Evolution of YY1, YY2, REX1 and DNA-binding motifs in vertebrate genomes

Christopher Don Faulk
Louisiana State University and Agricultural and Mechanical College
EVOLUTION OF YY1, YY2, REX1 AND DNA-BINDING MOTIFS IN VERTEBRATE GENOMES

A Dissertation

Submitted to the Graduate Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in

The Department of Biological Sciences

by

Christopher Don Faulk
B.A., Louisiana State University, 2001
May 2010
ACKNOWLEDGEMENTS

Beyond a simple list of those to whom I owe debts and gratitude, I must say that I did not do this on my own. I thank Dr. Leslie Kozak for encouraging me to attend graduate school and Dr. Thomas Moore for eventually letting me in. Once here I found an advisor under whose mentoring style I not only flourished but wish to emulate in the future. Dr. Joomyeong Kim saw in me the ability to become a good scientist, despite my achievements on paper. Without his tremendous knowledge, patience in the lab, and consistent good cheer, I would not have completed this program. Doubtless I owe him a great debt and will endeavor to pay it forward by producing good science and mentoring my own students in the future.

Too often committee members are underappreciated, and certainly I will not break that tradition. Dr. Donze’s coursework changed my mind regarding my area of specialization; he is a good friend as well as advisor. Dr. Whitehead made me earn every bit of credit that I can claim as my own. And Dr. Batzer has been no less than a father to me ever since I joined LSU. I also thank Dr. Masami Yoshimura for his enthusiastic questions and comments. My unofficial committee has many members: Drs. Kurt Svoboda, Greg Pettis, and John Larkin were instrumental in my admittance to the program and have given me useful advice throughout my years here. Hana Kim, Muhammad Ekram and Drs. Jennifer Huang, JeongDo Kim, and Sungryul Yu worked alongside me in the lab, reminding me that not every experiment works, and sometimes you need coffee. Josh Meyer, Deepa Srikanta, Charles Huang, Shurjo Sen, and Jeanine Braud made coffee a group event; in matters both personal and professional they are models for how problems can be solved creatively and not always with a pipette.

Lastly, I thank my family. Donald and Aleta Faulk did the best job possible while raising me and continue to be there for me providing support in innumerable ways. I must just say thank you and accept that they understand how much. My sister Constance Faulk forever makes me laugh. I hope to make them proud.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................ ii

LIST OF TABLES ................................................................................................................ iv

LIST OF FIGURES ............................................................................................................. v

ABSTRACT .......................................................................................................................... vi

CHAPTER ONE: BACKGROUND .......................................................................................... 1

CHAPTER TWO:  
RETROPOSITION AND EVOLUTION OF THE DNA-BINDING  
MOTIFS OF YY1, YY2 AND REX1 .................................................................................. 12

CHAPTER THREE:  
YY1’S DNA-BINDING MOTIFS IN MAMMALIAN OLFACTORY  
RECEPTOR GENES ............................................................................................................ 39

CHAPTER FOUR:  
POSITION-BASED CLUSTERING OF MICROARRAY  
EXPRESSION DATA ............................................................................................................ 59

CHAPTER FIVE: SUMMARY .................................................................................................. 71

APPENDIX: LETTERS OF PERMISSION ............................................................................ 78

VITA ..................................................................................................................................... 81
LIST OF TABLES

3.1 Dipeptide frequency in olfactory receptor proteins .......................................................... 48
## LIST OF FIGURES

1.1 Duplication of YY1 via mRNA mediated retroposition ........................................ 5

1.2 The 10 base pair consensus sequence of the YY1 binding motif .............................. 6

2.1 Global alignment of YY1, YY2, REX1, PHO and PHOL ........................................ 16

2.2 Gene trees connecting (A) PHO and PHOL and (B) YY1, YY2 and REX1 ............... 18

2.3 Exon structures of (A) PHO and PHOL and (B) YY1, YY2 and REX1 .................. 20

2.4 Sequence alignment of the zinc finger domains of YY1-related sequences ............... 21

2.5 DNA-binding motifs of YY1, YY2 and REX1 ..................................................... 24

2.6 Gel shift assays of DNA-binding motifs of (A) YY1, (B) YY2 and (C) REX1 .......... 27

3.1 YY1 binding sites in genomic context .................................................................. 44

3.2 Histogram of genes with increasing YY1 density ................................................. 45

3.3 YY1 DNA-binding motifs as translated dipeptides .............................................. 47

3.4 Correlation of observed and expected dipeptides and statistical significance .......... 49

3.5 YY1 binding sites within the ORFs of the olfactory receptor and histone 4 gene families ............................................................................................................. 50

4.1 Schema for generating position-based data vs. traditional method ......................... 61

4.2 Note on fold change calculation .......................................................................... 62

4.3 Heat map of position-based expression .................................................................. 68

4.4 Four clusters of unusual response ......................................................................... 69
ABSTRACT

Transcription factors are important for many aspects of gene regulation in eukaryotes. YY1 (Yin-Yang 1) is a particularly interesting example of a highly conserved zinc-finger transcription factor, involved in transcriptional activation, repression, initiation, and in chromatin modification. YY1 is ubiquitously expressed in mammals, and its binding sites are found in ~10% of human genes as well as in repetitive elements. It is a targeting protein of the Polycomb complex and is involved in mammalian genomic imprinting.

First, we explored the evolutionary history of YY1 using 62 species and formation of its paralogs, YY2 and REX1, which are found in mammals, and Pho and Phol, which are found in Drosophila. We confirmed the specificity of the consensus YY1 binding site and the differences of the target binding motifs of YY2 and REX1 which are reflected in their amino acid sequences. We found that the core motif, CCAT, is conserved for all three homologs and that YY2 and REX1 were produced via retrotransposition events early in the mammalian lineage.

Second, we identified unusual clusters of YY1-binding motifs found in the coding regions of olfactory receptor genes (OLFRs) in mammals but not in fish. Olfactory genes provide scent detection and are the largest class of genes in mammals. Statistical analysis indicates that the core of the YY1-binding motifs cannot be accounted for by conserved amino acid motifs or overall protein homology. Thus selection has acted at the DNA level rather than at the protein level in preserving these YY1-binding sites within coding regions. Therefore, YY1 is likely to play a crucial role in regulating the expression of OLFRs.

Third, we produced a new method of microarray data analysis predicated on the positions of genes along a chromosome as well as their expression levels. This technique is supplementary to traditional microarray data analysis and adds a new dimension to finding target genes of interest by looking for co-regulation.
Overall, this work provides a coherent background to the evolution of YY1 and its homologs. It provides strong evidence that coding sequences of genes can encode information both at the DNA level and the protein level.
CHAPTER ONE:

BACKGROUND
Transcription Factors

Eukaryotic genomes consist of two major types of genes: one type exhibits constitutive expression and the other with tissue-specific and/or stage-specific expression. Most of the tissue-specific and stage-specific expression patterns are mediated by transcription factors, i.e., proteins which recognize a specific pattern of nucleotides in the genomic DNA and attach to these “motifs” (Farnham, 2009). The trans-activating domain functions in protein-protein binding. Transcription factors then recruit other proteins for a number of tasks: activation, initiation, repression, DNA methylation, or chromatin modification. Approximately 10% of the human protein-coding genes encode transcription factors, many of which affect the expression of large numbers of downstream genes (Lander et al., 2001; Venter et al., 2001). Transcription factors are trans-regulatory elements; that is, they act on distant genes regardless of the position of the gene which encodes the transcription factor. Since they can act globally on many other genes, a small change in the function or binding site specificity of a single transcription factor can have large changes on the development or health of an organism by affecting the regulation of many downstream genes (Nowick et al., 2009). This provides a pathway for rapid evolution of a population since the initial coding changes are small, yet the modified binding may affect all of the downstream co-regulated genes. Alternatively, some genes such as Yin-Yang-1 (YY1) have evolved such important roles in the life cycle of many organisms that changes to their structure, activity, or copy number are extremely deleterious and have been strongly selected against (Shi et al., 1997). Hence, some transcription factors have undergone rapid duplication and divergence in different lineages while others have remained practically unchanged in amino acid sequence and copy number for hundreds of millions of years. YY1, which we discuss in chapter two, has very few duplicate copies, though interestingly, two of them are mammal-specific.
YY1 Evolution

Charles Darwin noted in 1859 that organisms appear to have replicated parts which diverge in function (Darwin, 1859). Over 100 years later Susumu Ohno posited gene duplication and divergence as the main process by which eukaryotic organisms obtain new genes (Ohno, 1970). Genes can undergo duplication by several methods including unequal crossing over, whole genome duplication, gene conversion, and RNA-mediated retroposition (Taylor and Raes, 2004). In humans, gene duplication can have profound effects on our health (Mefford and Eichler, 2009). Though most of these new copies are thought to rapidly pseudogenize and become non-functional, they are free of the purifying selection imposed upon the extant original copy and can undergo divergence and adopt new functions. Transcription factors tend to have dosage-sensitive effects and have been preferentially retained after whole genome duplications (Edger and Pires, 2009). Similarly, this retention applies to specific domains within a group of paralogous genes (Bornberg-Bauer et al., 2005). In chapter two we describe the evolutionary history of YY1 and its homologs and the domains which comprise them.

The YY1 family is ancient, having homologs in all studied animal species. In flying insects, the gene is known as Pleiohomeotic (PHO), while the Drosophila genus also harbors a duplicate copy, Pho-like (PHOL). Most vertebrate genomes have a single copy of YY1. Fish genomes maintain two active copies of YY1 which resulted from an ancient genome-wide duplication event and the two copies are nearly identical. Mammals possess two additional independently retroposed copies, Yin-Yang 2 and Reduced Expression 1 (REX1, also known as zfp42) (Fig 1.1). YY1 was discovered independently in 1991 by Shi and Shenk as YY1 and also by Park and Atchison as NF-E1 and was found to be the human ortholog of the mouse δ protein (Hariharan et al., 1991; Park and Atchison, 1991; Shi et al., 1991). It is a C2H2 Gli-Krüppel zinc-finger protein with four zinc fingers that serve to bind to the major groove of DNA in a
YY2 was discovered in 2004 and, like YY1, acts as a repressor and activator (Nguyen et al., 2004). YY2, while functionally similar to YY1, is unusual in that it is located within the intron of another gene, membrane-bound transcription factor protease, site 2 (MBTPS2) on the X-chromosome (Luo et al., 2006). The zinc-finger domain and its corresponding binding site in YY2 are very similar to those of YY1, leading to a similar target motif (Nguyen et al., 2004). REX1’s sequence similarity to YY1 also implies its formation from YY1. Compared to YY2, REX1’s DNA-binding domain has a greater number of amino-acid changes from YY1 which is reflected in its changed binding specificity, and its protein-protein binding domain has significant differences as well. REX1 was originally discovered by Gudas in 1989 and was characterized as a gene whose expression was reduced during retinoic acid induction of differentiation in F9 teratocarcinoma stem cells (Hosler et al., 1989). However, it is dispensable for pluripotency and its deletion is not lethal in mice (Masui et al., 2008). REX1 is highly studied as a stem cell marker and as such, has implications for human health (Mongan et al., 2006).

**YY1 Function**

The name YY1 stands for Yin-Yang 1 which reflects its ability to both activate and repress gene expression. Initially it was shown to be a repressor of the adeno-associated virus by attaching to the P5 promoter region, yet also act as an activator in the same system when exposed to E1A oncogene protein (Shi et al., 1997). It is constitutively expressed in vertebrates and is essential for viability as well as being implicated in the biology of cancer (Gordon et al., 2006). Approximately 10% of human genes have YY1 binding sites within their promoters and YY1 has a critical role in the expression of these genes (Schug et al., 2005). Additionally, YY1 acts as a surveillance gene, silencing the transcription of repetitive elements.
Figure 1.1 Duplication of YY1 via mRNA mediated retroposition. YY1 has duplicated twice in the mammalian lineage resulting in the paralogs YY2 and REX1. These copies all reside on different chromosomes and the coding regions are made up of a single exon. Both bear the features of genes which have duplicated by retroposition through an mRNA intermediate. Chromosome numbers refer to the human genome.

Interestingly, YY1 binding activity is also methyl sensitive, meaning that binding of YY1 to its target motif is reduced in the presence of a 5-methyl cytosine (Kim et al., 2003).

The N-terminal end of YY1 contains several protein-protein interaction domains which recruit additional transcription factors, histone modifying enzymes, and Polycomb Repression Complexes (PRCs). These serve to modify the chromatin structure and activate or repress gene activity at the level of transcription. The C-terminal end is required for nuclear localization which fits with YY1’s role in cell-cycle progression (Gordon et al., 2006). YY1 functions to
regulate genomic imprinting, repetitive element repression, viral infection, apoptosis, and cancer (Calame and Atchison, 2007; Kim and Kim, 2007; Kim et al., 2007; Kim, 2008).

