mers. We also note that there is no feature overlap between the two sets of promoters in the top 20 features listed in Table 4.4.

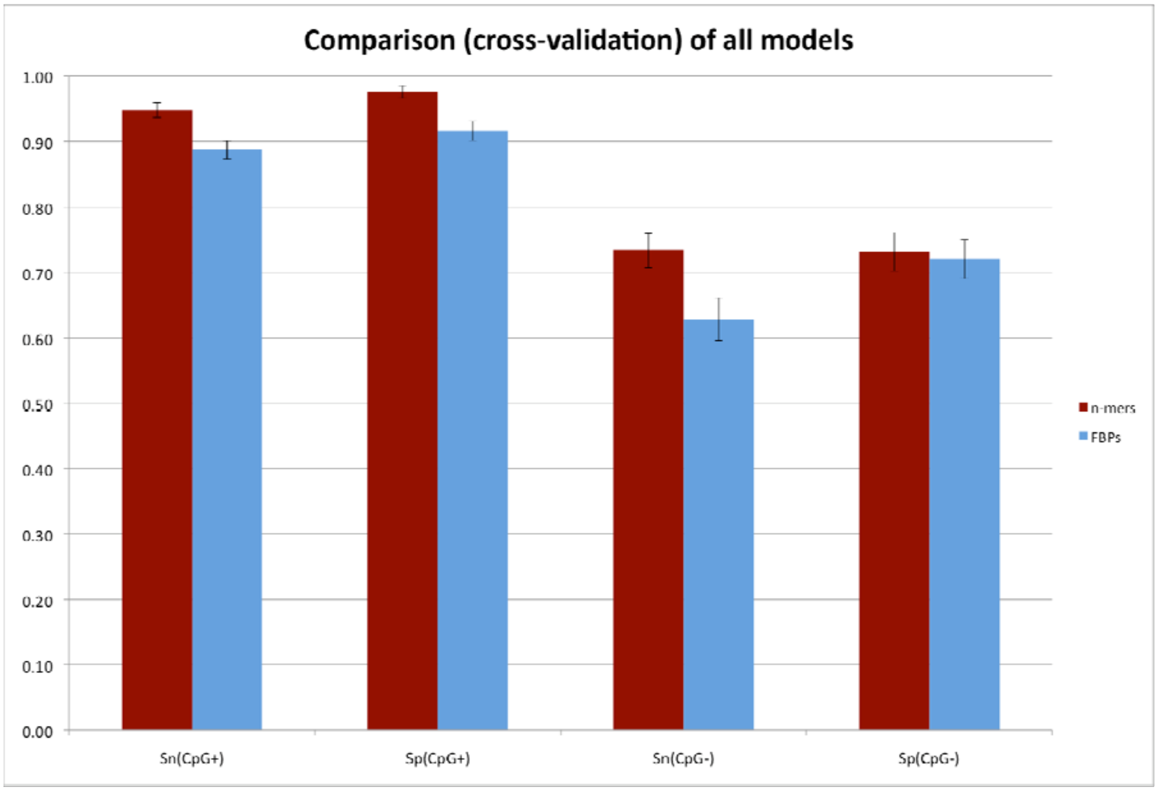


Figure 4.3 Performance of our SVM models in predicting CpG and non-CpG core promoters

4.2.5 Comparison of Core Promoters for Protein Coding Genes and miRNA with SVM Models

The ChIP-chip data showed that 35 of the intergenic miRNA genes had significant Pol II signals nearby. We used our SVM based model trained on protein coding genes to identify the region within the ChIP-chip significant peak that contains strong core promoter region features and therefore the likely TSSs of these miRNA genes. Our model used two SVMS trained separately on CpG and non-CpG protein coding promoters. For the training and scoring, we used the dot kernel function with intergenic sequences as background; the feature space was comprised of *n*-mer frequencies (*n*=3,4). For identifying the TSSs of the intergenic miRNA genes with Pol II peaks, the 3kb regions surrounding the windows with the most significant Pol II peak were collected and the presence or absence of CpG islands was determined using the same method as in Ioshikhes and Zhang (95). The corresponding SVM model was then used across the significant ChIP-chip region to identify a 500bp window classified as a core promoter by our model. The model was able to identify a TSS in the upstream region of 32 out of the 35 intergenic miRNA genes (Table 4.1). The three intergenic miRNAs for which our model was unable to identify a true core promoter region did each contain a 500bp region that scored just below the threshold cutoff for identifying a core promoter from a background sequence (*data not shown*).

The number of Pol II associated intergenic miRNA genes is not large enough to retrain the SVM models and calculate significant sequence features. However, we can test whether the most significant features in the promoters of the protein coding genes (Table 4.4) are also overrepresented in the miRNA promoters. Comparison of all *n*-mer frequencies of the CpG promoters of protein coding genes with those of the miRNA genes resulted in a statistically significant (Wilcoxon signed-rank test; *p*-value < 0.05 after Bonferroni correction) difference in 5 *n*-mers ('CCGG', 'ATA', 'GGG', 'CAA', and 'GGGG'). However, the only 4-mer in the list of the top 50 most important features for the model as identified by '`gist-fselect`' was 'CCGG'. For the non-CpG promoters, we found no features with a statistically different frequency in the protein coding and miRNA core promoters.

Table 4.4 The top 20 most significant *n*-mers for each of the two models and the Fisher score as well as the –log10 of the *t*-test metric p-value as determined by the *Gist* package

non-CpG			**CpG**		
Feature	-$\log_{10}$(p-value)	Fisher Score	Feature	-$\log_{10}$(p-value)	Fisher Score
CCCT	29.7925	0.152704	GCG	4.34E+09	2.90813
AGGG	26.8574	0.136658	CGG	4.34E+09	2.70749
GCCC	23.9996	0.12122	CGC	4.34E+09	2.67387
CCC	23.6638	0.119395	CCG	4.34E+09	2.41605
TGTA	23.7021	0.119389	TCG	4.34E+09	1.79033
CCCC	23.6248	0.119181	GCGC	4.34E+09	1.74254
AAT	23.4104	0.117827	CCGG	4.34E+09	1.61845
GAAG	22.2979	0.111908	CGGC	4.34E+09	1.6182
AGC	21.1428	0.105734	GGCG	4.34E+09	1.55504
TAC	20.8754	0.104254	GCGG	4.34E+09	1.55492
ATT	19.5344	0.0971108	CGA	4.34E+09	1.54674
TAAT	19.1561	0.0950535	ACG	4.34E+09	1.48203
ATTA	19.1021	0.0947959	CGCC	4.34E+09	1.44506
TACA	19.0021	0.0942557	CGCG	4.34E+09	1.44053
GTA	18.6868	0.0925992	CGT	4.34E+09	1.422
AATA	18.6051	0.092075	CCGT	4.34E+09	1.42109
GGG	18.0089	0.0890836	CGCT	4.34E+09	1.36182
CTGC	18.0034	0.0890473	GCCG	4.34E+09	1.31601
CAGC	16.8798	0.0831072	TCCG	4.34E+09	1.30007
CTG	16.3289	0.0801748	AGCG	4.34E+09	1.26542

4.2.6 Computational Analysis of Potential Promoters of Intronic Genes

Intronic miRNA genes are generally believed to be transcribed by their host genes (5). However, the ChIP-chip data indicated that there might be autonomous transcription for 15 of them (Table 4.2). The ChIP identified regions were scanned with the CpG+ and CpG- SVM models and it was found that 11 of the 15 intronic genes contained a significant region, offering additional evidence that these miRNAs may have their own internal promoter. One of the genes that is predict to have its own promoter is miR-32, which has been previously shown to have a negative correlation with its host gene, *C9orf5* (4). This is an important and interesting finding about the transcriptional regulation of miRNAs, although further biochemical validation is required.

4.3 CONCLUSION

The prediction of miRNA TSSs and the understanding of the mechanisms that play a role in their transcription is an important step towards deciphering their role in regulatory networks. In this study, high-throughput Pol II ChIP-chip data was collected for first time and used to infer the actively transcribed miRNA genes in lung cells. Analysis of these data showed that the miRNA TSSs can be located as far as 40 kb from the pre-miRNA genes, indicating that pri-miRNA transcripts might be much longer than originally thought (13, 92, 93, 95)

In addition, we compared two types of features that are commonly used for the identification of Pol II core promoters: *n*-mer frequencies and PSSM models. We used the SVM framework to compare these types of features. We found the *n*-mer frequencies to be generally better than the generalized PSSM models, but at the cost of additional parameters. Also, in agreement with other studies (95), we found that promoters with CpG islands are much easier to predict than those without and that core promoter prediction is more efficient when two models are used (CpG+ and CpG- trained models). Using the best performing SVM model for core promoter prediction on our ChIP-chip data, we found that miRNA Pol II core promoters have the same features as those of the protein coding genes. Finally, the ChIP-chip data and the SVM predictions together indicate that a number of intronic miRNA genes that may be transcribed by their own promoter, and that these promoters can be located as far as 40 kb away. Taken together, these observations suggest that the transcription of miRNA genes is more complicated than initially thought.

A tool for large-scale searches for core promoter regions based upon the SVM is currently under development. Further use of the miRNA promoter array in multiple cell types should allow for the identification of all pri-miRNA transcription start sites as well as verification of those found in this study.

4.4 MATERIALS AND METHODS

4.4.1 Chromatin Immunoprecipitation (ChIP-chip)

Approximately 10^8 A549 cells (American Type Culture Collection, Manassas, VA) were grown in F12K medium (Invitrogen, Carlsbad, CA) with 2 mM L-glutamine and 10% fetal bovine serum. Cells were incubated at 37 °C in a humidified chamber supplemented with 5% CO_2. Once 80% confluent, cells were serum starved overnight. Proteins were cross-linked to the DNA using fresh formaldehyde solution (50 mM Hepes-KOH pH 7.5, 100 mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 11% Formaldehyde) for 10 min at room temperature. The formaldehyde was quenched with 2.5 M glycine for 5 min at room temperature. Cells were washed twice in PBS and harvested using a silicone scraper. Cells were centrifuged at 1,350 x g for 5 minutes at 4°C and the pellet washed twice with PBS. The pellet was resuspended in 5 ml of lysis buffer 1 (50 mM Hepes-

KOH pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100) and rocked at 4°C for 10 min. The cells were centrifuged at 1,350 x g for 5 minutes at 4°C and the pellet resuspended in 5 ml of lysis buffer 2 (10 mM Tris-HCl, pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA), rocked at room temperature for 10 min. The nuclei were pelleted by centrifuging at 1,350 x g for 5 minutes at 4°C. The pellet was resuspended in 5 ml of lysis buffer 3 (10mM Tris-HCl, pH 8.0, 100mM NaCl, 1mM EDTA, 0.5mM EGTA, 0.1% Na-deoxycholate, 0.5% N-lauroylsarcosine). The cells were sonicated for 7 cycles of 30 seconds ON and 60 seconds OFF at a power 7 using a sonic dismembrator Model 100 (Fisher Scientific, Waltham, MA). The cells were centrifuged at 20,000 x g for 10 minutes at 4°C and 50µl of the supernatant was set aside as the whole cell extract (WCE). The rest of the supernatant was incubated overnight at 4°C with 100 µl of Dynal Protein G magnetic beads that had been pre-incubated with either 10 µg RNA polymerase II antibody (Abcam, Cambridge, MA) or 10 µg E2F-4 antibody (Santa Cruz Biotechnology, Santa Cruz, CA). The beads were washed 7 times in RIPA buffer (50 mM Hepes-KOH pH 7.6, 500 mM LiCl, 1 mM EDTA pH 8.0, 1% NP-40, 0.7% Na-deoxycholate) and once in Tris-EDTA containing 50 mM NaCl. Elution was done in elution buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA pH 8.0, 1% SDS) for 15 min at 65°C. Reversal of crosslinks of the immunoprecipitate (IP) and the WCE was done at 65°C overnight. Cellular RNA was digested with 0.2 mg/ml RNaseA (Invitrogen) at 37°C for 2 h followed by protein digestion with 0.2 mg/ml proteinase K (Invitrogen) at 55°C for 30min. The DNA was purified by phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation. Purified DNA was blunted using T4 DNA polymerase (New England Biolabs, Ipswich, MA) and ligated to 2 µM linkers using T4 DNA ligase (New

England Biolabs). The IP and the WCE was amplified in two stages of PCR and purified by phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation). 2 µg each of IP and WCE was labeled with Cy5-dUTP and Cy3-dUTP (Perkin Elmer, Waltham, MA) respectively. Labeling was carried out by random-primed Klenow-based extension using the CGH Labeling kit (Invitrogen). The samples were cleaned up using Invitrogen's CGH columns included in the kit. 5µg each of IP and WCE were combined with cot-1 DNA and the 10x blocking agent and 2x hybridization buffer supplied in the Agilent Oligo aCGH/ChIP-on-chip Hybridization Kit (Agilent, Santa Clara, CA). Hybridization was carried out in Agilent's SureHyb chambers at 65°C for 40 h in the DNA Microarray Hybridization Oven (Agilent). The slides were washed using Oligo aCGH/ChIP-on-chip wash buffer 1 and 2 (Agilent) and scanned in the DNA microarray scanner (Agilent). The scanned images were processed using Agilent's Feature Extraction software version 9.5.3

4.4.2 miRNA Location Array Design

The miRNA location array was custom-made by Agilent with AMADID (Agilent Microarray Design Identifier) 014119. The array is available on the 44K design. The probes are 60mers and T_m balanced. The probe design tiles 50kb (~100 bp spacing) or 100 kb (~200 bp spacing) regions surrounding each miRNA.

4.4.3 Analysis of ChIP-chip Data

Median normalization of the $\log_2$ values of the ratio of signal to mock was performed across the three-ChIP-chip arrays followed by a mean centralization to 0. Regions of Pol II binding were identified by the ChIPOTle sliding window method (99); a window size of 1kb was used with a step size of 250bp. The window was reported as significant if the *p*-value was below 0.05 after adjustment by the conservative Bonferroni correction method for multiple testing; overlapping statistically significant windows were combined.

4.4.4 Gene Coordinate and Sequence Collection

Pol II core promoters were extracted from two databases: Eukaryotic Promoter Database (113) and DBTSS (119). Between the two databases there were 3,015 unique human TSSs (1,744 from Eukaryotic Promoter Database and 1,271 from DBTSS as originally collected by Zhao *et al* (13)). The core promoter regions were partitioned into 1,445 that contained CpG islands and 1,570 that did not according to the method and threshold used in Zhao *et al* (13). For the training and testing of the various SVM models the area [-450, +50] surrounding the TSS was used as the positive dataset. An equal number of 500bp genomic sequences, randomly selected from the intergenic regions of all chromosomes were used as the negative dataset. Special care was given so that the randomly selected regions were not located within 3kb from the 5' end of any annotated gene.

We collected all genomic coordinates for the mRNA TSSs, mRNA introns and pre-miRNAs from the UCSC table browser (49, 50). Intragenic miRNAs were identified as those found within an intron, exon or UTR of a mRNA and transcribed in the same orientation. All other miRNAs were labeled as intergenic.

5.0 IDENTIFICATION OF FEED-FORWARD LOOPS INVOLVING MICRORNA GENES IN TGFβ SIGNALLING

David L. Corcoran[1], Hanadie Yousef[2], Kusum V. Pundit[1,2], Naftali Kaminski[2], Panayiotis V. Benos[3]

[1]Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA; [2]Division of Pulmonary, Allergy and Critical Care Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA; [3]Department of Computational Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

5.1 BACKGROUND

MicroRNA genes (miRNAs) are short non-coding RNAs that function as post-transcriptional gene regulators (6). Recent studies have implicated these RNAs in a variety of human diseases (34, 36, 89) leaving researchers with the question as to which specific cellular pathways miRNAs could be involved. Previous chapters in this dissertation have begun to identify the promoter regions where *cis*-regulatory elements may be able to bind and regulate miRNA expression. Identification of the specific transcription factors that regulate a given miRNA will provide a strong clue as to which cellular pathways a given miRNA may be a component.

To date, only a few biochemical studies that have been able to identify which transcription factors are regulating specific miRNAs (10, 12, 91, 120-122), but any study looking at large scale identification of networks have relied on purely computational predictions of the transcription start sites and therefore an incomplete understanding of where exactly the promoter regions for these miRNAs may be located (90, 123, 124). This is the first study to be able to use biochemically and computationally verified promoter regions to identify putative regulatory networks involving miRNAs.

The type of network that will be the focus of this study is called the feed-forward loop. The functional properties of this type of network are that a transcription factor is responsible for the up-regulation of a given gene as well as repression of all miRNAs that targets that gene. An example of this network has been shown in O'Donnell *et al* (120) in which the TF c-Myc regulates miR-17-5p and miR-20a which both target E2F, a gene also regulated at the same time by c-Myc resulting in cell cycle progression.

The cellular pathway that will be the focus of this study involves stimulation of lung epithelial cells by *transforming growth factor beta* (TGFβ). The TGFβ pathway has been studied in human diseases such as renal disease and idiopathic pulmonary fibrosis (125, 126). The TGFβ pathway is typically characterized by activation of the SMAD family of TFs, which in cooperation with each other or other factors activate or repress hundreds of genes (127, 128).

The purpose of this chapter is to identify putative feed-forward loops involved in the TGFβ signaling pathway. We will accomplish this task by searching miRNA promoters for binding sites of factors known to play a role in the TGFβ, including those promoters verified in Chapter 3 of this dissertation by the RNA Polymerase II chromatin immunoprecipitation data. We will then search the promoter regions of all predicted targets of those miRNA genes for the same set of transcription factors, hopefully identifying protein coding genes that are targets of both a miRNA gene and a transcription factor that regulates that targeting miRNA gene.

5.2 RESULTS AND DISCUSSION

5.2.1 Identification of miRNAs that may play a role in TGFβ signaling

The ability of miRNA genes to regulate coding genes at a translation level makes it difficult to verify the accuracy of miRNA target prediction methods because the majority of high-throughput gene expression data available is for transcripts levels and not protein levels. miRNA target prediction algorithms currently identify hundreds of genes that are as targets for each miRNA and identify many genes as being targeted by multiple miRNAs, making it difficult to assemble putative regulatory networks (7-9).

To identify which miRNAs may play a role in TGFβ signaling we stimulated A549 epithelial lung cancer cells with TGFβ and then used a gene expression microarray to determine the change in gene expression prior to and 4 hours post-stimulation. The function of the feed-forward loop requires that in order for a gene to be expressed it needs to both be activated by a TF as well as have its post-transcriptional regulatory miRNA down regulated. To identify the miRNA genes that may play a role in the TGFβ pathway the top 10% of genes that showed the largest increase in expression levels post TGFβ stimulation were inputted into the TARGETSCAN algorithm (40) and the most overrepresented miRNA genes (see *Materials and Methods)* that target those up regulated protein coding genes were identified (Table 5.1).

From that analysis we identified seventeen miRNA families to be significantly over-represented (p-value < 0.01) in the top 10% of genes that showed the highest fold change 4 hours post TGFβ stimulation as compared to all protein coding genes. Of those 17 miRNA families, 11 were found to have at least 1 member expressed in A549 cells prior to TGFβ stimulation, suggesting that if their target genes are to be expressed they may need to be down regulated. All of the 11 families were found to be expressed in A549s showed a decrease in expression 2 hours post TGFβ stimulation (see *Materials and Methods*) according to an miRNA expression microarray. Interestingly, the majority of the miRNAs whose targets are up regulated and were expressed prior to stimulation show a decrease in expression level after stimulation reaffirming the target prediction method.

Table 5.1 miRNAs likely to target coding genes upregulated post TGFβ stimulation
p-value corresponds to how likely that miRNA family was in being found in the top 10% of upregulated genes post TGFβ stimulation as compared to all genes. Expression in A549 cell line is determined by miRNA microarray at baseline. Change in expression shows the direction that each miRNA family member'a expression changed post stimulation.

miRNA Family	***p*-value (overrepresentation in upregulated genes)**	**Expressed in A549 Cell Line**	**Change In Expression Post TGFβ Stimulation**		
			Up-regulated	**No Change**	**Down-regulated**
hsa-let-7a/b/c/d/e/f/g	0.0090	YES (all)	■	b/c	a/d/e/f/g
hsa-miR-124	0.0078	NO	■	■	■
hsa-miR-506	0.0023	NO	■	■	■
hsa-miR-30b/c/d/a-5p/e-5p	0.0008	YES (d)	■	■	Yes
hsa-miR-381	0.0001	NO	■	■	■
hsa-miR-17-5p	0.0004	YES	■	■	Yes
hsa-miR-519a/b/c/d	0.0003	NO	■	■	■
hsa-miR-20a/b	0.0001	YES (all)	■	■	a/b
hsa-miR-106a/b	0.0001	YES (all)	■	■	a/b
hsa-miR-27a/b	0.0013	YES (all)	■	■	a/b
hsa-miR-607	0.0016	NO	■	■	■
hsa-miR-29a/b/c	0.0019	YES (a/b)	■	■	a/b
hsa-miR-181a/b/c/d	0.0017	YES (b/d)	■	c/d	a/b
hsa-miR-23a/b	0.0015	YES (all)	■	■	a/b
hsa-miR-19a/b	0.0093	YES (all)	■	■	a/b
hsa-miR-496	0.0038	NO	■	■	■
hsa-miR-301	0.0005	YES	■	■	Yes

5.2.2 Assembling Putative Feed-Forward Loops Involving miRNAs

While the SMAD family of transcription factors is the most often mentioned TFs when discussing TGFβ signaling, other factors have been shown to play a role such as c-JUN/c-FOS (also known as AP-1) as well as SP1, p53 and NF-kB. To identify potential feed-forward loops, the 3kb upstream region of each upregulated gene as well as intergenic miRNAs were analyzed with the FOOTER algorithm for putative binding sites for these 6 factors. If a miRNA transcription start site had been identified in Chapter 4 of this dissertation, the 3kb upstream of that location was analyzed with FOOTER as well. Previous studies have demonstrated that TFBS can regulate miRNAs from either being upstream of the intergenic pre-miRNA or of the pri-miRNA transcription start site (91, 121). Promoter regions of intronic miRNAs were identified as the 3kb upstream of their host gene. Analysis of the promoter regions for the 23 miRNAs that show a decrease in expression post TGFβ stimulation as well as all of their targets identified a variety of putative feed-forward loops used by the cell (Table 5.2).