**YY1 Binding Motifs**

Like nearly all transcription factors, YY1 has a degenerate target motif. The consensus motif is 10 base pairs in length (CCGCCATnTT); however, the core of the binding site (CCAT) is absolutely required for binding (Fig 1.2). Some genes have been shown to contain a longer YY1 binding motif, allowing enhanced binding when compared to the typical 10 base pair motif (Kim, 2009). The initial CG dinucleotide can be methylated which greatly reduces the binding affinity of YY1 protein to the target motif.

![Figure 1.2. The 10 base pair consensus sequence of the YY1 binding motif.](image)

This WebLogo represents the proportion of each possible base found at every position within the 10 base pair YY1 binding motif. The core, CCAT, is invariable and required for YY1 binding affinity.

Recent work has shown that YY1 binding sites are found within the first intron of the YY1 gene itself and are active in controlling the expression of the protein; it is effectively autoregulated (Kim et al., 2009). Some YY1 binding sites are found clustered near imprinting control regions where they have been maintained at high density in mammals (Kim et al., 2006).

In chapter three we describe the enrichment of YY1 binding sites found in the olfactory receptor genes of mammals. The radiation of mammals coincides with a large increase in the number of olfactory receptor genes and requires a conserved regulatory mechanism (Buck and Axel, 1991; Niimura and Nei, 2006). Olfactory receptor genes are known to contain regulatory
sequences within their coding regions, but the identity of these sequences is unknown (Merriam and Chess, 2007). We analyzed coding regions of the genome for YY1 binding motifs under the assumption that DNA can encode information not only for proteins but also for regulation (Itzkovitz and Alon, 2007). Statistical analysis suggests that the high density of YY1 binding motifs within the coding regions of the olfactory receptor genes is maintained as part of a regulatory mechanism for this class of genes. As predicted, the enrichment of YY1 binding motifs is independent of selection pressure at the amino acid level. Intriguingly, fish have no such enrichment of YY1 binding sites in their more limited number of olfactory receptor genes suggesting that this is a tetrapod innovation.

**Microarray Analysis**

In chapter four we present an auxiliary method to analyze microarrays. Traditionally, microarray data interpretation has focused on the change in gene transcript levels with different experimental conditions (Trevino et al., 2007). The challenge inherent in microarray analysis is with the very large amount of data which is produced and not analyzed since only a small number of genes with large changes in expression are typically studied (Quackenbush, 2006). Because genes are located in physical proximity along a chromosome, they can be subject to coordinated regulation by epigenetic marks. To efficiently search for position-based regulation, we developed a protocol which can be applied to many types of microarray data, often publicly available in databases such as the Gene Expression Omnibus. Using this protocol researchers can create a “heat map” of gene expression in which genes are aligned by chromosome position and a color assigned based upon the magnitude of expression change. Previous methods to detect regions of low but consistent fold change are more complex and thus unsuited to widespread use. Our position-based clustering protocol provides a straightforward method to identify new targets of interest.
References


CHAPTER TWO:

RETROPOSITION AND EVOLUTION OF THE DNA-BINDING MOTIFS OF YY1, YY2 AND REX1*

*Reprinted by permission of Nucleic Acids Research
Introduction

The transcription factor YY1 (Yin Yang 1) is a Gli-Krüppel type zinc finger protein, and can function as a repressor, activator or transcription initiator depending upon the sequence context of YY1-binding sites with respect to other regulator elements (Shi et al., 1997). The protein has a DNA-binding domain at the C-terminus and other modulating domains at the N-terminus displaying repression, activation and protein/protein interaction activities (Thomas and Seto, 1999). YY1 interacts with several key transcription factors, including TBP, TAFs, TFIIB and Sp1, as well as histone-modifying complexes, such as p300, HDACs, PRMT1 and Polycomb complexes (Thomas and Seto, 1999; Gordon et al., 2006). Many cellular and viral genes are controlled by YY1. A recent survey estimated that 10% of human genes contain YY1 binding sites near their promoter regions (Schug et al., 2005). Another set of studies has revealed that some of mammalian imprinted genes contain very unusual tandem arrays of YY1 binding sites in their controlling regions, suggesting potential roles in mammalian genomic imprinting (Kim et al., 2003; Kim et al., 2006; Kim et al., 2007). A series of mouse mutagenesis experiments demonstrated the dosage-dependent essential roles of YY1 during mouse development as well as in cell cycle control (Donohoe et al., 1999; Affar el et al., 2006).

YY1 is evolutionarily well conserved throughout the vertebrate and invertebrate lineages. It has been identified in several vertebrate species (Shi et al., 1997; Satijn et al., 2001; Sui et al., 2004), and two genes very similar to YY1 are found even in flies, Pleiohomeotic (PHO) and Pho-like (PHOL) (Brown et al., 1998; Brown et al., 2003). PHO is one of the DNA-targeting proteins for the Polycomb complex and the phenotypes of pho-deficient mutants can be rescued by mammalian YY1 (Srinivasan et al., 2005). In mammalian genomes, two other YY1-related genes have been identified, YY2 (Yin Yang 2) and Reduced Expression 1 (REX1). YY2 is functionally very similar to YY1 (Nguyen et al., 2004), and is a retroposed copy duplicated from
YY1 based on its intronless structure and location in the intron of another X-chromosomal gene, Mbtps2 (Luo et al., 2006). REX1 was independently discovered, before the identification of YY1, due to its unique expression profile: dramatic decline of expression after retinoic acid-induced differentiation of F9 murine teratocarcinoma stem cells (Hosler et al., 1989). Subsequently, REX1 has been mainly studied as a stem cell marker that is controlled by Oct3/4 (Hosler et al., 1993; Ben-Shushan et al., 1998). A recent comparative study, however, emphasized that REX1 is a member of the YY1 subfamily (Mongan et al., 2006).

Despite the significant roles and evolutionary conservation of YY1-related sequences in animals, there has not been any systematic analysis of these sequences in terms of their origins, evolutionary patterns and implications for functional diversification. To address this, we have analyzed YY1-related sequences identified from genome sequences ranging from flying insects to placental mammals. We have identified two evolutionarily conserved protein domains within YY1 which were previously unrecognized. We have uncovered independent retroposition events that have been responsible for forming duplicate copies, such as PHOL from PHO in flies, and YY2 and REX1 from YY1 in placental mammals. Our analyses revealed that the zinc finger domains of YY2 and REX1 have been under different selection pressures compared to YY1. Their DNA-binding properties have evolved from YY1 by weakening DNA-binding affinity in both YY2 and REX1, and changing DNA-binding motifs in REX1. The evolution patterns of YY1 and other YY1-related genes described in the current study provide a unique paradigm for gene duplication and functional diversification.

Results

Identification of YY1-related Sequences from Invertebrates and Vertebrates

The protein sequence of human YY1 (GenBank accession no. NP_0034941, 414 amino acid long) was used to search databases to identify YY1-related sequences from all available
genome sequences. One YY1 homolog, known as PHO, was identified from each of the flying insects, including mosquitoes, honeybees, beetles and 10 different species of flies. In flies, a similar sequence, known as PHOL, was identified from each of the 10 different fly species. This totals to 23 different YY1-related sequences from insects. Database searches identified 39 different YY1-related sequences in chordates, ranging from urochordates (sea squirts) to placental mammals: one each from sea squirts and purple sea urchins, six from fish, one from frog, one from chicken, 29 from mammals. In fish, two copies of YY1 sequences were identified from each of three sequenced genomes, zebrafish, puffer fish and spotted pufferfish whereas three copies of YY1-related sequences were identified from each placental mammal. Database searches have identified a total of 62 YY1-related sequences. Based on sequence similarity, these are categorized into five groups: the PHO and PHOL groups from flying insects, the YY1 group from vertebrates, and the YY2 and REX1 groups from placental mammals (Fig. 2.1). Individual sequences and other related information are available through the following website (http://JooKimLab.lsu.edu/JooKimLab/Data.html).

Comparison of the amino acid sequences derived from the YY1-related sequences identified three evolutionarily conserved protein domains (Fig. 2.1). These include two domains in the middle of the protein (amino acid position 203–226 and 250–281 in the human YY1, respectively), and one DNA-binding zinc finger domain at the C-terminus (aa 298–414). The two domains in the middle, Domains I and II, are located within the region previously known as the Spacer between several N-terminal domains and the C-terminal DNA-binding domain (Thomas and Seto, 1999).

These two domains are found in the YY1-related sequences of most, but not all, vertebrates and insects. In flies, only Domain I is found in both PHO and PHOL sequences.
Figure 2.1. Global alignment of YY1, YY2, REX1, PHO and PHOL. Protein sequences derived from 62 YY1-related sequences are aligned using the ClustalW program. The zoom-out view of this result is shown for global representation. The actual sequence alignment is available as Supplementary Data 2 through the following website (http://JooKimLab.lsu.edu/JooKimLab/Data.html). Each row represents one individual sequence and these sequences are categorized into different groups indicated by parentheses on the left. Different amino acids are represented by different background colors, and thus a vertical line with the same color indicates the conservation (or identical amino acid residue) at that position amongst all the sequences analyzed. The different levels of evolutionary conservation throughout the entire region of YY1-related sequences are represented by a graph underneath the alignment. Three regions are evolutionarily conserved, and thus highlighted by underlines with arrows. These include Domain I and II that are located within the previously defined Spacer region, and the DNA-binding zinc finger domain at the C-terminus. The protein domain structure of human YY1 is shown as a reference at the bottom (Thomas and Seto, 1999).

In placental mammals, two domains are found in YY2, but only Domain II is found in the REX1 sequences. However, the zinc finger domain is found in all the YY1-related sequences with high levels of sequence conservation, ranging from 66 to 100% similarity. The relative positions of these three protein domains are also conserved among all the identified sequences. The conservation of these three domains in the YY1-related sequences suggests that these three domains constitute the original domain structure of the YY1 protein.
**Retroposition-mediated YY1 Duplications in Flies and Placental Mammals**

Several lineages have more than one copy of YY1-related sequences, including flies, fish and placental mammals. Two copies of YY1-related sequences, PHO and PHOL, are found in all the fly species examined to date while only one copy, PHO, is found in the other flying insects. This suggests a gene duplication unique to the fly lineage. According to the results of phylogenetic analyses (Fig. 2.2A), the topology of the two gene trees corresponding to the PHO and PHOL groups in flies is very similar to that of the known species tree of the fruitfly genus *Drosophila*, indicating that this gene duplication predates the radiation of all fly species. The PHO sequences of the other flying insects show slightly greater levels of sequence similarity to the PHO rather than PHOL sequences in flies, suggesting that PHO is the original sequence that gave rise to the duplicated copy PHOL. This is further confirmed by the different exon structures of PHO and PHOL (Fig. 2.3A). The coding region of PHO is split into five exons, and a similar split exon structure is also found in the PHO of other insects, such as beetles and honeybees. In contrast, the entire coding region of PHOL is located within one exon, an intronless structure of its coding region. This intronless genomic structure is usually observed in the sequences that have been duplicated through an RNA-mediated mechanism, retroposition, by which processed mRNAs are reverse-transcribed and transposed to other genomic loci without introns in germ cells (Brosius, 2003). These data therefore indicate that PHOL has been duplicated from PHO through retroposition.