Table 5.2 Number of genes in possible feed-forward loops for the given miRNA gne and transcription factor

miRNA	**Transcription Factors**	**Number of Genes Predicted Targets of Both the TFs and miRNA Family Upregulated Post** TGFβ stimulation
hsa-let-7a/d/e/f/g	SMAD3/4	6
	Sp1	16
	p53	0
	NF-kB	4
hsa-miR-30d	SMAD3/4	8
	Sp1	20
	p53	1
hsa-miR-17-5p	SMAD3/4	5
	AP-1	10
	NF-kB	3
	p53	0
hsa-miR-20a/b	SMAD3/4	5
	AP-1	9
	NF-kB	1
	p53	0
hsa-miR-106a/b	SMAD3/4	4
	AP-1	9
	NF-kB	3
	p53	1
	Sp1	13
hsa-miR-27a/b	SMAD3/4	6
	NF-kB	9
hsa-miR-29a/b/c	NF-kB	4
	Sp1	21
	p53	2
hsa-miR-181a/b	SMAD3/SMAD4	8
	Sp1	26
	p53	1
hsa-miR-23a/b	SMAD3	4
	NF-kB	3
hsa-miR-19a/b	SMAD3/4	3
	AP-1	9
	NF-kB	5
	p53	0
hsa-miR-301	SMAD3/4	2
	AP-1	7
	NF-kB	4

5.3 CONCLUSION

Use of a combination of biochemical experiments and *in silico* prediction methods allows for a first step toward the understanding of the complex regulatory networks involving miRNAs during cell signaling pathways. While the networks we identified in this study are reliant upon the transcription factor binding site prediction algorithm as well as the miRNA target prediction algorithm it still provides the best idea of which miRNAs may be involved in feed-forward loops post TGFβ stimulation. The list of putative networks should be further explored by experimental study to confirm the findings in this study.

5.4 MATERIALS AND METHODS

5.4.1 Identification of over-represented miRNA target genes

Gene expression data of A549 cells was collected from Ranganathan *et al.* (129). Duplicate experiments were performed by their group in which transcript levels were collected 4 hours post TGFβ stimulation and compared to a non-stimulated control. The top 10% of genes, as averaged between both duplicate experiments, that showed up-regulation post TGFβ stimulation were used to identify miRNA target sites based upon 'conserved site' results from the TARGETSCAN algorithm (40). Of these up-regulated genes, 426 were identified as targets of 433 different miRNAs.

To identify which of the miRNAs were over-represented in the set of up regulated genes a method was set up in which random sets of genes were chosen from the expression array until 426 were found to be targets of miRNAs. The number of instances that a given miRNA was found in that set was counted and recorded to create a distribution as to the probability of finding the identified number of instances in the up-regulated genes by chance. This procedure was repeated 10,000 times for each miRNA. A cutoff of 0.1% was used to determine significance of miRNA over-representation.

5.4.2 Cell Culture

A549 cells (CCL-185, American Type Culture Collection, Manassas, VA) were grown in F12K medium (Invitrogen, Carlsbad, CA) with 2 mM L-glutamine and 10% fetal bovine serum. Cells were incubated at 37 °C in a humidified chamber supplemented with 5% CO_2. Once 80% confluent, cells were serum starved overnight and stimulated with 10 ng/ml TGFβ (R&D, Minneapolis, MN)

5.4.3 RNA Isolation

Total RNA was isolated using the miRNeasy Mini kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. The quantity of the RNA was determined by optical density, measured at 260nm by Nanodrop spectrophotometer and its quality was measured using Agilent Bioanalyzer 2100.

5.4.4 miRNA Expression Array

miRNA profiling was carried out using the Agilent Human miRNA Microarray. These microarrays have an 8 x 15K design with 470 miRNAs based on Release 9.1 of Sanger miRBASE. The manufacturer's instructions were followed in the labeling and hybridization of the RNA. Briefly, 100 ng of total RNA was dephosphorylated using calf intestine alkaline phosphatase (GE Healthcare, Piscataway, NJ), denatured with DMSO, and labeled with pCp-Cy3 using T4 RNA ligase (New England Biolabs, Ipswich, MA) at 16°C for 2h. The labeled RNA was purified using Micro Bio-spin 6 columns and hybridized onto the Agilent miRNA microarrays at 55°C for 20h. The arrays were washed with Gene Expression Wash Buffers 1 and 2 (Agilent) and scanned using the

Agilent Microarray Scanner. The scanned images were processed by Agilent's Feature Extraction software version 9.5.3.

5.4.5 miRNA Expression Array Analysis

miRNAs were identified as expressed if all probes for each of the miRNAs had a value greater than 90th percentile of the negative controls on each of the arrays. The median value of the negative controls was used to normalize the data across the arrays; this value was used because of the small number of genes being expressed and the inability to assume that the majority of those genes are not changing expression post stimulation. miRNAs were considered to have a change in expression if all of the probes for a specific miRNA showed a change of expression in the same direction. No statistical significance can be derived from the miRNA expression array analysis because of the single replicate experiment.

5.4.6 Identification of Transcription Factor Binding Sites

Genomic coordinates of all miRNAs whose targets were identified to be over-represented in the set of up-regulated genes were obtained from the UCSC Genome Browser (49, 50). The 3kb upstream sequence of the intergenic human miRNAs as well as the 3kb upstream of the transcription start site identified in chapter 3, when applicable, were collected along with the as well as the 3kb upstream of the host gene for all intronic miRNAs. The aligned mouse sequence of each of the human 3kb regions was also collected as identified by the UCSC genome browser. The host gene promoter sequence was used for intronic miRNAs because previous reports have shown that the miRNA and host gene are co-transcribed and share the same promoter region(4). Genomic coordinates and the 3kb upstream sequence for all human and mouse genes identified as a target of a significant miRNA family were collected from the UCSC genome browser. SMAD3, SMAD4, p53, Sp1, AP1 and NF-kB binding site prediction was carried out with the FOOTER algorithm using default parameters (28).

6.0 REGULATION AND ROLE OF LET-7D MICRORNA IN EPITHELIAL MESENCHYMAL TRANSITION

David L. Corcoran[1*], Kusum V. Pundit[1,2*], Hanadie Yousef[2], Daniel Handley[1,2], Kazuhisa Konishi[2], Selman Moises[3], Pardo Annie[4], Ahmy Ben-Yehuda[2], Oliver Eickelberg[5], Panayiotis V. Benos[6], Naftali Kaminski[2]

[1]Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA; [2]Division of Pulmonary, Allergy and Critical Care Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA; [3]Agilent Technologies Inc., Santa Clara, CA, USA; [3]Instituto Nacional de Enfermedades Respiratorias, Mexico DF, Mexico; [4]Facultad de Ciencias, Universidad Nacional Autonoma de Mexico, Mexico; [5]Universitat Gieben, Giessen, Germany; [6]Department of Computational Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA; [*]These authors contributed equally.

6.1 BACKGROUND

Transforming growth factor beta (TGFβ) has been identified to induce alveolar epithelial cells to undergo epithelial mesenchymal transition (EMT) (130-132). EMT is the phenomenon in which polarized epithelial cells are converted to motile mesenchymal cells and occurs during embryogenesis and tumor invasion. EMT has also been described to occur in idiopathic pulmonary fibrosis (IPF) during which injured alveolar epithelial cells migrate and proliferate to form myofibroblasts in the fibroblast foci (130, 132).

Morphologically, the cells acquire an elongated phenotype and are characterized by loss of the epithelial cell marker *E-cadherin* and expression of mesenchymal cell markers such as *vimentin*, *N-cadherin* and *alpha smooth muscle actin* (133). The TGFβ signaling cascade is initiated by its binding to TGFβ-RI and TGFβ-RII resulting in phosphorylation of the effectors SMAD2 and SMAD3. These molecules form a complex with SMAD4, which is translocated to the nucleus whereupon they modulate gene transcription.

TGFβ mediates EMT by induction of HMGA2 (131), a chromatin-associated protein present during embryogenesis and carcinogenesis. HMGA2 lacks intrinsic transcriptional activity but modulates transcription by altering chromatin structure and facilitating assembly of transcription factors (134). HMGA2 regulates *snail*, *slug*, *twist* and Id2, the key players in EMT (131).

Recent studies have identified miRNAs as key components in many complex human diseases including Fragile X syndrome (89), B cell lymphoma (34) and chronic heart failure (36). Considering that the lung in IPF is characterized by profound changes in the phenotype of epithelial cells, as well as drastic changes in gene expression (135, 136), this study hypothesizes that miRNAs will be differentially expressed during EMT. miRNAs are short, ~22nt, RNAs and act as post-transcriptinal gene regulators that function by binding to specific sequences in the 3' untranslated region of the target mRNAs. Once bound, miRNAs are capable of blocking translation or causing the rapid degradation of the target transcript (6). Each miRNA is predicted to target a large number of genes and many genes are predicted to be the target of multiple miRNAs (137). It is the focus of this study to identify miRNA genes that are involved in the SMAD signaling pathway and therefore potentially play an important role in the induction of EMT. The previous chapter of this dissertation identified putative feed-forward loops involving the TGFβ signaling pathway. One of the loops we identified involved the HMGA2 gene, it is this loop involving the SMAD transcription factors, the let-7d miRNA gene, and HMGA2 that will be further explored with biochemical verification.

6.2 RESULTS AND DISCUSSION

6.2.1 Identification of miRNA genes Regulated by SMAD Transcription Factors

Previous chapters of this dissertation have demonstrated that many miRNAs are likely transcribed by RNA polymerase II (Chapters 2 and 4), share a similar sequence conservation structure in the immediate upstream region as protein coding genes (Chapter 2), and have similar sequence features to protein coding genes in their core promoter regions (Chapter 4). These observations suggest that miRNAs have similar transcriptional regulatory mechanisms as other RNA polymerase II transcribed genes. The previous chapter of this dissertation identified putative feed-forward loops involving the SMAD family of transcription factors. One of our identified loops involved HMGA2, a gene known to play a regulatory role in EMT. The putative binding site in the upstream region of hsa-let-7d was further investigated after qRT-PCR analysis in A549 cells revealed that the let-7d miRNA did have a change in expression post TGFβ stimulation (Figure 6.1), confirming the results of the miRNA expression microarray (see Chapter 5).

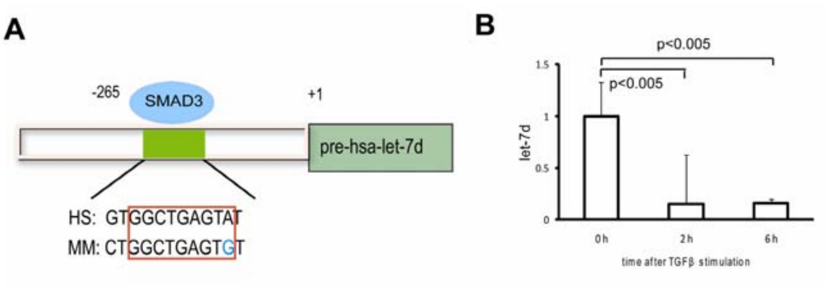
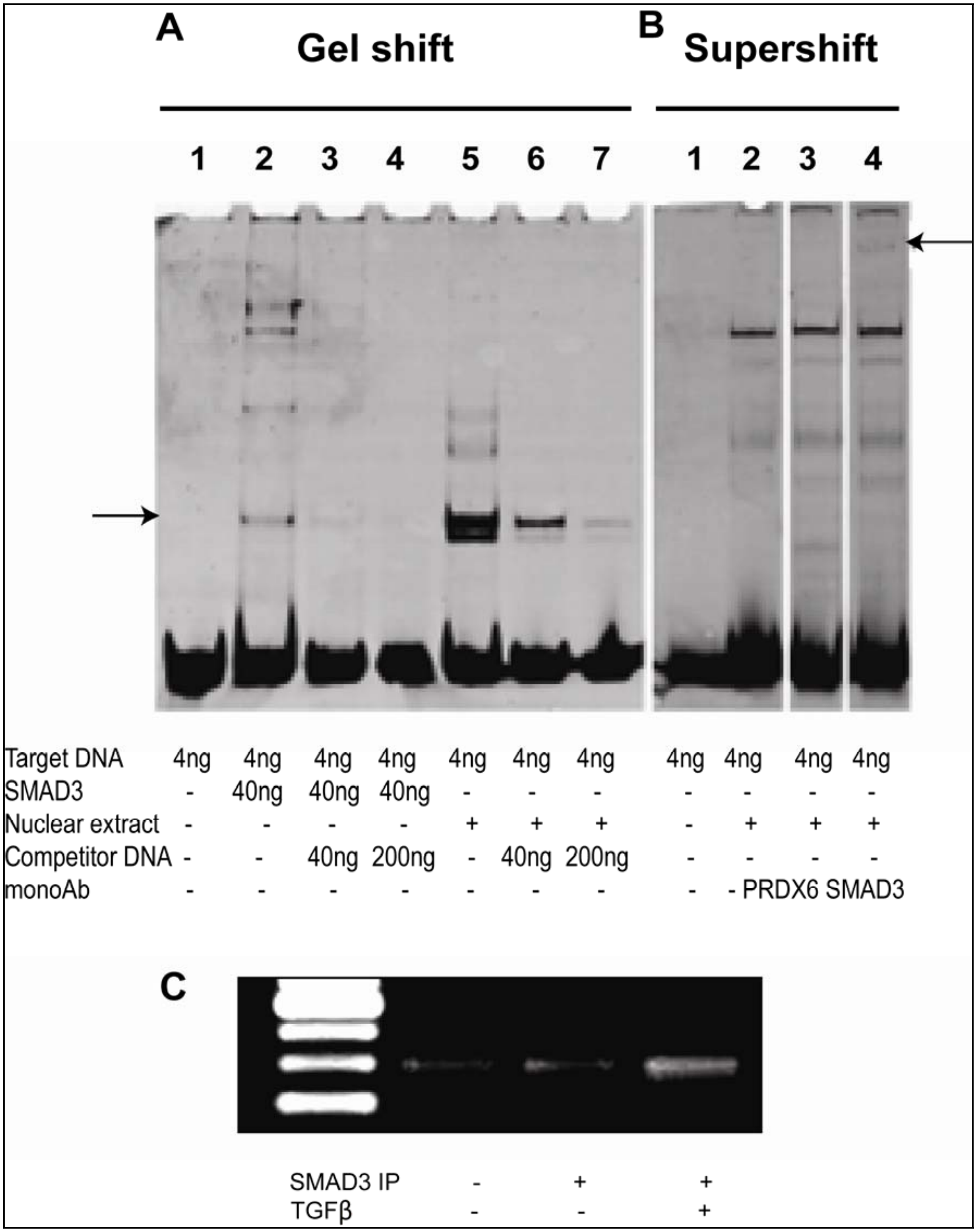