The two copies of YY1 sequences found in the fish lineage show an almost identical sequence and exon structure to each other (data not shown). Chromosome-wide duplications are known to have been prevalent at the early stage of the fish genome evolution (Taylor et al., 2003; Christoffels et al., 2004). Therefore, the two copies of YY1 present in each fish genome are
Figure 2.2. Gene trees connecting (A) PHO and PHOL and (B) YY1, YY2 and REX1. Alignments were first created using a subset of sequences, the protein sequences of which are available or can be predicted with certainty. Later, the trees were constructed with the neighbor-joining method using the Mega3 program. In each tree, the bootstrap values calculated from 1000 replicates are indicated above each branch. The trees constructed with the maximum parsimony method are also available as Supplementary Data 6 at http://JooKimLab.lsu.edu.
thought to be another outcome of this chromosome-wide duplication event. In contrast, the two additional copies in placental mammals, YY2 and REX1, show quite different evolution patterns. First, like PHOL, the coding regions of both YY2 and REX1 are also located within one exon while the YY1 genes of all vertebrates show a very similar split exon structure with five coding exons (Fig. 2.3B). This suggests that both YY2 and REX1 were also duplicated from YY1 by retroposition. The detection of YY2 and REX1 exclusively in placental mammals further suggests relatively recent formation of these two copies during mammalian evolution with the estimated time being about 60–100 million years ago. In mammals, both YY2 and REX1 are transcribed and maintain their Open Reading Frames (ORFs), confirming the functionality of these two retroposed copies. Second, despite this recent origin, inter-species sequence divergence levels of YY2 and REX1 are much greater than those of YY1, as reflected on the phylogenetic tree shown in figure 2.2B. Very low levels of sequence divergence are observed between all the YY1 sequences of different vertebrates whereas each sequence from the YY2 and REX1 groups exhibits average 20% divergence between different species. This indicates relaxation of evolutionary constraints on both the YY2 and REX1 genes. As compared to REX1, YY2 displays greater levels of similarity to YY1 in terms of its overall sequence and protein domain structure, suggesting that the retroposition of YY2 may have occurred in more recent times than that of REX1. Pairwise sequence comparison also revealed that both YY2 and REX1 share higher sequence identity with YY1 than each other (Supplementary Data 3), suggesting that both REX1 and YY2 have been independently derived from YY1. The presence of two conserved domains, Domains I and II, in YY2 also supports the idea that YY2 has been derived from YY1, not from REX1, since REX1 has only Domain I. Overall, exon structure and sequence conservation levels suggest that the two retroposed copies, YY2 and REX1, have been under different levels of functional constraints than the original gene, YY1.
Figure 2.3. Exon structures of (A) PHO and PHOL and (B) YY1, YY2 and REX1. The protein coding regions of PHO and YY1 both are split into five different exons depicted by boxes. Three conserved domains are marked by different colors: green for Domain I, blue for Domain II and gray for the zinc finger domain. This multi-exonic structure of both PHO and YY1 is conserved throughout all studied lineages. In contrast, the entire coding region of each PHOL, YY2 and REX1 is localized within one exon, suggesting the retroposition-driven formation of these duplicates in both the fly and placental mammal lineages. This retroposition-mediated duplication also resulted in the different chromosomal positions among these duplicates as shown in the right column.

**Different Selection Pressures on the DNA-binding Domains of YY1, YY2 and REX1**

All the YY1-related sequences show very unusual levels of sequence conservation in the DNA-binding domain of the predicted proteins (Fig. 2.4). The zinc finger domains of PHO and PHOL from all of the different fly species share 5 and 18 amino acid differences, respectively, as compared to those of vertebrate YY1. The zinc finger domains of the other flying insects, however, show an almost identical sequence to those of vertebrate YY1. Thus, the observed amino acid differences in flies represent the substitutions that had occurred in the fly lineage.
Apparently, the overall consensus sequence of flying insects’ PHO is still identical to that of vertebrate YY1. Similarly, the zinc finger domains of vertebrates’ YY1 also do not show any shared substitution except for one or two species-specific amino acid changes. Thus, YY1 is believed to have maintained its DNA-binding domain without any amino acid changes in the past 600 million year period, representing one of the most extreme cases for functional selection imposed on an eukaryote gene.

**Figure 2.4. Sequence alignment of the zinc finger domains of YY1-related sequences.** The zinc finger domains of YY1-related sequences are compared with that of human YY1 (aa 298–411). The amino acid residues identical to human YY1 are indicated by dashes (–). The residues that differ from human YY1 are indicated by the single letter amino acid code with colors: green for conservative substitution and red for non-conservative substitution. Several entries underneath the YY1 sequence correspond to the sequences from different lineages. The entries in the middle and bottom of the alignment represent the YY2 and REX1 sequences, respectively. The amino acid residues known to contact directly with the bases of target DNAs are indicated on the last row. The duplex sequence of a known YY1 target DNA from the Adeno-Associated Virus (AAV) P5 promoter is shown along with contacting amino acid residues, which are indicated by the single letter code with the amino acid position information based on human YY1 sequence.
As described earlier, YY2 and REX1 have been under different levels of evolutionary constraints since their formation in placental mammals. This is in stark contrast to the extreme conservation of YY1. The zinc finger domains of different species’ YY2 protein show an average of 6–11 amino acid differences as compared to that of YY1 (Fig. 2.4). None of these changes are shared among different mammals, indicating that these changes represent independent substitutions that occurred in each species. Similarly, the zinc finger domains of different species’ REX1 proteins also show an average of 11–20 amino acid differences between each other, implying a slightly higher level of relaxation of evolutionary constraint on REX1. As compared to vertebrate YY1, however, the zinc finger domains of all REX1 sequences share 8 amino acid substitutions (Fig. 2.4). These substitutions represent the changes that occurred and were fixed before the radiation of eutherian mammals. The sudden fixation of these substitutions might be an evolutionary remnant suggesting positive selection that might have occurred in the early stages of REX1 evolution, although our analyses point toward purifying selection with relaxed constraints for the REX1 evolution (Supplementary Data 5). Interestingly, most of these changes are localized within Fingers 1 and 4, and are also non-conservative amino acid substitutions from the original amino acid residues of YY1. In particular, the amino acid change T398N in Finger 4 is localized within the region known to contact directly with the bases of target DNAs (Houbaviy et al., 1996). Therefore, this change along with other amino acid substitutions in REX1 may have a functional outcome possibly allowing REX1 to bind to DNA motifs divergent from the YY1 DNA-binding motif. Similarly, the amino acid substitutions within YY2 also appear to be slightly more frequent in Fingers 1 and 4, suggesting the presence of different selection pressures on each zinc finger. However, none of YY2 changes appear to be located within critical regions for its DNA binding, predicting no major difference between the DNA binding motifs between YY1 and YY2.
DNA-binding Motifs of YY1, YY2 and REX1

We have further investigated the functional consequences of different selection pressures imposed on the zinc finger domains of YY1, YY2 and REX1 by characterizing their DNA-binding motifs. For this experiment, the zinc finger domain of each protein was subcloned into the downstream region of the GST protein, expressed as part of a fusion protein in bacteria, fixed on agarose beads, and finally we allowed them to bind to duplex DNAs derived from randomized oligonucleotide sequences (4n=15). After five rounds of selection, the bound DNAs were subcloned and sequenced (Fig. 2.5). In the case of YY1, 20 of 34 bound DNAs contain DNA motifs that have either a perfect match or 1 base difference from the known YY1 consensus sequence. All of the remaining 14 bound DNAs still show an almost identical sequence as YY1 but have an average of two base differences from YY1. Our approach used only the zinc finger domain of YY1, but most of the bound DNAs are identical to the known consensus sequence of YY1. This confirms the modular nature of the zinc finger domain of YY1, and subsequently the feasibility of this approach.

In the case of YY2, 16 of 46 sequences contain DNA motifs similar to the YY1 consensus sequence. As with the YY1 fusion protein, the remaining sequences also contain a motif similar to YY1 with two base differences, confirming our initial prediction: there is no major difference between YY1 and YY2 motifs. Interestingly, however, most of the YY2-bound sequences have more than two binding motifs within the randomized portion of each sequence. About half of the bound sequences show two motifs in an opposite orientation, with the other half in the same orientation. In contrast to YY2, the DNAs bound by REX1 seem to be slightly different from those bound by either YY1 or YY2. The sequences bound by REX1 can be divided into two groups. These two groups can be represented by two slightly different
Figure 2.5. DNA-binding motifs of YY1, YY2 and REX1. The sequences of DNAs bound by YY1 (left), YY2 (middle) and REX1 (right) are shown with the clone numbers on the right. The uppercase sequences are derived from the randomized portion of the input DNAs for binding whereas the lowercase dinucleotides represent the surrounding, fixed portion of the input DNAs. The majority of the DNAs bound by both YY1 and YY2 contain the known YY1 consensus motif (CGCCAT.TT), which is marked blue in the forward direction and by red in the reverse direction. The DNAs bound by REX1 are divided into two groups: one group indicated by blue and the other by bold-type. The total number of analyzed DNA molecules for each individual protein is indicated inside the parenthesis. For YY1 and YY2, the first number corresponds to the number of bound DNAs with either a perfect match or one base difference, while the second number to bound DNAs with more than two base differences.

Consensus sequences (Fig. 2.5): Type 1 (5'-GGCAGCCATTA-3') and Type 2 (5'-GGCCATTA-3'). The consensus sequences of these two groups differ by the presence (or absence) of three bases (GGC) at the 5'-side. These two consensus sequences also show one unique difference at their 3'-side final position: all the DNAs bound by REX1 contain A instead of T. This is consistent with the amino acid change detected in the critical DNA binding region of REX1,
T388N in figure 2.4. Despite these changes, the core sequences of the YY2 and REX1 binding motifs are still the same as that of YY1 (5'-CCAT-3'), suggesting that the conservation of the two fingers, Fingers 2 and 3, may be responsible for maintaining a similar core target motif among these three genes.

**Gel Shift Assays of DNA-binding Motifs of YY1, YY2 and REX1**

The DNA-binding motifs of YY1, YY2 and REX1 were further analyzed using gel shift assays (Fig. 2.6). In the case of YY1, we have used the same set of duplex oligonucleotides used in a previous study to demonstrate the subtle but unique property of YY1, methylation-sensitive DNA-binding (Kim et al., 2003). As expected, the DNA-binding domain, as part of the GST-YY1 fusion protein, showed an almost identical pattern of DNA binding as endogenous YY1 protein (Fig. 2.6A). The GST-YY1 protein is methylation-sensitive: methylation on the upper strand is inhibitory to the binding (Fig. 2.6A, Lanes 1–4). One base change in this CpG site, either CpA or TpG, somewhat reduced the affinity of the YY1 binding, but still allowed YY1 binding to these probes (Fig. 2.6A, Lanes 5–6). The DNA-binding domain of YY2 also showed a similar pattern of DNA binding: methylation-sensitive binding and subtle effects by single base changes caused by the CpG site (Fig. 2.6B). However, the DNA-binding affinity of YY2 is much weaker than YY1 based on the results derived from our control experiments for gel shift assays (Supplementary Data 4). We have also tested some of the DNAs that contain two motifs within the randomized portion of the target DNAs (Fig. 2.5B). We did not observe any difference in binding between the duplex DNAs with two binding motifs versus single binding motif (data not shown). Overall, the DNA-binding patterns of YY1 and YY2 appear to be similar except for the fact that the binding affinity of YY2 is much weaker than YY1, consistent with the observed relaxation of evolutionary constraint on the DNA-binding domain of YY2.
Several sets of gel shift assays were performed for the identified DNA-binding motifs of REX1 (Upper panel in Fig. 2.6C). The REX1 and YY1 fusion proteins were individually allowed to bind to seven duplex probes. These include three consensus motifs, the consensus of YY1 (Probe 7), the consensus of REX1 Type 1 (Probe 3) and Type 2 (Probe 4). We have also included four other probes containing one or two base variations from the three consensus motifs to further dissect the binding specificity of REX1 and YY1. The REX1 protein bound to the four probes containing REX1 motifs (Probes 1–4), but not to the YY1 or related probes (Probes 5–7). On the other hand, the binding of the YY1 protein to the REX1 probes was detected but very marginal compared to its binding to the YY1 or related probes (Probes 5–7). This indicates the different binding specificity between the YY1 and REX1 proteins. This different binding specificity is originated from three key differences found in the REX1 binding motifs as compared to the YY1 binding motifs. First, the REX1 motifs have A instead of T at the 8th position of the YY1 consensus (CGCCATNTT). This change reduced dramatically the binding affinity of the YY1 protein, but increased the binding affinity of the REX1 protein (Probe 4 versus Probe 5). Second, the REX1 motifs do not show any base preference at the 9th position of the YY1 consensus (CGCCATNTT). Interestingly, the T base at this position reduced slightly the binding affinity of the REX1 protein, but is required for the binding of the YY1 protein (Probe 1 versus Probe 2). Third, one of the REX1 motifs contains additional three bases (5'-GGC-3’) at the 5'-side of its sequence. The addition of these three bases reduced the binding affinity of the YY1 protein, but increased the affinity of the REX1 protein (Probe 2 versus Probe 6). The significance of these key differences was further demonstrated by competition assays using three representative probes (Lower panel in Fig. 2.6C). Overall, these data clearly demonstrate the different binding specificity between the YY1 and REX1 proteins, and also
Figure 2.6. Gel shift assays of DNA-binding motifs of (A) YY1, (B) YY2 and (C) REX1. Identified DNA-binding motifs were further confirmed with gel shift assays using three fusion proteins. For the gel shift assays of the YY1 and YY2 fusion proteins, we have used a set of the six different duplex probes that have been previously used for testing the methylation-sensitive DNA-binding activity of endogenous YY1 (Kim et al., 2003). Four different probes have an identical sequence as the CSE2 probe containing one YY1 binding site indicated by an underline. However, their methylation status at the CpG site is different: u (−/−), unmethylated on both strands; hm (+/−), methylated on the upper strand; hm (−/+), methylated on the bottom strand; and m (+/+), methylated on both strands. For the DNA-binding motif studies of (C) REX1, we have used seven probes: the YY1 consensus motif probe (Probe 7), and two YY1-related probes with one base difference (Probe 6&5), the Type 2 and 1 motifs of REX1 (Probe 4&3), and two variants of the Type 2 motif (Probe 2&1). The REX1 and YY1 proteins were individually used for the left and right sets of gel shift assays, respectively (Upper panel). Three representative probes were also used for competition assays (Lower panel). One minor band below the REX1 protein is from non-specific binding by an unidentified Escherichia coli protein in crude extracts. The sequences of these probes are shown on the bottom, and the relevant binding motifs within these sequences are bold-typed.
prove that the two identified motifs, Types 1 and 2, represent bona fide DNA-binding motifs for REX1. The positions of these three critical base differences in the surrounding regions of the core motif (5'-CCAT-3') are consistent with an observed evolution pattern (Fig. 2.4), differential selection pressures on each of the four zinc finger units of the REX1 protein.