Figure 6.1 miRNA gene let-7d potentially regulated by SMAD transcription factors
(A) Putative SMAD3 binding sites identified by the FOOTER algorithm upstream of hsa-let-7d. HH: human sequence, MM: mouse sequence. (B) A549 cells were treated with 10 ng/ml recombinant TGFβ and let-7d expression determined at 0h, 2h and 6h post-stimulation. The results represent an average expression of triplicate experiments.

6.2.2 SMAD3 Binds to the Let-7d Promoter

We used EMSA to verify the computational finding of SMAD3 binding at the let-7d promoter (Figure 6.2A). Incubation of target DNA with either recombinant SMAD3 protein or nuclear extract of A549 cells revealed a distinct band representing the binding of SMAD3 to the let-7d promoter sequence. The intensity of these bands diminished in the presence of increasing concentrations of competitor DNA. We also used a supershift assay to ascertain the specificity of the SMAD3 binding site; demonstrated by incubating the target DNA and nuclear extract in the presence of antibodies to SMAD3 and *peroxiredoxin 6*. The supershift band representing the DNA-protein-antibody complex was visible with SMAD3 antibody but not with the control *peroxiredoxin 6* antibody (Figure 6.2B).

We then used chromatin immunoprecipitation (ChIP) to demonstrate the presence of SMAD3 on the let-7d promoter *in vivo* (Figure 6.2C). Immunoprecipitation with anti-SMAD3 antibody following TGFβ stimulation yielded a distinct band by gene-specific PCR demonstrating the association of SMAD3 with the let-7d promoter. In contrast, immunoprecipitation with an irrelevant antibody resulted in the absence of this band confirming the specificity of the interaction.



	1	2	3	4	5	6	7	1	2	3	4
Target DNA	4ng	4ng	4ng	4ng	4ng	4ng	4ng	4ng	4ng	4ng	4ng
SMAD3	-	40ng	40ng	40ng	-	-	-	-	-	-	-
Nuclear extract	-	-	-	-	+	+	+	-	+	+	+
Competitor DNA	-	-	40ng	200ng	-	40ng	200ng	-	-	-	-
monoAb	-	-	-	-	-	-	-	-	-	PRDX6	SMAD3

SMAD3 IP	-	+	+
TGFβ	-	-	+

Figure 6.2 Verification of SMAD binding to the let-7d promoter region
(A) Electromobility shift assay and (B) supershift assays of recombinant SMAD3 protein and nuclear extracts. (C) SMAD3 ChIP assay revealed *in vivo* association with let-7d in A549 lung cells.

6.2.3 Inhibition of Let-7d Results in EMT

Thuault *et al.* showed that in mouse mammary epithelial cells TGFβ mediates EMT by induction of HMGA2 (131). A similar result is demonstrated in a human lung epithelial lung cell line, A549, (Figure 6.3A) where a 2-fold increase in HMGA2 is observed at 2h and remains elevated at 6h post-stimulation (p value<0.005) while its post-transcriptional regulator, let-7d, is down-regulated (Figure 6.1B).

We transfected A549 cells with a let-7d inhibitor for 24 and 48 hours to demonstrate the involvement of let-7d in EMT. Let-7d inhibition led to a 3-fold increase in HMGA2 at 24h that remained elevated at 48 hrs (*data not shown*). qRT-PCR of the mesenchymal markers *N-cadherin, vimentin* and *alpha smooth muscle actin* revealed an increase of 5-fold (p<0.05) , 8-fold (p<0.05) , and 6-fold (p< 0.05) respectively (Figure 6.3B). Immunofluorescence was performed to confirm the results at the protein level. A549 cells transfected with anti-let-7d undergo EMT as evidenced by positive staining to *N-cadherin, vimentin* and *alpha smooth muscle actin* 48h post-transfection. This staining was absent in mock transfected cells (Figure 6.3C). Our findings demonstrate a new component of the TGFβ signaling cascade that results in the induction of EMT. We have shown in this study that for the cell to up regulate expression of HMGA2 it also must reduce the amount of let-7d (Figure 6.4).

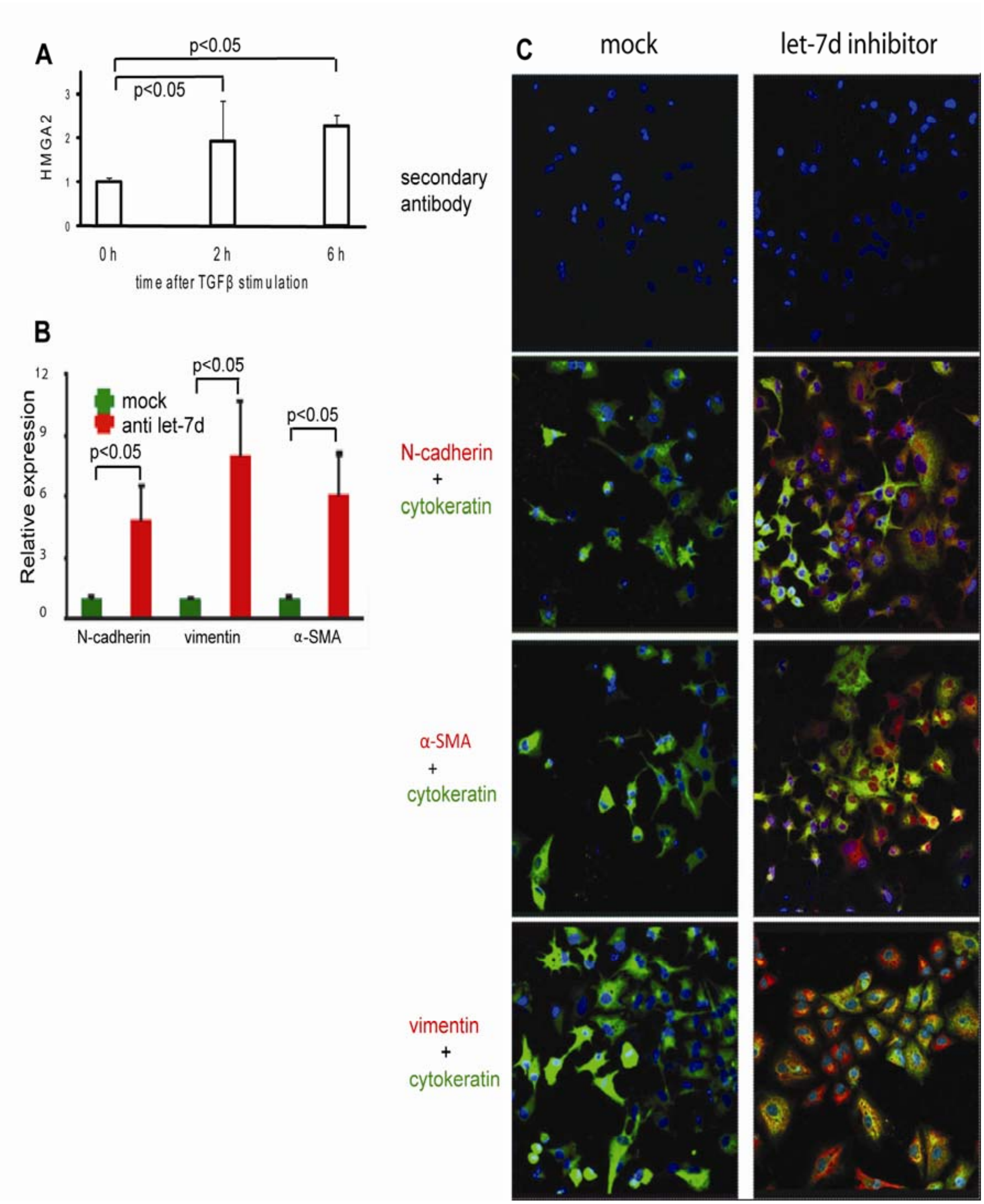


Figure 6.3 Cellular response to changes in let-7d expression

(A) HMGA2 mRNA levels determined by qRT-PCR in A549 cells at 0h, 2h and 6h post-stimulation with 10ng/ml recombinant TGFβ. The results represent an average expression of triplicate experiments. (B) HMGA2 mRNA levels determined by qRT-PCR in A549 cells at 24h and 48h post-transfection with 50 nM of let-7d inhibitor. (C) Immunofluorescence imaging of A549 cells transfected with 50 nM of let-7d inhibitor. The green fluorescent antibody was tagged to cytokeratin, an epithelial marker. The red fluorescent antibody was tagged to the mesenchymal markers. Nuclei were counterstained with DAPI. While red staining is observed in cells transfected with let-7d inhibitor (right panel), there is no staining in cells transfected with a control oligonucleotide (left panel)

6.3 CONCLUSION

With recent studies identifying important roles for miRNAs in complex human diseases (34, 36, 89) it has been the focus of this study to determine putative roles for miRNAs in EMT (131). EMT has previously been shown to be induced by the TGFβ signaling pathway in which the SMAD family of transcription factors is activated by phosphorylation where upon they enter the nucleus to regulate transcription of their target genes. One specific gene the SMAD factors have been shown to activate is HMGA2, a known key regulator of EMT. We have demonstrated in this study that not only do the SMAD factors up-regulate HMGA2 upon TGFβ stimulation, but they also repress its post-transcriptional regulator, the miRNA let-7d. Hence, our data indicates that SMAD proteins induce the expression of HMGA2 by a feed-forward mechanism. This is the first study to demonstrate the direct influence of a growth factor on the transcriptional regulation of miRNAs.