Discussion

In the current study, we have analyzed all the YY1-related sequences identified from genome sequences of invertebrates and vertebrates. We have identified two other protein domains, besides the zinc finger domain, that are conserved throughout all the YY1 and YY1-related sequences. Our analyses also confirmed that independent retroposition events have been responsible for forming duplicated copies, such as PHOL from PHO in flies, and YY2 and REX1 from YY1 in placental mammals. The zinc finger domains of YY2 and REX1 have been under different selection pressures than YY1, and consequently their DNA-binding properties have evolved from those of YY1 by weakening DNA-binding affinity in YY2 and REX1, and changing DNA-binding motifs in REX1. The evolution patterns of YY1 and other YY1-related proteins appear to be unique in several regards, as discussed subsequently.

Besides the zinc finger domain, two other protein domains, Domains I and II, are evolutionarily well conserved throughout all the YY1-related sequences ranging from flying insects to mammals (Fig. 2.1). The conservation of these two domains is somewhat less obvious within the sequences of flies, but the detection of these domains within the PHO sequences of honeybees and beetles undoubtedly indicates that these two domains are part of the original domains of YY1. Database searches with these two domains did not find any proteins other than YY1 or YY1-related sequences, suggesting that these two domains are unique to YY1-related sequences (data not shown). According to previous studies analyzing protein–protein interactions, the Spacer region, a relatively large region of YY1 (aa 201–298 in human YY1)
encompassing these two domains, is responsible for the interaction with the viral oncoprotein E1A and the p53-interacting partner Hdm2 (Shi et al., 1991; Sui et al., 2004). It should be interesting to test whether YY2 and REX1 also interact with the above two proteins. Nevertheless, the functional roles played by these two domains are predicted to be essential for YY1 functions based on their conservation in most of the YY1-related sequences.

There are several key transcription factors with similar evolutionary ages as YY1, such as Sp1 and the E2F family of proteins. These transcription factors have increased their gene copy numbers along with the increase of complexity and genome size of animals (Carroll et al., 2001; Davidson, 2001), but the duplication of these genes has been mainly driven by DNA-mediated mechanisms involving the entire genomic fragments surrounding individual genes (Dynlacht et al., 1994; Kaczynski et al., 2003). That is, in the Sp1 and E2F families, the whole gene structure has been duplicated with exons, introns and promoters intact. In the case of YY1, however, retroposition has been the primary mechanism for its duplication: PHOL duplication from PHO, and YY2 and REX1 duplications from YY1 (Fig. 2.3), which is quite different from the general duplication mode observed in other key transcription factors. A gene copy duplicated through retroposition is subject to transcriptional controls different from those of its original gene due to its random insertions at other genomic regions. As an outcome, the duplicate copy tends to show different expression patterns compared with its original gene. Consistently, both YY2 and REX1 also display expression patterns quite different from that of YY1. As compared to the ubiquitous expression patterns of YY1, YY2 shows more germ cell-specific expression patterns (Luo et al., 2006), and REX1 exhibits stem cell-specific expression (Mongan et al., 2006). It is still puzzling why YY1 duplication has been driven by retroposition, but different expression patterns resulting from this duplication mode may have been one major factor contributing to the success of YY1 duplications in placental mammals.
The evolutionary patterns observed with YY2 and REX1 are quite different from that of YY1 (Fig. 2.4). YY1 shows high levels of sequence conservation throughout its coding region. In particular, the zinc finger domain of YY1 has maintained its amino acid sequence without any changes in the past 600-million year period, implying that the YY1 homologs, insect PHO and vertebrate YY1, may still bind to similar DNA motifs. This turns out to be the case based on DNA-binding motif studies (Brown et al., 1998; Brown et al., 2003). In contrast, the zinc finger domains of YY2 and REX1 show much higher levels of inter-species sequence divergence, suggesting relaxed constraints on their DNA-binding domains. Consequently, both YY2 and REX1 display much weaker DNA-binding affinity than YY1 (Fig. 2.6 and Supplementary Data 4). The loosened DNA-binding affinities of YY2 and REX1 may have allowed these duplicates to bind to slightly different binding motifs, as seen in REX1 (Fig. 2.5), and subsequently to bind to new sets of downstream genes. Together different expression patterns, loosened affinities and different DNA-binding motifs may have contributed to the functional diversification of the two duplicates, YY2 and REX1, in the mammalian lineage.

Successful gene duplication is still regarded as a rare evolutionary event (Kondrashov and Kondrashov, 2006), which is further supported by the single-copy status of YY1 in the majority of animal lineages. Then, what could be the main reason(s) underlying the sudden formation of two YY1 duplicates in placental mammals? This may be indirectly answered by observations drawn from other gene duplicates in mammals. For instance, DNMT3L is a member of the DNA methyltransferase family, which is found only in mammals (Yokomine et al., 2006). Yet, DNMT3L has been found to be involved in genomic imprinting (Bourc'his et al., 2001), a gene dosage control mechanism unique to placental mammals (Reik and Lewis, 2005). CTCFL (or BORIS), a mammal-specific duplicate of the vertebrate insulator protein CTCF (Loukinov et al., 2002), might be involved in establishing the gametic imprinting mark of DNA
methylation for H19 during germ cell development (Jelinic et al., 2006). In both cases, gene duplicates appear to play specific roles in mammal lineage-specific novelties, such as genomic imprinting and epigenetic modification. These two duplicates, interestingly, share some similarities with the YY1 duplicates, YY2 and REX1, such as recent formation, rapid evolution, lineage-specific conservation in mammals and germ cell-specific expression (Bestor and Bourc'his, 2004). Furthermore, recent studies suggest that several imprinted domains may be controlled by YY1 or related transcription factors (Kim et al., 2006; Kim et al., 2007). This entices the speculation that both YY2 and REX1 may be also involved in novel placental mammal-specific functions, such as genomic imprinting. This idea needs to be tested, but the evolutionary patterns presented in this study clearly indicate the tight linkage of both YY2 and REX1 to the biology of placental mammals.

Materials and Methods

Database Search and Sequence Analyses

A series of database searches were conducted using the BLAST program (http://www.ncbi.nlm.gov/BLAST) to obtain YY1-related sequences. Human YY1 (NP_003394.1) was first used as a query sequence to search sequence databases, including NCBI, the Genome Browser at University of California Santa Cruz and Ensembl. Later, human REX1 (NP_777560.2) and YY2 (NP_996806.1) were used to further characterize the identified YY1-related sequences from chordates, while Drosophila melanogaster PHO (NP_524630.1) and PHOL (NP_648317.1) were used for the identified insect sequences. The detailed information regarding all the YY1-related sequences described in this study is available as Supplementary Data 1 through the following website (http://JooKimLab.lsu.edu/JooKimLab/Data.html).
Multiple sequence alignments were performed with ClustalW using the following parameters: gap opening penalty = 10, gap extension penalty = 0.1 (0.2 for multiple alignment), Gonnet Protein Weight Matrix, residue specific penalties = ON, hydrophilic penalties = ON, gap separation distance = 4, end gap separation = OFF (Thompson et al., 1994). Sequences were edited manually in Mega3 V3.1 to remove spurious introns from some sequences (Kumar et al., 2004). Separate multiple alignments were performed for insects’ and chordates’ sequences. Subsequently, two phylogeny gene trees were constructed and analyzed using both the neighbor-joining and maximum parsimony methods as implemented in Mega3 V3.1 with Poisson correction and confirmed by bootstrapping 1000 iterations (Saitou and Nei, 1987). Synonymous and non-synonymous substitution rates were estimated using two different approaches: Nei–Gojobori (Nei and Gojobori, 1986) and Yang–Nielsen methods (Yang and Nielsen, 2000).

**Expression of Fusion Proteins and DNA-binding Motif Study**

The zinc finger regions of YY1 (NM_009537.2), YY2 (NM_178266) and REX1 (NM_009556.2) were amplified from either mouse brain cDNAs or genomic DNAs by the following primer sets: YY1 (mYY1Zn5, 5'-CCAAGAACATAGCTTGCCCTC-3' and mYY1Zn3, 5'-TCACTGGTTGTTTTTTGGCTTTAGCG-3'), YY2 (mYY2Zn5, 5'-CCAAGACCTAGCATGCTCTC-3' and mYY2Zn3, 5'-TTACTGGTCATTCTT-3'), YY3 (mYY3Zn5, 5'-TTATCGATGCTGGAGTGTCCTCAAGC-3' and mYY3Zn3, 5'-TTACTGGTCTATCGACCATGCTCCTC-3'). The amplified products were first cloned into the pCR4-TOPO vector (Invitrogen, Carlsbad, CA, USA), and later transferred to the EcoRI site of the pGEX-4T-2 vector (Amersham Biosciences, Piscataway, NJ, USA) after sequence confirmation. The constructed vectors were transformed into BL21 (DE3) competent cells for bacterial expression (Strategen, La Jolla, CA, USA). The optimum induction of the constructs by
IPTG was monitored through SDS-PAGE (Supplementary Data 4 from http://JooKimLab.lsu.edu/JooKimLab/Data.html).

DNA-binding motif studies were conducted as described in the previous studies (Hyde-DeRuyscher et al., 1995; Yant et al., 1995) with slight modifications. Briefly, the transformed cells were grown at 37°C in LB media (100 ml) to an optical density of 0.6 at 600 nm, and protein expression was induced with 0.4 mM IPTG for additional 3.5 h. Cells were harvested by centrifugation at 4000g for 10 min at 4°C. Lysates were prepared from the cell pellets by sonication in 6 ml of ice-cold NETN buffer (100 mM NaCl, 1 mM EDTA, 20 mM Tris–HCl, pH 8.0, 0.5% NP-40). Protein concentration in cell lysates was determined using the Bradford assay (Pierce, Rockford, IL, USA). Aliquots of 500 µg/100 µl were frozen at –80°C.

Immobilized glutathione agarose (Pierce) was washed three times with 1 ml ice-cold NETN buffer and used to isolate fusion proteins by incubating 500 µg lysate with 50 µl washed agarose beads at 4°C for 30 min while rotating. The agarose beads were precipitated by centrifugation, and washed twice first with 1 ml ice-cold NTEN buffer and later with 1 ml 1x binding buffer (12 mM HEPES, pH 7.9, 60 mM KCl, 5 mM MgCl2, 1 mM DTT, 0.5 mM EDTA, 0.05% NP-40, 50 µg/ml bovine serum albumin, 10% glycerol). The final pellet was resuspended in 100 µl 1x binding buffer. Randomized duplex DNAs were prepared with PCR using following oligonucleotides (10 ng of NT55, 5’-
CTGTCGGAATTCGCTGACGT(N)15CGTCTTATCGGATCCTACGT-3’, 0.1 µg of UpNt, 5’-
CTGTCGGAATTCGCTGACGT-3’ and 0.1 µg of DwNt, 5’-ACGTAGGAT CCGATAAGACG-3’ as a template and primers for extension reaction, respectively). Duplex DNAs were labeled 10 µCi [α-32P] dATP for the easy chase of the bound DNAs with the PCR reaction containing 5 U of i-StarTaq DNA polymerase (Intron Biotech), 0.2 mM each of dGTP, dTTP and dCTP and 10 µM dATP. PCR was performed for 25 cycles (95°C 30 s; 65°C 1 min; 72°C 1
The labeled DNAs were allowed to bind to the fusion protein immobilized on the agarose beads at room temperature for 30 min with rotation. The bound DNAs were washed three times with 1 ml of 1x binding buffer, eluted by phenol: chloroform extraction, and finally precipitated ethanol. The eluted DNAs were amplified again with the same conditions described earlier for another round of DNA-binding. The following PCRs were performed only for 10 cycles. After five rounds of DNA-binding and amplification (Supplementary Data 4), the DNAs were subcloned into pCR4-TOPO vector (Invitrogen). For each fusion protein, 40–60 clones were purified and sequenced.

**Gel Shift Assay of DNA-binding Motifs**

The identified DNA motifs for each fusion protein were further analyzed with gel shift assays (Gel shift Assay System, Promega, Madison, WI, USA). About 10–20 µg of each fusion protein, as part of the cell lysate, was used for each experiment with the [{gamma}-32P] ATP-labeled duplex probes prepared from the following oligonucleotides: CSE2-A, 5'-CCCACCCACCTGGGCGCCATCTTTAATGAAAG-3', and CSE2-B, 5'-CTTTCATTTAAAGATGGGCGCCATCTTTAATGAAAG-3'; 2a-A, 5'-CCCACCCACCTGGGCGCCATCTTTAATGAAAG-3', and 2a-B, 5'-CTTTCATTTAAAGATGGGCGCCATCTTTAATGAAAG-3'; 2b-A, 5'-CCCACCCACCTGGGCGCCATCTTTAATGAAAG-3', and 2b-B, 5'-CTTTCATTTAAAGATGGGCGCCATCTTTAATGAAAG-3'; Probe1-A, 5'-CTTTCATTTAAAGATGGGCGCCATCTTTAATGAAAG-3', and Probe1-B, 5'-CTTTCATTTAAAGATGGGCGCCATCTTTAATGAAAG-3'; Probe2-A, 5'-GATAAGACGCACCATTTGGAACGTACACACG-3', and Probe2-B, 5'-CGCTGACGTTCCAAATGGCTGCCCGTCTTATC-3'; Probe3-A, 5'-GATAAGACGCACCATTTGGAACGTACACACG-3', and Probe3-B, 5'-CGCTGACGTTCCAAATGGCTGCCCGTCTTATC-3';
CGCTGACGTTCTAATGGCTGCGCGTCTTATC-3'; Probe4-A, 5'-
GATAAGACGGCCATTATGAGGCCCACGTGACCG-3', and Probe4-B, 5'-
CGCTGACGTGGCCTCATAATGGCCGTCTTATC-3'; Probe5-A, 5'-
GATAAGACGGCCATTATGAGGCCCACGTGACCG-3', and Probe5-B, 5'-
CGCTGACGTGGCCTCATAATGGCCGTCTTATC-3'; Probe6-A, 5'-
GATAAGACAGCCATTTTGAGGCCCACGTGACCG-3', and Probe6-B, 5'-
CGCTGACGTGGGCTTCAAAATGGCCGTCTTATC-3'; Probe7-A, 5'-
GATAAGACAGCCATTTTGAGGCCCACGTGACCG-3', and Probe7-B, 5'-
CGCTGACGTGGGCTTCAAAATGGCCGTCTTATC-3'. To monitor our gel shift assays, we also performed a set of control experiments using endogenous YY1 from HeLa nuclear extracts (Promega).

References


CHAPTER THREE:

YY1'S DNA-BINDING MOTIFS IN MAMMALIAN OLFACTORY RECEPTOR GENES*

*Reprinted by permission of BioMed Central Genomics
Introduction

The transcription factor YY1 is a Gli-Krüppel type zinc finger protein that is highly conserved from insects through vertebrates (Kim et al., 2007). YY1 can function as an activator, repressor, or initiator depending upon the other regulatory elements in the region (Shi et al., 1997). YY1 also interacts with a variety of proteins including components of RNA polymerase II complex, transcription factors, and histone-modifying complexes (Thomas and Seto, 1999; Gordon et al., 2006; Wilkinson et al., 2006). According to genome-wide surveys, about 10% of all human genes contain YY1 binding motifs in their promoter regions (Schug et al., 2005). Functionally, YY1 is involved in many biological processes, including embryonic development, cell cycle progression, apoptosis, B cell development, Polycomb group Gene (PcG)-mediated repression, genomic imprinting, and X chromosomal inactivation (Shi et al., 1997; Sui et al., 2004; Gordon et al., 2006). YY1 was also initially identified as a factor controlling the transcriptional activity of the murine retrotransposon 'Intracisternal A Particle' (Satyamoorthy et al., 1993). Since then, many retroposons, including SINE, LINE, and endogenous retrovirus families, have been shown to contain YY1 binding sites in their promoter regions (Thomas and Seto, 1999; Gordon et al., 2006). Due to this ubiquitous presence of YY1 binding sites in genome-wide repeats, YY1 has also been regarded as a surveillance gene that is responsible for repressing transcriptional background noise from these repeats (Humphrey et al., 1996).

The olfactory receptor (OLFR) genes of mammals encode short, single coding exon, G protein-coupled receptors that are responsible for sensing a large number of airborne scents (Buck and Axel, 1991). This gene family is comprised of over 800 and 1,300 gene members in human and mouse respectively, forming the largest gene family in mammalian genomes (Buck and Axel, 1991; Zozulya et al., 2001; Zhang et al., 2004; Niimura and Nei, 2006). The aquatic vertebrates, the teleost fish lineage, also have a similar odorant receptor gene family (Alioto and
Ngai, 2005). However, the odorant receptor (OR) family of the fish lineage consists of a much smaller number of genes than that of the mammals, and these OR genes are also much more diverse in sequence identity than those of mammals. Mammalian OLFR genes are divided into Class I and Class II groups based on sequence identity (Kambere and Lane, 2007). Class II genes make up ~90% of OLFRs and are thought to have expanded during the transition to land-based living.

In mammals these olfactory receptors presumably expanded due to the selective advantage conferred by a well developed sense of smell (Kambere and Lane, 2007). While mice and other mammals retain function and expression of almost all OLFRs, the majority of these are pseudogenized in humans (Keller and Vosshall, 2008). The mammalian OLFR genes are highly tissue-specific and are expressed primarily in the olfactory epithelium though a subset expresses in a chemosensory role in other tissues such as kidney and sperm (Feldmesser et al., 2006; Spehr et al., 2006; Pluznick et al., 2009). Furthermore, only one copy (allele) out of all 1,000 OLFR genes is selected and expressed in each neuron cell of the olfactory tissue (Chess et al., 1994). The unusual transcriptional control of the OLFR gene family is likely mediated through unknown trans-acting factors (Shykind, 2005). The tissue-specific nature of their expression coupled with their widespread duplication requires a mitotically-stable global silencing mechanism in all cell types. According to recent studies, potential cis-regulatory elements recruiting these trans factors are hypothesized to be located within the protein- coding regions of the OLFR genes rather than their surrounding genomic regions (Merriam and Chess, 2007; Nguyen et al., 2007).

While performing genome-wide searches of the DNA- binding motifs of YY1, we discovered that the mammalian OLFR genes contain unusual clusters of YY1 binding sites within their protein-coding regions, whereas most YY1 binding sites are solitary and upstream of a regulated
gene. In this study, we further analyzed the significance of this discovery with several bioinformatic and statistical measures, which will be described below. Specifically we test whether the presence of the YY1 binding sites could be explained by DNA sequence or amino acid motif conservation.

Results

YY1 DNA-binding Motifs in the Mammalian OLFR Genes

YY1 is predicted to be a global epigenetic regulator based on its ubiquitous expression and interaction with many histone-modifying enzymes (Thomas and Seto, 1999). As part of the efforts exploring this possibility, we have performed several series of YY1 binding motif searches using the genome sequences of mammals (human, mouse, and cow) (Kim et al., 2006; Kang et al., 2009; Kim, 2009). Here we first scanned the genome sequences of human and mouse using a Position Weight Matrix (PWM)-based Perl script. Repetitive elements are known to contain YY1 binding sites so the RepeatMasked genome was used (Smit, 1996-2004). Later, the results of these searches were visualized using the repeat-masked Custom Track of the UCSC genome browser. While inspecting global localization patterns of YY1 binding motifs in each genome, we noticed that clusters of YY1 binding motifs are co-localized with the genomic regions harboring olfactory receptor (OLFR) genes. The mammalian OLFR genes show a single coding exon structure, and they are also localized as gene clusters in specific regions of mammalian chromosomes. One such example is shown using the 100-kb genomic region from Mmu 7 chromosome (Fig. 3.1). This figure shows a representative sample of 5 OLFR genes, and the locations of these genes coincide with those of YY1 binding motifs. Each OLFR gene appears to contain a range of 4 to 8 YY1 binding motifs within its 1-kb-long Open Reading Frame (ORF). As expected, the coding regions correlate well to the placental mammal conservation plot, a default track available on the UCSC genome browser. Also, the identified
YY1 binding motifs appear to be random in location within the coding regions but do show a bias in orientation with respect to the direction of OLFR gene transcription, which will be described later.

The unusual clustering of YY1 binding motifs within the protein-coding regions of the OLFR genes was further analyzed to test if this pattern is unique to only the OLFR genes or also found in the ORFs of other genes. For this analysis, the entire set of the mouse mRNA database was scanned with the PWM-based Perl program to identify YY1 binding motifs. The number of the identified YY1 binding motifs within a given DNA sequence corresponding to the transcript was further divided by the size of the mRNA sequence, yielding a YY1 density score. The mouse mRNA sequences (total number = 20,191) were subsequently binned based on their relative YY1 density scores (the X-axis on figure 3.2a). The Y-axes of figure 3.2 represents the number of genes within a given range of the YY1 density score. This analysis indicated that the majority of the non-olfactory mouse mRNA sequences (19,083 sequences) were distributed evenly and randomly throughout the varying ranges of the YY1 density score (0 to 0.114) (Fig. 3.2a). In contrast, our detailed inspection revealed that about half of the OLFR gene set (1,108) show very high YY1 density scores. To better visualize this unusual pattern, we separated only the OLFR gene set from the rest of the mRNA set, and derived another histogram (Fig. 3.2b). As shown in figure 3.2b, about 45% (or 496 of 1,108) of the OLFR genes have YY1 density scores ranging from 0.032 to 0.094, which are equivalent to 4 to 8 (or more) YY1 binding sites per a 1-kb mRNA sequence. In contrast, a scan of the 1kb upstream regions of 20,419 Refseq genes revealed that only 10% of these putative promoter containing regions had a density greater than 4 YY1 binding sites.

We also repeated the above analyses using the mRNA data sets derived from human, cow, and zebrafish transcriptomes to test the evolutionary conservation of this unusual clustering of
YY1 binding motifs within OLFR genes (Fig. 3.2c & 3.2d). The results from human and cow (data not shown) mRNA data sets also showed a pattern consistent with the mouse data set: an unusual clustering of YY1 binding motifs within many OLFR genes.

Figure 3.1. YY1 binding sites in genomic context. YY1 binding sites are mapped onto the Custom Track of the UCSC Genome Browser (http://genome.ucsc.edu/). YY1 binding sites are marked by blue and red bars indicating forward and reverse directions respectively. Height on the Y-axis represents the score of each YY1 binding site in relation to the consensus sequence. This 100 kb segment of the genome contains five olfactory receptor genes and shows association of several YY1 binding sites correlated with the exons of the genes. Mammalian conservation is shown in the plot at the bottom of the map. This map was created through scanning the RepeatMasked sequence with a scoring matrix representing the YY1 binding site consensus. Mammalian conservation correlates with exons as expected, but also with many YY1 binding sites in this region.

In humans we found 57% (or 215 of 380) of OLFRs contain more than 4 YY1 binding sites while percentage in cow rose to 82% (or 740 of 900). It is important to note that the total number of the human OLFR genes (380) in the figure is smaller than those of the other mammals since a large fraction of the human OLFR genes are known to have become pseudogenes in recent evolutionary times and we removed all genes annotated as hypothetical. In contrast, the OLFR genes of zebrafish do not show a similar pattern to mammals. The total 25 OLFR genes of zebrafish show as wide a range of the YY1 density scores as seen in the other non-OLFR genes.
These results confirm that the unusual clustering of YY1 binding motifs within the OLFR genes is a feature of the mammalian lineage.