Importance of the let-7d branch of the TGFβ/SMAD/HMGA2 pathway in EMT was demonstrated by our study in the A549 lung epithelial cell line by change in let-7d and HMGA2 expression upon TGFβ stimulation that resulted in up-regulation of mesenchymal proteins. Inhibition of let-7d alone was sufficient to cause EMT in A549 cells.

A previous study has identified the role of the miR-200 family in the direct regulation of two transcription factors that regulate *E-cadherin*, an epithelial cell marker (138). While the current study does not directly overlap the putative pathway that they have identified, there is no doubt that there may be multiple signaling cascades likely to be at work in a complex cellular transition such as EMT.

While this study has demonstrated on a cellular level the importance of let-7d in induction of EMT, further research with the use of animal models is required for certainty. The advancement in understanding of this TGFβ pathway that may play an important role in the complex and chronic disease IPF may better the chances for a new therapeutic target.

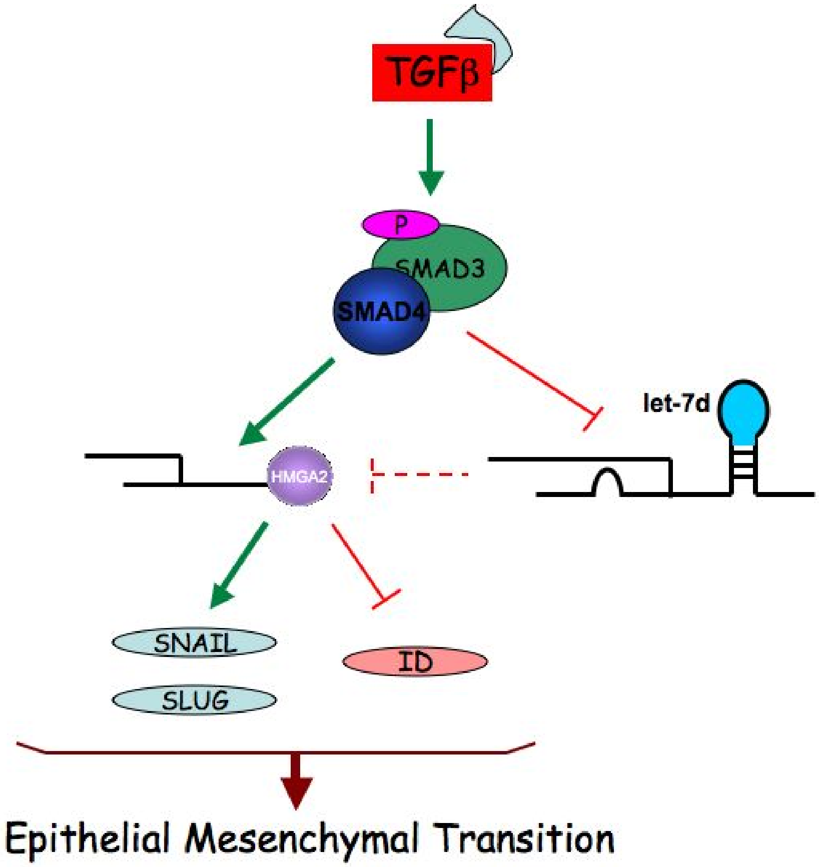


Figure 6.4 Cartoon depiction of the let-7d pathway involved in EMT.

6.4 MATERIALS AND METHODS

6.4.1 miRNA Promoter Analysis

Genomic coordinates of all miRNAs identified to be differentially expressed between IPF and control lungs as well as the coordinates of the murine orthologues were obtained from the UCSC Genome Browser (49, 50). The 1kb upstream sequence of the intergenic human and mouse miRNAs were collected as well as the 1kb upstream of the host gene for all intronic miRNAs. The host gene promoter sequence was used for intronic miRNAs because previous reports have shown that the miRNA and host gene are co-transcribed and share the same promoter region(4). SMAD3 and SMAD4 binding site prediction was carried out with the FOOTER algorithm using default parameters (28).

6.4.2 Cell culture

A549 cells (CCL-185, American Type Culture Collection, Manassas, VA) were grown in F12K medium (Invitrogen, Carlsbad, CA) with 2 mM L-glutamine and 10% fetal bovine serum. Cells were incubated at 37 °C in a humidified chamber supplemented with 5% CO_2. Once 80% confluent, cells were serum starved overnight and stimulated with 10 ng/ml TGFβ (R&D, Minneapolis, MN).

6.4.3 Chromatin Immunoprecipitation (ChIP)

The ChIP protocol (139) was a performed according to the published protocol from the Young laboratory (73). A549 cells were grown to 5×10^7–1×10^8 cells per analysis condition. Cells were either untreated (control) or stimulated with 2 ng/mL TGFβ for 30 minutes. Chromatin cross-linking was performed by adding 1/10 volume of freshly prepared 11% formaldehyde solution for 15 minutes at room temperature. The crosslinking reaction was then quenched by adding 1/20 volume of 2.5M glycine. Cells were rinsed twice with PBS, collected with a silicon scraper, flash frozen in liquid nitrogen, and stored at −80°C until use. Upon thawing, cells were resuspended in a lysis buffer and sonicated at 4°C to solubilize cellular components and shear crosslinked chromatin. The cell lysate was incubated overnight at 4°C with 100 µl of Dynal Protein G magnetic beads that had been preincubated with 10 µg of either anti-flag (mock IP) or anti-SMAD3 antibodies (Millipore, Billerica, MA). Protein G magnetic beads were washed five times with RIPA buffer and one time with TE buffer containing 50 mM NaCl. Cross-linked promoter fragment/transcription factor complexes were eluted from the beads by heating at 65°C with vortexing at 2 minute intervals for 15 minutes. Crosslinking was reversed by incubation at 65°C overnight. Recovered promoter fragments were treated with RNaseA, proteinase K digestion, and purified by phenol:chloroform:isoamyl alcohol extraction/ethanol precipitation. Gene-specific PCR was performed on a portion of the purified recovered nucleic acid (25 cycles) to verify the presence of the upstream sequence of pre-hsa-let-7d. The primers used for gene-specific PCR are: *let-7d forward*: 5' - CAC TTA AAC CCA GGA GGC AGA GGT T - 3' and *let-7d reverse*: 5' - ACC ACG TAT TAC TGG AGT CGC TGA - 3'.

6.4.4 Electromobility Shift Assay (EMSA)

Cultured A549 lung alveolar epithelial carcinoma cells at 60-70% confluence were treated with 2 ng/mL recombinant human TGFβ_1 (R&D Systems) for 60 minutes. Nuclear proteins were isolated using a standard rapid micropreparation technique described previously (140). The supernatant was reserved and snap frozen in liquid nitrogen as the nuclear protein fraction. Nuclear extracts and recombinant full length SMAD3 protein (Santa Cruz Biotechnology, Santa Cruz, CA) were incubated with 5'-end Cyanine-5 labeled probe and/or non-labeled competitor oligonucleotide for 20 minutes at room temperature in a binding buffer consisting of 20% glycerol, 5 mM $MgCl_2$, 2.5 mM EDTA, 25 mM DTT, 200 mM NaCl, 50 mM Tris HCl pH 7.6, and 0.25 mg/mL poly(dI-dC). The complementary oligonucleotides (5' - GATAATTAAATGTTAAAAGTCAGC - 3', 5' - GCTGACTTTTAACATTTAATTATC - 3') were synthesized by Integrated DNA Technologies (Coralville, IA), and consisted of a sequence upstream of the predicted SMAD3/let7d binding site (GGCTGAGTA). Additionally, a supershift assay was performed by incubating nuclear extract with 0.1 μl rabbit monoclonal antibody [EP568Y] to SMAD3 (Abcam, Cambridge, MA) or 1.0 μl mouse monoclonal [4A3] to peroxiredoxin 6 as a control (Abcam) prior to incubating with the target oligonucleotide. The protein/DNA complexes were run on a 6% native polyacrylamide gel and visualized on a Typhoon imaging and documentation system using Cyanine-5 dye excitation and fluorescence settings.

6.4.5 RNA Isolation

Total RNA from A549 cells was isolated using the miRNeasy Mini kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. The quantity of the RNA was determined by optical density, measured at 260nm by Nanodrop spectrophotometer and its quality was measured using Agilent Bioanalyzer 2100.

6.4.6 Quantitative RT-PCR

TaqMan MicroRNA assays (ABI, Foster City, CA) were used to determine the relative expression levels of hsa-let-7d, miR-30c, miR-30d and miR-30e-5p. For RT reactions, 50 ng of total RNA was used in each 15 µl reaction. The conditions for the RT reaction were: 16 °C for 30 min; 42 °C for 30 min; 85 °C for 5 min; and then held on 4 °C. The cDNA was diluted 1:14 and 1.33 µl of the diluted cDNA was used with the TaqMan primers in the PCR reaction. The conditions for the PCR were: 95 °C for 10 min followed by 40 cycles of 95 °C for 15 s and 60 °C for 1 min in the ABI 7300 real-time PCR system. The results were analyzed by the ΔΔCt method using RNU43 control RNA to normalize the results. Fold change was calculated taking 0h as the baseline. TaqMan gene expression assays (ABI) were used to determine the relative expression levels of HMGA2, N-cadherin, vimentin and alpha smooth muscle actin. 500 ng of RNA was reverse transcribed using the SuperScript First-Strand Synthesis System for RT-PCR (Invitrogen) in a total reaction volume of 20 µl, the cDNA diluted 1:5 and 3 µl of this cDNA was used in a total volume of 31 µl for the PCR. PCR conditions were as follows: 12 min at 95°C, followed by 40 cycles with 15 s at 95°C and 1 min at 60°C in the ABI 7300 real-time PCR system. The results were analyzed by the ΔΔCt method and GUSB was used for normalization. Fold change was calculated taking 0h as the baseline.

6.4.7 Transfection

A549 cells were plated in 6-well plates at 50% confluence in F12K medium containing 10% fetal bovine serum. After the cells were adherent, the medium was changed to Opti-MEM I reduced serum medium (Invitrogen). Transfection of hsa-let-7d inhibitor and the negative control (Ambion, Austin, TX) was carried out at 50 nM using Lipofectamine

2000 (Invitrogen) according to the manufacturer's instructions. RNA was isolated 24h and 48h post-transfection.

6.4.8 Immunofluorescence

A549 cells were plated on cover slips. Cells were starved for 24 hours by the removal of serum and transfected with 50 nM anti-let7d for 48 hours. Cover slips were removed, washed with PBS three times for 5 minutes each and then fixed in 1% paraformaldehyde (Sigma) for 40 minutes. Permeabilization of cells was carried out by using 0.1% Triton X in PBS for 15 minutes, air-dried and then dehydrated with three washes in PBS each for five minutes followed by three washes each for five minutes with 0.3% BSA and 5% goat serum in PBS. Cover slips were blocked with 2% BSA in PBS for 60 minutes, and incubated with anti cytokeratin, anti vimentin, anti N-Cadherin, or anti alpha-smooth muscle actin (Abcam Inc., Cambridge, MA) in 0.5% BSA in PBS for 60 minutes, followed by three washes with 0.5% BSA in PBS (5 minutes each), and incubated with rabbit anti mouse conjugated to Alexia 488 (Invitrogen, Carlsbad, CA) for one hour. After staining, cover slips were washed with 0.1% Triton in PBS for 2 times 5 minutes each followed by three washes with PBS 5 minutes each. Coverslips were inverted onto slides and mounted in Vectashield anti-fade medium that contained DAPI for nuclei staining (Vector Laboratories, Burlingame, CA) to prevent photobleaching. Slides were examined using a Leica TCS-SP2 laser scanning confocal microscope equipped with appropriate lasers for simultaneous imaging of up to four fluorophores. Digital data was archived to compact disk or DVD and prepared for publication using Adobe Photoshop software (Adobe Systems Inc., MountainView, CA).

7.0 CONCLUDING REMARKS

7.1 MOTIVATION

The advancement in the knowledge of miRNA function and their role in human diseases have created a large need for the understanding of how the transcription of these small, non-coding RNAs are regulated. Previous studies have demonstrated that miRNAs are transcribed by RNA polymerase II (2), suggesting that they will have promoter regions similar to those found regulating protein coding genes. This allows for a similar set of biochemical and computational tools to be applied in the study of miRNAs promoters. The first step in identifying the regulators of miRNAs is to identify their transcription start site, which will then provide for the location of their core promoter and a possible location for the proximal promoter and enhancer regions. The promoter regions can then be analyzed with tools for identifying transcription factor binding sites, providing potential cellular networks in which miRNAs may participate.

7.2 SUMMARY OF MAJOR FINDINGS

The studies described in this dissertation have helped advance the understanding of the transcriptional regulation of miRNAs. The study began in Chapter 2 in which we were able to demonstrate that the upstream region of intergenic miRNA genes share the same evolutionary conservation features as protein coding genes, which are transcribed by RNA Polymerase II. This lead us to use an RNA Polymerase II ChIP-chip to identify the true transcription start site and therefore core promoter region of 35 intergenic miRNA genes or polycistronic gene clusters. It is the first study to date to use a high-throughput biochemical experiment to identify miRNA transcription start sites. Analysis of miRNA core promoter regions in Chapter 4 with a support vector machine model demonstrated that they share the same features as the core promoters of protein coding genes. That same analysis also provided us with evidence to suggest that some intronic miRNA genes may be transcribed by their own, unique promoter region.

The similarity in core promoters between miRNA and protein coding genes provided us incentive in Chapter 5 to search for transcription factor biniding sites resulting in the identification of putative feed-forward loops that may be involved in the TGFβ signaling pathway. One of the putative loops identified contained the gene HMGA2, which is a known key player in epithelial mesenchymal transition, a cellular phenotype known to occur in response to the TGFβ signaling pathway. In Chapter 6 we further investigated this feed-forward loop involving HMGA2, the miRNA let-7d and the SMAD family of transcription factors. We were able to demonstrate that the repression of let-7d was required in order for lung to express HMGA2 and allow the epithelial cells to undergo mesenchymal transition. This study is the first to demonstrate a direct link between a growth factor and the transcriptional regulation of a miRNA gene.

7.3 FUTURE CONSIDERATIONS

While the studies in this dissertation have provided a good first step into the understanding of the transcriptional regulation of miRNAs and the regulatory networks in which they participate, there is still much left to be discovered. The high throughput RNA polymerase II ChIP-chip should be repeated on many different cell types and under a variety of conditions in an attempt to identify transcription start sites for all human miRNAs. Those same arrays can also be useful for performing ChIP-chip with a variety of different transcription factors to accumulate a better understanding of just which factors may be the main regulators of miRNAs. The next important step in the identification of regulatory networks involving miRNAs is a high throughput method to locate all of the mRNA targets of miRNAs. To date, only a handful of these targets have been experimentally verified.

7.4 PUBLIC HEALTH SIGNIFICANCE

The identification of a growth factor directly regulating the transcription of a miRNA gene has big implications for public health. This finding demonstrates that miRNA expression should be studied in human diseases just as extensively as protein coding genes, if not more extensively because of their ability to post-transcriptionally regulate the expression of large numbers of other genes. The importance of understanding how these genes are regulated will likely have a large impact in the health sciences, especially as researchers continue to search for regulatory networks in which therapies or treatments can be developed for human diseases.

BIBLIOGRAPHY

1. Lee RC, Feinbaum RL, and Ambros V 1993 The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell **75**(5)**:**843-54.
2. Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, and Kim VN 2004 MicroRNA genes are transcribed by RNA polymerase II. EMBO J **23**(20)**:**4051-60.
3. Borchert GM, Lanier W, and Davidson BL 2006 RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol **13**(12)**:**1097-101.
4. Baskerville S and Bartel DP 2005 Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. Rna **11**(3)**:**241-7.
5. Rodriguez A, Griffiths-Jones S, Ashurst JL, and Bradley A 2004 Identification of mammalian microRNA host genes and transcription units. Genome Res **14**(10A)**:**1902-10.
6. Bartel DP 2004 MicroRNAs: genomics, biogenesis, mechanism, and function. Cell **116**(2)**:**281-97.
7. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, and Enright AJ 2006 miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res **34**(Database issue)**:**D140-4.
8. John B, Enright AJ, Aravin A, Tuschl T, Sander C, and Marks DS 2004 Human MicroRNA targets. PLoS Biol **2**(11)**:**e363.
9. Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, and Hatzigeorgiou A 2004 A combined computational-experimental approach predicts human microRNA targets. Genes Dev **18**(10)**:**1165-78.
10. Woods K, Thomson JM, and Hammond SM 2007 Direct regulation of an oncogenic micro-RNA cluster by E2F transcription factors. J Biol Chem **282**(4)**:**2130-4.
11. Fujita S and Iba H 2008 Putative promoter regions of miRNA genes involved in evolutionarily conserved regulatory systems among vertebrates. Bioinformatics **24**(3)**:**303-8.
12. Taganov KD, Boldin MP, Chang KJ, and Baltimore D 2006 NF-kappaB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. Proc Natl Acad Sci U S A **103**(33)**:**12481-6.
13. Zhou X, Ruan J, Wang G, and Zhang W 2007 Characterization and identification of microRNA core promoters in four model species. PLoS Comput Biol **3**(3)**:**e37.
14. Nikolov DB and Burley SK 1997 RNA polymerase II transcription initiation: a structural view. Proc Natl Acad Sci U S A **94**(1)**:**15-22.
15. Yang C, Bolotin E, Jiang T, Sladek FM, and Martinez E 2007 Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene **389**(1)**:**52-65.

16. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, and Hayashizaki Y 2006 Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet **38**(6)**:**626-35.

17. Antequera F 2003 Structure, function and evolution of CpG island promoters. Cell Mol Life Sci **60**(8)**:**1647-58.

18. Gershenzon NI and Ioshikhes IP 2005 Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. Bioinformatics **21**(8)**:**1295-300.

19. Smale ST and Kadonaga JT 2003 The RNA polymerase II core promoter. Annu Rev Biochem **72:**449-79.

20. Ince TA and Scotto KW 1995 A conserved downstream element defines a new class of RNA polymerase II promoters. J Biol Chem **270**(51)**:**30249-52.

21. Lee KW, Lee Y, Kwon HJ, and Kim DS 2005 Sp1-associated activation of macrophage inflammatory protein-2 promoter by CpG-oligodeoxynucleotide and lipopolysaccharide. Cell Mol Life Sci **62**(2)**:**188-98.

22. Ng SY, Gunning P, Liu SH, Leavitt J, and Kedes L 1989 Regulation of the human beta-actin promoter by upstream and intron domains. Nucleic Acids Res **17**(2)**:**601-15.

23. Stormo GD 2000 DNA binding sites: representation and discovery. Bioinformatics **16**(1)**:**16-23.

24. Mahony S, Auron PE, and Benos PV 2007 DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. PLoS Comput Biol **3**(3)**:**e61.

25. Sandelin A and Wasserman WW 2004 Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. J Mol Biol **338**(2)**:**207-15.

26. Umetani M, Mataki C, Minegishi N, Yamamoto M, Hamakubo T, and Kodama T 2001 Function of GATA transcription factors in induction of endothelial vascular cell adhesion molecule-1 by tumor necrosis factor-alpha. Arterioscler Thromb Vasc Biol **21**(6)**:**917-22.

27. Buck MJ and Lieb JD 2004 ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics **83**(3)**:**349-60.

28. Corcoran DL, Feingold E, Dominick J, Wright M, Harnaha J, Trucco M, Giannoukakis N, and Benos PV 2005 Footer: a quantitative comparative genomics method for efficient recognition of cis-regulatory elements. Genome Res **15**(6)**:**840-7.

29. Quandt K, Frech K, Karas H, Wingender E, and Werner T 1995 MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res **23**(23)**:**4878-84.

30. Bussemaker HJ, Li H, and Siggia ED 2001 Regulatory element detection using correlation with expression. Nat Genet **27**(2)**:**167-71.

31. Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, and Wasserman WW 2003 Identification of conserved regulatory elements by comparative genome analysis. J Biol **2**(2)**:**13.

32. Loots GG and Ovcharenko I 2004 rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Res **32**(Web Server issue)**:**W217-21.

33. Roth FP, Hughes JD, Estep PW, and Church GM 1998 Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol **16**(10)**:**939-45.

34. Pardo A, Gibson K, Cisneros J, Richards TJ, Yang Y, Becerril C, Yousem S, Herrera I, Ruiz V, Selman M, and Kaminski N 2005 Up-regulation and profibrotic role of osteopontin in human idiopathic pulmonary fibrosis. PLoS Med **2**(9)**:**e251.

35. Jin P, Alisch RS, and Warren ST 2004 RNA and microRNAs in fragile X mental retardation. Nat Cell Biol **6**(11)**:**1048-53.

36. Thum T, Galuppo P, Wolf C, Fiedler J, Kneitz S, van Laake LW, Doevendans PA, Mummery CL, Borlak J, Haverich A, Gross C, Engelhardt S, Ertl G, and Bauersachs J 2007 MicroRNAs in the human heart: a clue to fetal gene reprogramming in heart failure. Circulation **116**(3)**:**258-67.

37. Sandelin A, Wasserman WW, and Lenhard B 2004 ConSite: web-based prediction of regulatory elements using cross-species comparison. Nucleic Acids Res **32**(Web Server issue)**:**W249-52.

38. Siddharthan R, Siggia ED, and van Nimwegen E 2005 PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. PLoS Comput Biol **1**(7)**:**e67.

39. Ambros V 2004 The functions of animal microRNAs. Nature **431**(7006)**:**350-5.

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

41. Chen CZ, Li L, Lodish HF, and Bartel DP 2004 MicroRNAs modulate hematopoietic lineage differentiation. Science **303**(5654)**:**83-6.

42. Farh KK, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, and Bartel DP 2005 The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science **310**(5755)**:**1817-21.

43. Krichevsky AM, King KS, Donahue CP, Khrapko K, and Kosik KS 2003 A microRNA array reveals extensive regulation of microRNAs during brain development. RNA **9**(10)**:**1274-81.

44. Lee CT, Risom T, and Strauss WM 2006 MicroRNAs in mammalian development. Birth Defects Res C Embryo Today **78**(2)**:**129-39.

45. Wingender E 2004 TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. In Silico Biol **4**(1)**:**55-61.

46. Nichols M, Bell J, Klekamp MS, Weil PA, and Soll D 1989 Multiple mutations of the first gene of a dimeric tRNA gene abolish in vitro tRNA gene transcription. J Biol Chem **264**(29)**:**17084-90.