**Figure 3.2. Histogram of genes with increasing YY1 density.** Each olfactory receptor gene was scanned for the number of YY1 binding sites found within its transcript. The combined score was divided by the gene length, resulting in the YY1 density for the given olfactory receptor gene. Note that the quality of the YY1 scores varies based on the match generated by the Position Weight Matrix, and thus the calculation of YY1 binding sites per 1 kb is approximate. Genes were ordered based on increasing YY1 density and the histograms were derived by the relative position of a gene within the group of all genes. A) The mouse non-olfactory genes binned by relative order of increasing YY1 density (x-axis). The y-axis indicates the number of genes in each bin. B) Mouse olfactory genes binned in the same fashion, also from their position within the group of all genes. The highest group of bins corresponds to a density of 4-11 YY1 binding sites per 1 kb transcript for mouse though only 6 genes show a YY1 density corresponding to more than 8 YY1 sites per kb. For human and mouse (panels B&C), there is a large group of olfactory receptor genes with high YY1 density, whereas fish (D) does not show this pattern.
**Statistical Tests for the Clustering of YY1 Binding Motifs within OLFR genes**

We performed two different series of analyses to test the functional significance of the observed clustering of YY1 binding motifs within the mammalian OLFR genes. First, we tested whether some peptide motifs enriched within the protein sequences of the OLFR genes are responsible for a spurious display of the YY1 binding motifs in the nucleotide sequences of the OLFR genes. If some peptide motifs are enriched then we expect a significant difference in the number of those motifs found in OLFR genes versus the rest of the proteome. For this test, we identified 49 individual dipeptides that can be encoded by the core motifs (CCAT or ATGG) of the YY1 binding consensus sequence, and determined if any of these dipeptides were over-represented in the protein sequences of the OLFR genes (Fig. 3.3). This global comparison did not immediately identify any dipeptides that are unusually enriched within the protein sequences of the OLFR genes. However, according to the detailed analyses using the Z-score values, two dipeptides (PL and AI) among the 49 dipeptides for the YY1 core motif showed significant enrichment ($P < 0.01$, Fig. 3.4b). Even so, this enrichment is limited to the 4 bp YY1 core binding site and not the complete 10 bp motif which we used to calculate YY1 density in OLFR genes. Furthermore, the reverse-translated sequences of these two peptides (PL and AI) have 24 and 12 fold degeneracy, respectively. This means that only one out of 24 PL or 12 AI dipeptides may contain the actual core motif of the YY1 binding consensus sequence. Therefore, the detection of 4 to 8 YY1 binding motifs per one OLFR gene cannot be simply accounted for by the serendipitous overlap between the YY1 core motif and a subset of codon combinations encoding frequent peptide motifs of the OLFR genes.

As a second measure, we carefully analyzed the positions of the identified YY1 binding motifs within the ORFs of the OLFR genes (Fig. 3.5). Conservation in location of YY1 binding
49* Possible Dipeptides Can Form a Core YY1 Binding Site

. . C C A T . .
Frame 1 15 Possible dipeptides
2 8
3 6
Total 29

. . A T G G . .
Frame 1 12 Possible dipeptides
2 4
3 5
Total 21

* The dipeptide ProTrp (PW) can be encoded by two different combinations of codons shown.

Figure 3.3. YY1 DNA-binding motifs as translated dipeptides. The four base pair core motif of the YY1 binding site is CCAT (forward) and ATGG (reverse). In principle, this core can code for two codons in each of the three frames. In the forward direction, CCAT can encode 29 possible dipeptides while in the reverse direction ATGG can encode 21 dipeptides, with one duplicate. These 49 dipeptides are a subset of the 400 possible dipeptides which can be found in a protein (20 total amino acids)². The distribution frequency of these dipeptides has been examined to determine if any of the 49 YY1 binding sites is overrepresented in olfactory receptor genes. We first calculated the frequencies of all 400 possible combinations of dipeptides using all other protein sequences except OLFR genes of the mouse to derive a reference set of the expected frequencies for the 400 dipeptides. In parallel, we also separately calculated the frequencies of the 400 dipeptides, including the 49 dipeptides that can encode the YY1 core motif, using only the protein sequences of the mouse OLFR genes. The frequencies derived from the OLFR genes (Observed Values) were compared with the reference set (Expected Values) via the Z-score to identify any over-represented dipeptides within the OLFR genes (Fig. 3.4a) (Table 3.1).
Table 3.1. Dipeptide frequency in olfactory receptor proteins*

<table>
<thead>
<tr>
<th>Dipeptide</th>
<th>All proteins</th>
<th>Olfactory (Observed)</th>
<th>Olfactory (Expected)</th>
<th>Frequency</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL</td>
<td>77943</td>
<td>3222</td>
<td>1894</td>
<td>1.70</td>
<td>3.72</td>
</tr>
<tr>
<td>AI</td>
<td>41587</td>
<td>2784</td>
<td>1010</td>
<td>2.76</td>
<td>3.01</td>
</tr>
<tr>
<td>TI</td>
<td>35087</td>
<td>2444</td>
<td>852</td>
<td>2.87</td>
<td>2.46</td>
</tr>
<tr>
<td>SI</td>
<td>48732</td>
<td>2110</td>
<td>1184</td>
<td>1.78</td>
<td>1.92</td>
</tr>
<tr>
<td>PM</td>
<td>15764</td>
<td>1971</td>
<td>383</td>
<td>5.15</td>
<td>1.69</td>
</tr>
<tr>
<td>LH</td>
<td>42086</td>
<td>1961</td>
<td>1022</td>
<td>1.92</td>
<td>1.68</td>
</tr>
<tr>
<td>MA</td>
<td>30513</td>
<td>1817</td>
<td>741</td>
<td>2.45</td>
<td>1.44</td>
</tr>
<tr>
<td>SH</td>
<td>34455</td>
<td>1798</td>
<td>837</td>
<td>2.15</td>
<td>1.41</td>
</tr>
<tr>
<td>YG</td>
<td>25865</td>
<td>1130</td>
<td>628</td>
<td>1.80</td>
<td>0.33</td>
</tr>
<tr>
<td>AM</td>
<td>22983</td>
<td>1120</td>
<td>558</td>
<td>2.01</td>
<td>0.31</td>
</tr>
<tr>
<td>PF</td>
<td>29419</td>
<td>1065</td>
<td>715</td>
<td>1.49</td>
<td>0.23</td>
</tr>
<tr>
<td>MV</td>
<td>22565</td>
<td>939</td>
<td>548</td>
<td>1.71</td>
<td>0.02</td>
</tr>
<tr>
<td>SM</td>
<td>26009</td>
<td>926</td>
<td>632</td>
<td>1.47</td>
<td>0.00</td>
</tr>
<tr>
<td>TM</td>
<td>17689</td>
<td>786</td>
<td>430</td>
<td>1.83</td>
<td>-0.23</td>
</tr>
<tr>
<td>MG</td>
<td>19991</td>
<td>746</td>
<td>486</td>
<td>1.54</td>
<td>-0.29</td>
</tr>
<tr>
<td>PI</td>
<td>28078</td>
<td>689</td>
<td>682</td>
<td>1.01</td>
<td>-0.38</td>
</tr>
<tr>
<td>NH</td>
<td>14111</td>
<td>661</td>
<td>343</td>
<td>1.93</td>
<td>-0.43</td>
</tr>
<tr>
<td>IH</td>
<td>21733</td>
<td>609</td>
<td>528</td>
<td>1.15</td>
<td>-0.51</td>
</tr>
<tr>
<td>SW</td>
<td>17376</td>
<td>478</td>
<td>422</td>
<td>1.13</td>
<td>-0.73</td>
</tr>
<tr>
<td>MD</td>
<td>18778</td>
<td>475</td>
<td>456</td>
<td>1.04</td>
<td>-0.73</td>
</tr>
<tr>
<td>CH</td>
<td>10586</td>
<td>330</td>
<td>257</td>
<td>1.28</td>
<td>-0.97</td>
</tr>
</tbody>
</table>

*Only the 21/49 YY1 core forming dipeptides where observed count in olfactory receptor genes is greater than expected are shown.

site would indicate conservation of the encoded dipeptide motif whereas a lack of conservation in motif location is consistent with selection for the presence of the motif at the DNA level. The full length YY1 binding motifs within the OLFR genes do not show any patterns in their relative positions to the protein sequences of the OLFR genes as would be detected by conserved dipeptide motifs. However, there is a bias in the orientation YY1 binding sites in OLFR genes. OLFR genes contain an average of 3.4 forward sites for every reverse site per 1 kilobase. Non-OLFR genes show only a slight bias of 1.2 reverse sites for every forward site per 1 kilobase. In contrast, a similar analysis on the members of the histone 4 gene family (Hist1h4) resulted in a different outcome. Two potential YY1 binding motifs were found within the protein-coding regions of this gene family, but the relative positions and orientations of these two motifs are identical and fixed among all the members of this gene family in mouse. The two identified YY1 binding motifs also coincide with two conserved peptide motifs (AM, and IA). Given the high
Figure 3.4. Correlation of observed and expected dipeptides and statistical significance. The mouse proteome was divided into olfactory genes (1,178) and all other genes (33,788) and the frequency of each dipeptide was tabulated. Of the possible 400 dipeptides, 49 can encode the core of the YY1 binding site. A) Observed vs. expected plot of 400 dipeptides (hollow diamonds) as compared to the subset of 49 YY1 binding site core forming dipeptides (filled diamonds). The subset does not deviate from the overall pattern. Pro-Met, PM, is indicated as having a much larger observed than expected count. B) Z-score of the observed and expected counts of the subset of YY1 binding site core forming dipeptides. All but three fall within 1 standard deviation of their expected values. Only Ala-Ile, and Pro-Leu deviate significantly from their expected values.

levels of sequence conservation detected within the members of the histone 4 family (95% sequence identity), it is uncertain whether the identified YY1 binding motifs are genuine cis-regulatory elements or simply reflecting the serendipitous sequence match between the YY1 core motifs and some combinations of codons encoding conserved peptides. This pattern is in stark contrast to the random patterns associated with the position of YY1 binding motifs in the OLFR genes.
This random pattern supports our initial idea that the clusters of the YY1 binding motifs within the OLFR genes most likely have been formed and maintained as *cis*-regulatory elements by purifying selection at the DNA sequence level. The bias favoring forward orientation may be an indication that forward sites enhance the efficiency of YY1 suppression of olfactory receptor genes. It is conceivable that multiple YY1 binding sites have been selected in OLFR genes with a preference for directionality as well as number. Alternatively, the non-random placement and orientation in histone 4 genes along with their extreme conservation prohibits a determination of whether selection is occurring at the DNA level or the protein level.

**Figure 3.5. YY1 binding sites within the ORFs of the olfactory receptor and histone 4 gene families.** Four examples of olfactory receptor genes and histone 4 genes are shown with the location and direction of their YY1 binding sites. Large gray boxes indicate open reading frames (ORFs) of these single exon genes. Boxes above each ORF show forward direction (empty box) and reverse direction (filled box) of the YY1 binding sites. The arrows in each ORF represent transcriptional directions. The bottom plot shows amino acid sequence identity for all members of each family. A total of 1,178 olfactory receptor proteins have 35% sequence identity (the four shown have 52%) while six genes of the histone 4 family share 95% identity. Conservation of position of YY1 binding sites is non-existent in olfactory genes while it is perfectly preserved in the histone 4 gene family. Directional bias in olfactory genes reflected in the figure is consistent with the complete dataset which yields 3.4 forward sites for every reverse site per 1 kb. The dipeptides encoded by the core motif of the YY1 binding site are identical for all six homologs of the histone 4 gene family (four shown). Olfactory receptor genes, by contrast, have a high number of YY1 binding sites that do not correlate to the position of these dipeptides.
Discussion

In the current study, we have shown an unusual enrichment of YY1 DNA-binding motifs in the OLFR gene family of mammals. About half of the members of the OLFR gene family have a range of 4-8 YY1 binding motifs within their protein-coding regions (Fig. 3.1 & 3.2). Statistical analyses further confirmed that this enrichment of YY1 binding motifs is consistent with functional relevance and has likely been driven by unknown selection pressure at the DNA level. Overall, the current study suggests a potential role of YY1 or YY1-related transcription factors in the regulation of the mammalian OLFR genes. Also, this study provides further evidence that the protein-coding regions of vertebrate genes can both encode codon information and contain cis-regulatory elements for transcription factor binding.

The mammalian OLFR genes are expressed primarily in olfactory neurons, and only one single copy (allele) out of the entire 1,000 family members is expressed and functional in a given neuron cell (Chess et al., 1994). This highly tissue-specific expression pattern of the OLFR genes necessitates a global repression mechanism for the majority of the OLFR genes in neural cells and in the other cell types. This mechanism acts prior to, and is separate from the negative feedback which prevents bi- and multi-allelic expression in olfactory neurons (Lewcock and Reed, 2004; Serizawa et al., 2005). Since unknown mechanisms are believed to repress a large number of the OLFR genes all the time in most of the cell types, it is likely that these repression mechanisms are mediated through epigenetic modifications. Since YY1 is a well-known epigenetic regulator in the animal genome, it is plausible to propose that the identified YY1 binding motifs may play some roles in the predicted repression mechanisms for the OLFR genes (Gordon et al., 2006). In that regard, it is important to note that YY1 is one of the Polycomb group Gene (PcG) members found in both invertebrates and vertebrates (Shi et al., 1997). Furthermore, recent studies hinted at the possibility that PcG-mediated repression mechanisms
might be involved in the regulation of the OLFR genes (Alexander et al., 2007). In mutant mouse embryonic stem cells lacking the embryonic ectoderm development (EED) protein, some OLFR genes do not replicate asynchronously as they differentiate, suggesting a loss of the typical pattern of monoallelically expressed genes (Ohno, 1970). According to the results from another recent study, some cis-regulatory elements responsible for selecting one active OLFR copy in a given neuron cell are predicted to be located within the protein-coding regions of the OLFR genes (Nguyen et al., 2007). This is intriguing and consistent with the observation of the current study in that some critical cis-elements are located within the protein-coding regions of the OLFR genes. The genetic code is optimal for containing multiple layers of encoded information within protein-coding regions (Itzkovitz and Alon, 2007). In sum, although further investigation is warranted, it is likely that the identified YY1 binding motifs are functionally important cis-regulatory elements for the regulation of the OLFR genes.