47. Leaman D, Chen PY, Fak J, Yalcin A, Pearce M, Unnerstall U, Marks DS, Sander C, Tuschl T, and Gaul U 2005 Antisense-mediated depletion reveals essential and specific functions of microRNAs in Drosophila development. Cell **121**(7)**:**1097-108.
48. Johnston RJ, Jr. and Hobert O 2005 A novel C. elegans zinc finger transcription factor, lsy-2, required for the cell type-specific expression of the lsy-6 microRNA. Development **132**(24)**:**5451-60.
49. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, and Kent WJ 2004 The UCSC Table Browser data retrieval tool. Nucleic Acids Res **32**(Database issue)**:**D493-6.
50. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D 2002 The human genome browser at UCSC. Genome Res **12**(6)**:**996-1006.
51. Hsu PW, Huang HD, Hsu SD, Lin LZ, Tsou AP, Tseng CP, Stadler PF, Washietl S, and Hofacker IL 2006 miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes. Nucleic Acids Res **34**(Database issue)**:**D135-9.
52. Kent WJ 2002 BLAT--the BLAST-like alignment tool. Genome Res **12**(4)**:**656-64.
53. Hayashita Y, Osada H, Tatematsu Y, Yamada H, Yanagisawa K, Tomida S, Yatabe Y, Kawahara K, Sekido Y, and Takahashi T 2005 A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. Cancer Res **65**(21)**:**9628-32.
54. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, and Haussler D 2005 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res **15**(8)**:**1034-50.
55. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, and Miller W 2004 Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res **14**(4)**:**708-15.
56. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng

JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S and Chen YJ 2001 Initial sequencing and analysis of the human genome. Nature **409**(6822)**:**860-921.

57. 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**(7055)**:**69-87.

58. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy

M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC and Lander ES 2002 Initial sequencing and comparative analysis of the mouse genome. Nature **420**(6915)**:**520-62.

59. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, Okwuonu G, Hines S, Lewis L, DeRamo C, Delgado O, Dugan-Rocha S, Miner G, Morgan M, Hawes A, Gill R, Celera, Holt RA, Adams MD, Amanatides PG, Baden-Tillson H, Barnstead M, Chin S, Evans CA, Ferriera S, Fosler C, Glodek A, Gu Z, Jennings D, Kraft CL, Nguyen T, Pfannkoch CM, Sitter C, Sutton GG, Venter JC, Woodage T, Smith D, Lee HM, Gustafson E, Cahill P, Kana A, Doucette-Stamm L, Weinstock K, Fechtel K, Weiss RB, Dunn DM, Green ED, Blakesley RW, Bouffard GG, De Jong PJ, Osoegawa K, Zhu B, Marra M, Schein J, Bosdet I, Fjell C, Jones S, Krzywinski M, Mathewson C, Siddiqui A, Wye N, McPherson J, Zhao S, Fraser CM, Shetty J, Shatsman S, Geer K, Chen Y, Abramzon S, Nierman WC, Havlak PH, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Li B, Liu Y, Qin X, Cawley S, Cooney AJ, D'Souza LM, Martin K, Wu JQ, Gonzalez-Garay ML, Jackson AR, Kalafus KJ, McLeod MP, Milosavljevic A, Virk D, Volkov A, Wheeler DA, Zhang Z, Bailey JA, Eichler EE, Tuzun E, Birney E, Mongin E, Ureta-Vidal A, Woodwark C, Zdobnov E, Bork P, Suyama M, Torrents D, Alexandersson M, Trask BJ, Young JM, Huang H, Wang H, Xing H, Daniels S, Gietzen D, Schmidt J, Stevens K, Vitt U, Wingrove J, Camara F, Mar Alba M, Abril JF, Guigo R, Smit A, Dubchak I, Rubin EM, Couronne O, Poliakov A, Hubner N, Ganten D, Goesele C, Hummel O, Kreitler T, Lee YA, Monti J, Schulz H, Zimdahl H, Himmelbauer H, Lehrach H, Jacob HJ, Bromberg S, Gullings-Handley J, Jensen-Seaman MI, Kwitek AE, Lazar J, Pasko D, Tonellato PJ, Twigger S, Ponting CP, Duarte JM, Rice S, Goodstadt L, Beatson SA, Emes RD, Winter EE, Webber C, Brandt P, Nyakatura G, Adetobi M, Chiaromonte F, Elnitski L, Eswara P, Hardison RC, Hou M, Kolbe D, Makova K, Miller W, Nekrutenko A, Riemer C, Schwartz S, Taylor J, Yang S, Zhang Y, Lindpaintner K, Andrews TD, Caccamo M, Clamp M, Clarke L, Curwen V, Durbin R, Eyras E, Searle SM, Cooper GM, Batzoglou S, Brudno M, Sidow A, Stone EA, Payseur BA, Bourque G, Lopez-Otin C, Puente XS, Chakrabarti K, Chatterji S, Dewey C, Pachter L, Bray N, Yap

VB, Caspi A, Tesler G, Pevzner PA, Haussler D, Roskin KM, Baertsch R, Clawson H, Furey TS, Hinrichs AS, Karolchik D, Kent WJ, Rosenbloom KR, Trumbower H, Weirauch M, Cooper DN, Stenson PD, Ma B, Brent M, Arumugam M, Shteynberg D, Copley RR, Taylor MS, Riethman H, Mudunuri U, Peterson J, Guyer M, Felsenfeld A, Old S, Mockrin S and Collins F 2004 Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428**(6982)**:**493-521.

60. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, 3rd, Zody MC, Mauceli E, Xie X, Breen M, Wayne RK, Ostrander EA, Ponting CP, Galibert F, Smith DR, DeJong PJ, Kirkness E, Alvarez P, Biagi T, Brockman W, Butler J, Chin CW, Cook A, Cuff J, Daly MJ, DeCaprio D, Gnerre S, Grabherr M, Kellis M, Kleber M, Bardeleben C, Goodstadt L, Heger A, Hitte C, Kim L, Koepfli KP, Parker HG, Pollinger JP, Searle SM, Sutter NB, Thomas R, Webber C, Baldwin J, Abebe A, Abouelleil A, Aftuck L, Ait-Zahra M, Aldredge T, Allen N, An P, Anderson S, Antoine C, Arachchi H, Aslam A, Ayotte L, Bachantsang P, Barry A, Bayul T, Benamara M, Berlin A, Bessette D, Blitshteyn B, Bloom T, Blye J, Boguslavskiy L, Bonnet C, Boukhgalter B, Brown A, Cahill P, Calixte N, Camarata J, Cheshatsang Y, Chu J, Citroen M, Collymore A, Cooke P, Dawoe T, Daza R, Decktor K, DeGray S, Dhargay N, Dooley K, Dorje P, Dorjee K, Dorris L, Duffey N, Dupes A, Egbiremolen O, Elong R, Falk J, Farina A, Faro S, Ferguson D, Ferreira P, Fisher S, FitzGerald M, Foley K, Foley C, Franke A, Friedrich D, Gage D, Garber M, Gearin G, Giannoukos G, Goode T, Goyette A, Graham J, Grandbois E, Gyaltsen K, Hafez N, Hagopian D, Hagos B, Hall J, Healy C, Hegarty R, Honan T, Horn A, Houde N, Hughes L, Hunnicutt L, Husby M, Jester B, Jones C, Kamat A, Kanga B, Kells C, Khazanovich D, Kieu AC, Kisner P, Kumar M, Lance K, Landers T, Lara M, Lee W, Leger JP, Lennon N, Leuper L, LeVine S, Liu J, Liu X, Lokyitsang Y, Lokyitsang T, Lui A, Macdonald J, Major J, Marabella R, Maru K, Matthews C, McDonough S, Mehta T, Meldrim J, Melnikov A, Meneus L, Mihalev A, Mihova T, Miller K, Mittelman R, Mlenga V, Mulrain L, Munson G, Navidi A, Naylor J, Nguyen T, Nguyen N, Nguyen C, Nicol R, Norbu N, Norbu C, Novod N, Nyima T, Olandt P, O'Neill B, O'Neill K, Osman S, Oyono L, Patti C, Perrin D, Phunkhang P, Pierre F, Priest M, Rachupka A, Raghuraman S, Rameau R, Ray V, Raymond C, Rege F, Rise C, Rogers J, Rogov P, Sahalie J, Settipalli S, Sharpe T, Shea T, Sheehan M, Sherpa N, Shi J, Shih D, Sloan J, Smith C, Sparrow T, Stalker J, Stange-Thomann N, Stavropoulos S, Stone C, Stone S, Sykes S, Tchuinga P, Tenzing P, Tesfaye S, Thoulutsang D, Thoulutsang Y, Topham K, Topping I, Tsamla T, Vassiliev H, Venkataraman V, Vo A, Wangchuk T, Wangdi T, Weiand M, Wilkinson J, Wilson A, Yadav S, Yang S, Yang X, Young G, Yu Q, Zainoun J, Zembek L, Zimmer A and Lander ES 2005 Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature **438**(7069)**:**803-19.

61. Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A, Jurka J, Kamal M, Mauceli E, Searle SM, Sharpe T, Baker ML, Batzer MA, Benos PV, Belov K, Clamp M, Cook A, Cuff J,

Das R, Davidow L, Deakin JE, Fazzari MJ, Glass JL, Grabherr M, Greally JM, Gu W, Hore TA, Huttley GA, Kleber M, Jirtle RL, Koina E, Lee JT, Mahony S, Marra MA, Miller RD, Nicholls RD, Oda M, Papenfuss AT, Parra ZE, Pollock DD, Ray DA, Schein JE, Speed TP, Thompson K, VandeBerg JL, Wade CM, Walker JA, Waters PD, Webber C, Weidman JR, Xie X, Zody MC, Graves JA, Ponting CP, Breen M, Samollow PB, Lander ES, and Lindblad-Toh K 2007 Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature **447**(7141)**:**167-77.

62. 2004 Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432**(7018)**:**695-716.
63. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MD, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJ, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, and Brenner S 2002 Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science **297**(5585)**:**1301-10.
64. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biemont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigo R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quetier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, and Roest Crollius H 2004 Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature **431**(7011)**:**946-57.
65. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pedersen JS, Hsu F, Hinrichs AS, Harte RA, Diekhans M, Clawson H, Bejerano G, Barber GP, Baertsch R, Haussler D, and Kent WJ 2007 The UCSC genome browser database: update 2007. Nucleic Acids Res **35**(Database issue)**:**D668-73.
66. Margulies EH, Maduro VV, Thomas PJ, Tomkins JP, Amemiya CT, Luo M, and Green ED 2005 Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. Proc Natl Acad Sci U S A **102**(9)**:**3354-9.
67. Levy S and Hannenhalli S 2002 Identification of transcription factor binding sites in the human genome sequence. Mamm Genome **13**(9)**:**510-4.
68. Liu Y, Liu XS, Wei L, Altman RB, and Batzoglou S 2004 Eukaryotic regulatory element conservation analysis and identification using comparative genomics. Genome Res **14**(3)**:**451-8.

69. Sauer T, Shelest E, and Wingender E 2006 Evaluating phylogenetic footprinting for human-rodent comparisons. Bioinformatics **22**(4)**:**430-7.

70. Bulyk ML 2003 Computational prediction of transcription-factor binding site locations. Genome Biol **5**(1)**:**201.

71. Mahony S, Corcoran DL, Feingold E, and Benos PV 2007 Regulatory conservation of protein coding and microRNA genes in vertebrates: lessons from the opossum genome. Genome Biol **8**(5)**:**R84.

72. Zhang C, Xuan Z, Otto S, Hover JR, McCorkle SR, Mandel G, and Zhang MQ 2006 A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. Nucleic Acids Res **34**(8)**:**2238-46.

73. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, and Young RA 2002 Transcriptional regulatory networks in Saccharomyces cerevisiae. Science **298**(5594)**:**799-804.

74. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich JM, Beyne E, Bleykasten C, Boisrame A, Boyer J, Cattolico L, Confanioleri F, De Daruvar A, Despons L, Fabre E, Fairhead C, Ferry-Dumazet H, Groppi A, Hantraye F, Hennequin C, Jauniaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Muller H, Nicaud JM, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richard GF, Straub ML, Suleau A, Swennen D, Tekaia F, Wesolowski-Louvel M, Westhof E, Wirth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P, and Souciet JL 2004 Genome evolution in yeasts. Nature **430**(6995)**:**35-44.

75. Jareborg N, Birney E, and Durbin R 1999 Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res **9**(9)**:**815-24.

76. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, and Miller W 2003 Human-mouse alignments with BLASTZ. Genome Res **13**(1)**:**103-7.

77. Zhang X, Odom DT, Koo SH, Conkright MD, Canettieri G, Best J, Chen H, Jenner R, Herbolsheimer E, Jacobsen E, Kadam S, Ecker JR, Emerson B, Hogenesch JB, Unterman T, Young RA, and Montminy M 2005 Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues. Proc Natl Acad Sci U S A **102**(12)**:**4459-64.

78. Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, Volkert TL, Schreiber J, Rolfe PA, Gifford DK, Fraenkel E, Bell GI, and Young RA 2004 Control of pancreas and liver gene expression by HNF transcription factors. Science **303**(5662)**:**1378-81.

79. Hong M ZG, Rozowsky JS, Gerstein MB, and Snyder MP 2007 ENCODE ChIP-chip for Rel A (p65) using polyclonal antibody (N-) on TNF-alpha-stimulated HeLaS3 cells. *unpublished*

80. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaoz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Loytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Xu M, Haidar JN, Yu Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrimsdottir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC,

Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B and de Jong PJ 2007 Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature **447**(7146)**:**799-816.

81. Cooper SJ, Trinklein ND, Nguyen L, and Myers RM 2007 Serum response factor binding sites differ in three human cell types. Genome Res **17**(2)**:**136-44.
82. Bruce C KP, Euskirchen G, Zhang Z, Rozowsky JS, Gerstein MB, and Snyder MP 2005 ENCODE ChIP-chip for JUN on human Hela S3 cells. *unpublished*
83. Kim J BA, Iyer V, Singer M, Jiang N, and Green R 2006 E2F4 and Myc ChIP-chip performed on Nimblegen hg17 ENCODE arrays. *unpublished*
84. Jin VX, O'Geen H, Iyengar S, Green R, and Farnham PJ 2007 Identification of an OCT4 and SRY regulatory module using integrated computational and experimental genomics approaches. Genome Res **17**(6)**:**807-17.
85. SGD Project "Saccharomyces Genome Database." http://www.yeastgenome.org/ (July 2007)
86. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, and Young RA 2004 Transcriptional regulatory code of a eukaryotic genome. Nature **431**(7004)**:**99-104.
87. Mahony S, Benos PV, Smith TJ, and Golden A 2006 Self-organizing neural networks to support the discovery of DNA-binding motifs. Neural Netw **19**(6-7)**:**950-62.
88. R Development Core Team 2006 R: a language and environment for statistical computing. http://www.R-project.org
89. Jin P, Zarnescu DC, Ceman S, Nakamoto M, Mowrey J, Jongens TA, Nelson DL, Moses K, and Warren ST 2004 Biochemical and genetic interaction between the fragile X mental retardation protein and the microRNA pathway. Nat Neurosci **7**(2)**:**113-7.
90. Shalgi R, Lieber D, Oren M, and Pilpel Y 2007 Global and local architecture of the mammalian microRNA-transcription factor regulatory network. PLoS Comput Biol **3**(7)**:**e131.
91. Liu N, Williams AH, Kim Y, McAnally J, Bezprozvannaya S, Sutherland LB, Richardson JA, Bassel-Duby R, and Olson EN 2007 An intragenic MEF2-dependent enhancer directs muscle-specific expression of microRNAs 1 and 133. Proc Natl Acad Sci U S A **104**(52)**:**20844-9.
92. Megraw M, Baev V, Rusinov V, Jensen ST, Kalantidis K, and Hatzigeorgiou AG 2006 MicroRNA promoter element discovery in Arabidopsis. RNA **12**(9)**:**1612-9.
93. Saini HK, Griffiths-Jones S, and Enright AJ 2007 Genomic analysis of human microRNA transcripts. Proc Natl Acad Sci U S A **104**(45)**:**17719-24.
94. Prestridge DS 1995 Predicting Pol II promoter sequences using transcription factor binding sites. J Mol Biol **249**(5)**:**923-32.

95. Ioshikhes IP and Zhang MQ 2000 Large-scale human promoter mapping using CpG islands. Nat Genet **26**(1)**:**61-3.

96. Ohler U, Niemann H, Liao G, and Rubin GM 2001 Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. Bioinformatics **17 Suppl 1:**S199-206.

97. Down TA and Hubbard TJ 2002 Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res **12**(3)**:**458-61.

98. Zhao X, Xuan Z, and Zhang MQ 2007 Boosting with stumps for predicting transcription start sites. Genome Biol **8**(2)**:**R17.

99. Buck MJ, Nobel AB, and Lieb JD 2005 ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. Genome Biol **6**(11)**:**R97.

100. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, and Myers RM 2006 Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome Res **16**(1)**:**1-10.

101. Taylor J, Tyekucheva S, King DC, Hardison RC, Miller W, and Chiaromonte F 2006 ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements. Genome Res **16**(12)**:**1596-604.

102. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schonbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K,

Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J and Hayashizaki Y 2005 The transcriptional landscape of the mammalian genome. Science **309**(5740)**:**1559-63.

103. Prestridge DS and Burks C 1993 The density of transcriptional elements in promoter and non-promoter sequences. Hum Mol Genet **2**(9)**:**1449-53.

104. Bussemaker HJ, Li H, and Siggia ED 2000 Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. Proc Natl Acad Sci U S A **97**(18)**:**10096-100.

105. Rigoutsos I and Floratos A 1998 Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics **14**(1)**:**55-67.

106. van Helden J, Andre B, and Collado-Vides J 1998 Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol **281**(5)**:**827-42.

107. Wang T and Stormo GD 2005 Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. Proc Natl Acad Sci U S A **102**(48)**:**17400-5.

108. Gangal R and Sharma P 2005 Human pol II promoter prediction: time series descriptors and machine learning. Nucleic Acids Res **33**(4)**:**1332-6.

109. Loots GG, Ovcharenko I, Pachter L, Dubchak I, and Rubin EM 2002 rVista for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res **12**(5)**:**832-9.

110. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr., and Haussler D 2000 Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A **97**(1)**:**262-7.

111. Chapelle O, Haffner P, and Vapnik VN 1999 Support vector machines for histogram-based image classification. IEEE Trans Neural Netw **10**(5)**:**1055-64.

112. Abeel T, Saeys Y, Bonnet E, Rouze P, and Van de Peer Y 2008 Generic eukaryotic core promoter prediction using structural features of DNA. Genome Res **18**(2)**:**310-23.

113. Schmid CD, Perier R, Praz V, and Bucher P 2006 EPD in its twentieth year: towards complete promoter coverage of selected model organisms. Nucleic Acids Res **34**(Database issue)**:**D82-5.

114. Bird AP 1980 DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res **8**(7)**:**1499-504.

115. Bird A 1999 DNA methylation de novo. Science **286**(5448)**:**2287-8.

116. Razin A and Cedar H 1991 DNA methylation and gene expression. Microbiol Rev **55**(3)**:**451-8.

117. Levine A, Cantoni GL, and Razin A 1992 Methylation in the preinitiation domain suppresses gene transcription by an indirect mechanism. Proc Natl Acad Sci U S A **89**(21)**:**10119-23.

118. Saxonov S, Berg P, and Brutlag DL 2006 A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A **103**(5)**:**1412-7.

119. Wakaguri H, Yamashita R, Suzuki Y, Sugano S, and Nakai K 2008 DBTSS: database of transcription start sites, progress report 2008. Nucleic Acids Res **36**(Database issue)**:**D97-101.

120. O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, and Mendell JT 2005 c-Myc-regulated microRNAs modulate E2F1 expression. Nature **435**(7043)**:**839-43.

121. Zhao Y, Samal E, and Srivastava D 2005 Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. Nature **436**(7048)**:**214-20.

122. Sylvestre Y, De Guire V, Querido E, Mukhopadhyay UK, Bourdeau V, Major F, Ferbeyre G, and Chartrand P 2007 An E2F/miR-20a autoregulatory feedback loop. J Biol Chem **282**(4)**:**2135-43.

123. Lee J, Li Z, Brower-Sinning R, and John B 2007 Regulatory circuit of human microRNA biogenesis. PLoS Comput Biol **3**(4)**:**e67.

124. Tsang J, Zhu J, and van Oudenaarden A 2007 MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. Mol Cell **26**(5)**:**753-67.

125. Bottinger EP and Bitzer M 2002 TGF-beta signaling in renal disease. J Am Soc Nephrol **13**(10)**:**2600-10.

126. Willis BC and Borok Z 2007 TGF-beta-induced EMT: mechanisms and implications for fibrotic lung disease. Am J Physiol Lung Cell Mol Physiol **293**(3)**:**L525-34.

127. Dupont S, Zacchigna L, Adorno M, Soligo S, Volpin D, Piccolo S, and Cordenonsi M 2004 Convergence of p53 and TGF-beta signaling networks. Cancer Lett **213**(2)**:**129-38.

128. Moustakas A and Heldin CH 2005 Non-Smad TGF-beta signals. J Cell Sci **118**(Pt 16)**:**3573-84.

129. Ranganathan P, Agrawal A, Bhushan R, Chavalmane AK, Kalathur RK, Takahashi T, and Kondaiah P 2007 Expression profiling of genes regulated by TGF-beta: differential regulation in normal and tumour cells. BMC Genomics **8:**98.

130. Kasai H, Allen JT, Mason RM, Kamimura T, and Zhang Z 2005 TGF-beta1 induces human alveolar epithelial to mesenchymal cell transition (EMT). Respir Res **6:**56.

131. Thuault S, Valcourt U, Petersen M, Manfioletti G, Heldin CH, and Moustakas A 2006 Transforming growth factor-beta employs HMGA2 to elicit epithelial-mesenchymal transition. J Cell Biol **174**(2)**:**175-83.

132. Willis BC, Liebler JM, Luby-Phelps K, Nicholson AG, Crandall ED, du Bois RM, and Borok Z 2005 Induction of epithelial-mesenchymal transition in alveolar epithelial cells by transforming growth factor-beta1: potential role in idiopathic pulmonary fibrosis. Am J Pathol **166**(5)**:**1321-32.

133. Fan JM, Ng YY, Hill PA, Nikolic-Paterson DJ, Mu W, Atkins RC, and Lan HY 1999 Transforming growth factor-beta regulates tubular epithelial-myofibroblast transdifferentiation in vitro. Kidney Int **56**(4)**:**1455-67.

134. Sgarra R, Rustighi A, Tessari MA, Di Bernardo J, Altamura S, Fusco A, Manfioletti G, and Giancotti V 2004 Nuclear phosphoproteins HMGA and their relationship with chromatin structure and cancer. FEBS Lett **574**(1-3)**:**1-8.
135. Zuo F, Kaminski N, Eugui E, Allard J, Yakhini Z, Ben-Dor A, Lollini L, Morris D, Kim Y, DeLustro B, Sheppard D, Pardo A, Selman M, and Heller RA 2002 Gene expression analysis reveals matrilysin as a key regulator of pulmonary fibrosis in mice and humans. Proc Natl Acad Sci U S A **99**(9)**:**6292-7.
136. Selman M, Pardo A, and Kaminski N 2008 Idiopathic pulmonary fibrosis: aberrant recapitulation of developmental programs? PLoS Med **5**(3)**:**e62.
137. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, and Rajewsky N 2005 Combinatorial microRNA target predictions. Nat Genet **37**(5)**:**495-500.
138. Park SM, Gaur AB, Lengyel E, and Peter ME 2008 The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. Genes Dev **22**(7)**:**894-907.
139. Orlando V 2000 Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. Trends Biochem Sci **25**(3)**:**99-104.
140. Andrews NC and Faller DV 1991 A rapid micropreparation technique for extraction of DNA-binding proteins from limiting numbers of mammalian cells. Nucleic Acids Res **19**(9)**:**2499.