The YY1 binding motifs identified within the OLFR genes are unusual since they are localized within the protein-coding regions of these genes and are present at high density. This is quite different from the typical pattern in which transcription factor binding sites (cis-regulatory elements) are located in the genomic regions surrounding the protein-coding regions of genes. OLFR genes give a fitness advantage by duplication and divergence while remaining active, yet must also retain regulatory information. According to our statistical analyses (Fig. 3.3 & 3.4), the identified YY1 binding motifs within the OLFR genes likely represent evolutionarily selected cis-regulatory elements. Previous studies have shown high levels of YY1 binding affinity to the type of YY1 binding motifs found within OLFRs (Kim et al., 2007). Though the functionality of the identified YY1 binding motifs remains to be demonstrated in vivo, it is plausible that some functional constraints serve to maintain the YY1 binding motifs within the protein-coding regions of the OLFR genes. In one scenario these motifs might be linked to the sudden expansion
of this gene family in mammalian genomes. The copy number of the OLFR genes has increased dramatically in recent evolutionary times in mammalian genomes, providing a large number of receptor proteins for airborne scents (Niimura and Nei, 2006). The clustering of the OLFR genes in chromosomal regions also suggests this gene family may have been duplicated through in situ tandem duplications (Niimura and Nei, 2006). Tandem duplication is known to be the most frequent mechanism in increasing copy numbers for gene families (Ohno, 1970). However, it is not well understood how this mechanism carries over the proper information for the transcriptional regulation of duplicated gene copies. In the case of the mammalian OLFR genes, their protein-coding regions may have both information for codon and cis-regulatory elements so that the duplication of these genes would most likely guarantee their coding potential as well as associated transcriptional control. This might have been one functional constraint for the co-evolution of the YY1 binding motifs within the protein-coding potential of the OLFR genes.

**Conclusion**

The current study reports that an unusual enrichment of YY1 binding sites, 4-8 binding sites per gene, are located in the coding regions of olfactory receptor genes in mammals. According to statistical analyses, these YY1 binding sites most likely have been selected as cis-regulatory elements. Also, similar patterns are found in other mammals, but not in fish, suggesting a mammalian-specific phenomenon. This study further suggests YY1 or YY1-related transcription factors as regulators of mammalian OLFR genes.

**Materials and Methods**

**Visualization of YY1-binding Sites in Coding Regions**

A custom Perl script, matrix-bidirectional.pl, was run against mouse chromosome files available from Ensemble (version NCBIM37.49) ftp://ftp.ensembl.org/pub/current_fasta/mus_musculus/dna/ (see Additional file 1 at
http://jookimlab.lsu.edu (Flicek et al., 2008). The program calculates score by matching a 10 bp window to a matrix of the likelihood for each position (Additional files 2 & 3 at http://jookimlab.lsu.edu) along the 10 bp consensus sequence (the Position Weight Matrix, PWM). Each base pair is given a value equivalent to the decimal percentage of its match to the known YY1 binding motifs. The four base pair core, CCAT, is scored at 100% for each base plus 2. Flanking bases vary in score from 0 to 1 based upon their frequency in known YY1 binding motifs. The total score for each 10 bp window is calculated and compared to our cutoff score of 8.0 which indicates a good match. Our previous work revealed that scores above 8.0 correlate with good YY1 binding in vitro (Kim et al., 2007; Kim, 2008). Output was generated in the WIG format for each chromosome with each YY1 binding motif score represented by start and end position and bar height corresponding to the PWM score match. Position weight matrix scores and location information were loaded into the University of California, Santa Cruz (UCSC) Genome Browser for visualization of YY1 location (Fig. 3.1) (Kent et al., 2002).

**Motif Finding and Scoring**

Our Perl script was run against the mouse, human and zebrafish mRNA available from NCBI (ftp://ftp.ncbi.nih.gov/genomes/). Each gene was scored for number and quality of YY1 motifs. Results were sorted by a YY1 density score, the combined score of YY1 motifs divided by the length of the gene. Predicted and hypothetical genes were removed from the mouse yielding 40,009 total genes (19,818 removed, 20,191 remaining), with 19,083 non-olfactory and 1,108 olfactory genes. Hypothetical genes were removed from the human transcriptome, yielding 24,886 total genes with 24,506 non-olfactory and 380 olfactory genes. Hypothetical genes were removed from the zebrafish transcriptome, yielding 9,092 genes with 9,067 non-odorant genes and 25 odorant receptors. We removed the hypothetical and predicted genes because they may not contain complete ORFs. The upstream 1kb regions of 20419 Refseq genes in the mouse were
obtained from UCSC (http://hgdownload.cse.ucsc.edu/downloads.html).

Histograms were made in Microsoft Excel 2007 by sorting the genes by YY1 density from high to low, then assigning a count to order the genes. Olfactory receptor genes were separated from non-olfactory receptor genes and the count numbers were used as position information to make a histogram which shows the distribution of OLFRs along the range of YY1 containing genes (Fig. 3.2).

**Protein Motif Correlation Testing**

The mouse proteome was downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/M_musculus/protein/protein.fa.gz) which contains 34,966 peptide sequences and nomenclature. Olfactory receptor proteins (1,178) were separated from non-olfactory proteins (33,788). A Perl script, dipep-singlefile.pl was used to generate a count of each of the 400 possible dipeptides in each of these groups (see Additional file 4 at http://jookimlab.lsu.edu).

A Z-test was performed according to the formula $Z = (\text{observed} - \mu)/\delta$ where observed is the count of each possible dipeptide found in olfactory receptors, $\mu$ is the mean of the counts from the whole population of non-olfactory proteins, and $\delta$ is the standard deviation of the population count. Z-score units are given in standard deviations from the mean. Expected count was calculated by multiplying the frequency of each dipeptide from all non-olfactory receptor proteins by the total number of dipeptides seen in OLFR proteins. Figure 3.4 shows the plot generated in Microsoft Excel 2007 comparing the observed to expected ratio for all 400 possible dipeptides and the subset of 49 YY1 core-forming dipeptides which exhibits no difference in the distribution of the subset.

We found 21 of the 49 dipeptides which can make up a YY1 binding site had greater than expected values, but only 2 were over 3 standard deviations away from the mean (Table 3.1,
Figure 3.4). Table 3.1 shows only the dipeptides in which the observed count in olfactory genes was higher than the expected count in mouse.

Global alignment was done using ClustalW with 1,178 OLFR and 6 hist1h4 amino acid sequences from mouse (Fig. 3.5) (Larkin et al., 2007). All Perl scripts are available for download on our website at http://JooKimLab.lsu.edu.

References


CHAPTER FOUR:
POSITION-BASED CLUSTERING OF MICROARRAY EXPRESSION DATA*

*Reprinted by permission of Cold Spring Harbor Protocols

59
**Introduction**

Microarray expression analysis has traditionally focused only on genes with the highest expression level changes, ignoring the majority of genes with lower fold changes. To address this problem, we provide a simple method that can derive additional useful biological information from the rest of the genes included on an array. Chromosomal position information combined with expression data, either raw or in expression-level changes, gives a regional overview of expression from an array. This protocol is tailored to Affymetrix data, but can be used for other types of microarray results. The procedure is illustrated by the reanalysis of a data set, GSE5230, previously deposited in the Gene Expression Omnibus (GEO). Using this procedure, we identified several classes of chromosomal regions where expression levels were affected in concert. Linking expression data to chromosomal position also allowed us to identify several genomic regions displaying low but steady fold change, which were missed by traditional approaches. Overall, this method is useful in detecting regional regulatory changes. It should allow for greater use of the large quantity of previously overlooked microarray data.

**Method**

**Find and Download Microarray Data**

1. Using a Web browser, download microarray data from the National Center for Biotechnology Information (NCBI) GEO. Either query for a specific gene or browse the database by platform, data set, or series.

**Preprocess Microarray Data**

2. Create a spreadsheet in Excel with one row per gene, and collect the raw data from each individual array into columns. Expression data must be combined from biological and technical replicates in accordance with the original experimental design parameters. The minimum information for each gene is: gene ID, position, and score per array.
Figure 4.1. Schema for generating position-based data vs. traditional method. (A) Flowchart for microarray analysis. (B) Traditional methods yield high fold-change single gene hits without chromosomal context. (C) After sorting on chromosomal position and applying heat map colorization, regions of interest can be visually identified. The approach uses chromosome positional information to highlight possible regional effects.

3. Remove any genes from the analysis that contain unreliable values, and combine the scores from technical and biological replicates. (See Troubleshooting.)

4. Calculate the fold change by taking the log2 of the average value of each gene from each condition divided by the control average (Fig. 4.2). In the example data set, GSE5230, these are designated as aza-cytidine (AZA)/Control, trichostatin A (TSA)/Control, and Combination/Control.

(Log2 absolute values are equivalent for up-regulation and down-regulation and the range is continuous, thus simplifying later analysis.)
**Note on fold change calculation**

Fold change can be calculated as reciprocal fold change with the formula \( f_g = \frac{u_{g,\text{experiment}}}{u_{g,\text{control}}} \) if \( u_{g,\text{experiment}} \geq u_{g,\text{control}} \) or \( f_g = -1 \times \frac{u_{g,\text{control}}}{u_{g,\text{experiment}}} \) if \( u_{g,\text{experiment}} < u_{g,\text{control}} \) (Breese et al. 2003). However this gives values where the sign indicates either up-regulation (+ values) or down-regulation (- values); the set \( \{ x : 1 < x < -1 \} \). In this inverse transformation scheme, -1 fold = +1 fold, whereas a heat map is most easily displayed with a continuous range. Thus we convert the fold changes to a logarithmic transformation (base 2), where \( x < 0 \) = down-regulated, \( x > 1 \) = up-regulated via the formula \( \log_2 \left[ \frac{u_{g,\text{experiment}}}{u_{g,\text{control}}} \right] \).

**Figure 4.2. Note on fold change calculation.** Fold change can be calculated as reciprocal fold change using the formulas shown (Breese, Stephens et al. 2003). To generate values for a heat map, use a log base 2 transformation to convert the fold change ratios.

5. Use Excel’s text-to-columns tool to extract chromosome and position information from the annotation table, using the first entry in the annotation column as definitive. Unix regex commands can also be used: %s/::\.*//g for chromosome and %s/chr\\d\+:.*\/(\d\+\-\-\d\+)\//g for positions.

(These extracted texts can be further converted into numeric values with the Value command in the Excel file. These numeric values are used for sorting the entire set of microarray data, see Step 7.)

6. Generate heat maps using the following conditional formatting: Format Style 3-color style, Minimum Type (Number, -1.4), Maximum Type (Number, 1.4).

**Identification of Regions of Interest**

(The regions of interest are both small regions of contiguous genes with higher expression and regions of slightly lower contiguous expression.)

7. Sort the values in the spreadsheet based on position by using Excel’s custom sort procedure and choose the column containing position data.
8. Apply Excel’s conditional formatting function and set maximum and minimum values appropriate to your data set. Manually filter against the individual transcript fold changes to look for groupings of expression changes. Extract these groups and examine them further for biological significance.

**Troubleshooting**

**Problem:** There is ambiguous position information.

*[Step 3]*

**Solution:** Although position information for the majority of transcripts can be retrieved immediately as part of an annotation library file (in analyzing the data sets from GEO GSE5230, this was 92%), transcripts having ambiguous position information should be removed. (See Discussion for additional explanation.)

**Discussion**

Microarrays have revolutionized the large-scale capture of gene expression data, allowing tens of thousands of transcript expression levels to be quantified quickly and cheaply. With the expansion of microarrays to the study of single nucleotide polymorphisms (SNPs), gene copy number variation, and methylation, analysis and understanding of the results is even more critical (Trevino et al. 2007). The most direct way to analyze this avalanche of data is to look for the genes with the highest change between experimental conditions. This primary fold change is often the first data a researcher sees and is used to build “top gene lists” of the most affected genes. A secondary analysis is usually performed by grouping genes by Gene Ontology (GO) classification, by biological pathway, by cluster analysis, or by other methods (Beissbarth 2006; Quackenbush 2006). The purpose of the secondary analysis is to extract more biologically meaningful information from the vast tables produced.
In higher eukaryotes, genomic and chromatin context are often a contributing factor to the expression levels and patterns of individual genes. For example, a group of genes with similar functions are colocalized and form a domain structure that coordinates spatial and temporal expression of individual genes. Also, epigenetic modifications and chromatin structures are usually modulated at relatively large scales involving multiple genes rather than single genes. Thus, the aim of this work is to interpret expression of individual genes as part of a larger domain, which necessitates clustering expression data by chromosomal positions (Razin 1998). Although Affymetrix provides position annotation, it often goes unused due to the difficulty of integrating it with expression data.

Position-based clustering of microarray expression data is a rapid method that looks for regional or multigene effects in microarray results. It provides a straightforward preliminary analysis used to find targets suitable for further benchwork testing. Position-based clustering complements existing “top hit” traditional approaches as well as the biological process method, found at the Gene Ontology project (http://www.geneontology.org), and other clustering methods. Using this protocol, gene clusters that are slightly up- or down-regulated stand out, revealing region-specific mechanisms of gene control. While the fold-change approach is statistically insufficient to find biological inference by itself, the heat map showing the relative position of genes that are near each other and show similar expression change is a good method for identifying regions of interest (Allison et al. 2006). In our example analysis, we used Affymetrix oligonucleotide arrays, but other types of microarrays can be analyzed using this protocol. For example, tiling arrays used for regulatory region detection and chromosomal function can be scanned for regions of consistent changes. Even ChIP-chip data can be reorganized according to chromosomal position to highlight the density of particular proteins binding in certain regions or their binding target spacing. With the explosion of publicly
available data, such as NCBI’s GEO database, position-based reanalysis of existing arrays can be a quick first step to find new targets of study.

**Ambiguous Position Information**

One of the main causes of ambiguous position information stemmed from either multiple positions for one transcript probe or one gene transcript with multiple probes. In most cases, a single probe was assigned to a single chromosomal location. However, some probes were assigned to multiple locations due to their high levels of sequence similarity. Also, the majority of the transcripts were usually represented by more than one probe, but this was not applied consistently to all of the genes. Some transcripts were represented by more than four or five probes, whereas others were each represented by a single probe. These were sometimes true replicates, but often they represented isoforms—transcripts from a different region of a gene. This inconsistent representation of the transcripts in the microarray resulted in the faulty overrepresentation in our heat map.

**Comparison to Other Methods and Previous Applications**

Previous studies have used the position information of a gene for analyzing expression data, but the methods are more complex than ours (Sabatti et al. 2002; Su et al. 2004). For example, Persson et al. used a sliding window to identify regions of coordinated expression to correlate tumor grades (Persson et al. 2007). Their algorithm, although useful, requires more complex statistics and software than is available to the typical bench scientist. The CHROMOWAVE model, implemented in Matlab, uses positional information and the wavelet transform to isolate regions of coexpressed genes comparable to the targets provided by our method (Turkheimer et al. 2006). The approach that is most similar to the one described in this protocol was developed for the GenomeCrawler algorithm (http://www.rockefeller.edu/vaf/strearray.php), which calculates statistical significance for
neighboring clusters followed by visual inspection (Ryan et al. 2007). Although this method is similar in principle to ours, the GenomeCrawler software must be run in the R software environment, which may be an impediment for some researchers. Regions of coordinated expression are provided by these and other statistically rigorous methods, whereas the method described in this protocol uses commonly available software (i.e., Excel) to derive initial targets for further research.

**Reanalysis of HepG2 Microarray Data: An Example**

For demonstrating the current approach, we used a set of microarray data derived from the human hepatoma cell line HepG2 that was deposited by Dannenberg and Edenberg (GEO accession no. GSE5230) (Dannenberg and Edenberg 2006). The information for our analysis was downloaded from the following websites: microarray data from the GEO (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5230), and the library reference file from Affymetrix (http://www.affymetrix.com/products/arrays/specific/hgu133plus.affx). The reference file was used because the HepG2 arrays do not include position information in the output file, so position information was collated from the library reference file. Supplementary data from our laboratory can be found at http://jookimlab1.lsu.edu/?q=node/39. The data downloaded from GEO contained four biological replicates—identical microarrays hybridized with RNA from the same experimental condition—and four different conditions, for a total of 16 slides (arrayed on the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array). The GSE5230 data set was reanalyzed as follows. (Step numbers that follow refer to the protocol steps above.)

**Step 1.** The 16 raw data sets were downloaded as individual Excel files.
Step 2. Replicates were grouped together based on individual treatments or technical replicates. All raw values labeled “Absent” by the Affymetrix Microarray Suite, due to lack of statistical significance, were removed.

Step 3. An average value was calculated for each gene transcript that contained three or more expression values out of the four replicates of each experimental condition and the control condition as given in the Affymetrix output file. The Excel formula is IF(COUNT(D2,G2,J2,M2)>=3,AVERAGE(D2,G2,J2,M2),“”), where each column represents raw values of expression for each of the microarrays. Each of the three experimental condition averages was compared to the control set average; however, all genes were removed if there was no average value for the control (therefore less than three out of four “Present” values from the four control arrays).

Step 4. The fold changes were calculated with the log2 of the average value from one condition divided by that of the negative control (no treatment). These calculations provided each gene transcript with three fold changes in a simple log transformed scale (experimental/control): AZA/Control, TSA/Control, and Combination/Control as shown in figure 4.3. This process resulted in a final set of 20,118 transcripts from the initial set of 54,568 gene transcripts.

Step 5. These transcripts were arrayed again based on their chromosomal positions. For this task, position information was extracted for each gene transcript from a library reference file containing gene annotations, which was incorporated into the Excel file with the three fold changes. The resulting Excel file was reorganized based on chromosomes and positions.

Step 6. To visualize the fold changes of a group of adjacent gene transcripts more effectively, heat maps were generated using Excel’s conditional formatting. To generate the heat map in figure 4.3, a given cell within the spreadsheet of the Excel file was color-coded based on its log2 fold change, ranging from -1.4 (blue, representing down-regulation) to 1.4 (red, representing up-
regulation). Any values above or below these thresholds appear either blue if below -1.4 or red if above 1.4. The fold change of 0 indicates no expression change and was represented in yellow. The gradual increase of the fold changes correlated with the transition from the blue to yellow to red color with different levels of intensity as shown in figure 4.3.

Figure 4.3. Heat map of position-based expression. Ideogram global view with the expression ratios shown as a heat map in relative gene order. Next to each chromosome are the three columns of the heat map. AZA treatment is shown in the left column, TSA in the middle, and combination treatment in the right column. Red represents upregulated genes, blue represents down-regulated genes, and yellow shows gradations of expression in between.

Steps 7 and 8. Finally, the entire length of each chromosome was scanned to identify regions with steady high values or low values, which manifest as a group of vertically adjacent spreadsheet cells in red or blue. In figure 4.4, four regions are magnified with a chart of the column values for visual comparison. In each selected region, at least one experimental condition is consistently up- or down-regulated and is easily distinguished by sight. The objective while
scanning was to identify regions that contained several transcripts falling above or below the cutoff value that were also within a range of 2 Mb. We magnified regions of unusual response and illustrated the data points in a graph to the right (Fig. 4.4). Although our manual scans through the heat map located 24 genomic regions with interesting expression patterns, this method could be improved by incorporating some automation.

**Figure 4.4. Four clusters of unusual response.** Four selected regions are magnified here and the columns are represented in chart format to the right of each region. In these four regions, at least one experimental condition is consistently up- or down-regulated. (A) Metallothionein family (120 kb) shows increased expression across all three conditions. (B) Aldo-keto reductase family (131 kb) shows increased expression across all three conditions, with strongest response in the combination treatment. (C) Histone cluster (102 kb) shows up-regulation in response to demethylation by AZA, and down-regulation in response to acetylation by TSA. Methylation response appears dominant. (D) Down regulated region (1 Mb) had slight but consistent down-regulated response to acetylation by TSA and a synergistic effect downward when combined with demethylation by AZA.

**References**


CHAPTER FIVE:

SUMMARY
Given the great diversity of transcription factors encoded in the human genome, their role in gene regulation is crucial to our understanding of organismal development (Wray, 2003). Nearly 10% of human genes encode DNA-binding transcription factors, and their importance in understanding evolution both within and between species is widely accepted (Babu et al., 2004). Small changes in a single transcription factor can affect many downstream target genes, either by changing their DNA target motif or by recruiting different proteins to a genomic location. Thus mutations in transcription factors induce pleiotropic effects. Changes in transcription factors can impact human health, as is seen in several diseases such as cancer and developmental disorders (Semenza, 1998; Darnell, 2002). Transcription factors also tend to reinforce speciation as their genes duplicate and diverge rapidly between lineages (Edger and Pires, 2009). Likewise, the hundreds of olfactory receptor genes (OLFRs) found in vertebrates appear to be in a state of flux, with dramatic differences in copy number between lineages (Kambere and Lane, 2007). The functional consequences of a change in an olfactory receptor gene are likely to be less dramatic than those of duplicated or mutated transcription factors, which can have pleiotropic effects. Both families have been studied as models of gene duplication and divergence. Here we present the evolutionary history of transcription factor YY1 and the first systematic analysis of its binding sites within the coding regions of olfactory receptor genes.

In this dissertation, I examined the transcription factor YY1 and its mammal-specific homologs, YY2 and REX1. The data reveal an overabundance of YY1-binding sites in olfactory receptors. We can ask, then, how do genes duplicate and adopt lineage-specific functions? Our findings suggest that changes in the DNA-binding domain of YY1 duplicates correspond to changes in binding site specificity. We also observed that these binding sites can be found within coding regions and selectively maintained through evolution.
In chapter two, we confirmed via phylogenetic analysis that YY1 has been essentially unchanged over 600 million years and, at the same time, has remained as a single-copy gene in most vertebrates. Over this time, YY1 and the version found in insects, called PHO, retained the same target DNA-binding motif. However, during the radiation of placental mammals 65 million years ago, two duplicate copies were acquired by mammals and recruited for different functions. The DNA-binding motif was now free to change due to relaxation of purifying selection on the two new copies. The binding affinity was reduced in both YY2 and REX1, and the target motif has coevolved away from the YY1 consensus in the case of REX1.

Gene duplication via retrotransposition rarely leads to a functional copy (Hurles, 2004). One plausible pathway to generate a new functional copy is for the duplicate copy to adopt a different expression profile. Indeed, whereas YY1 is ubiquitously expressed, YY2 is expressed tissue-specifically (in germ cells), and REX1 is found primarily in stem cells; their limited expression likely allows them to avoid deleterious interaction with YY1. In addition, REX1 is emerging as a strong candidate for mediating imprinting control and placenta formation, both mammalian innovations. Furthermore, our phylogenetic analysis hints at important differences in the evolution of these two homologs (Fig 2.2). YY2 appears to have duplicated and subsequently diverged along with each lineage, whereas REX1 underwent a period of rapid change prior to the radiation of mammals, before lineage-specific divergence took hold. In the greater context of biology, these new duplicate genes provide a quick mechanism to increase genomic complexity while providing flexibility to adapt new structures. Mammals have a number of unique characteristics such as hair, milk, and placentas, which have allowed them to become the most successful group of large vertebrates on land. These adaptations required new genes and expression control proteins. Based on the evidence I have shown, I posit that YY2 and REX1 are likely integral parts of the success of the mammalian lineage. With the coming
availability of genome sequences from more marsupials and basal eutherian mammals, future studies should be able to distinguish the order of the insertion events for YY2 and REX1 in evolutionary history. These genomes may additionally aid in the detection of new lineage-specific duplicates, giving further insight into the properties necessary for successful insertion of a YY1 paralog.

Having explored the YY1 family of transcription factors, we progressed to the study of their target binding motifs in chapter three. We scanned the transcriptome of the human, mouse, and cow and found an unusually high density of putative YY1-binding sites within the coding regions of olfactory receptor genes. A scan of the zebrafish, however, did not reveal the same pattern, suggesting that selection for a high YY1-binding site occurred after the development of land vertebrates. During the radiation of land vertebrates, a subset of olfactory receptor proteins underwent a dramatic expansion, leading to over 800 and 1300 copies in the human and mouse, respectively. The adaptation of OLFRs to detect airborne scents coincides with the expansion of OLFR genes in terrestrial animals and with the loss of genes which detect water soluble odorants (Niimura, 2009).

Olfactory receptors in land vertebrates are under an unusual selection pressure. First, unlike YY1 and other dosage-critical genes, new duplicate OLFRs are more likely to be retained in the genome since each new functional copy has the potential to bind a novel ligand and expand the repertoire of scents that an animal can detect. Second, in giving animals a wider array of scents to detect, rapid divergence of a copy’s coding sequence is favorable under natural selection; however, the sequence must retain its transmembrane-spanning properties and any control regions which allow it to remain functional. Correspondingly, we found the overall sequence identity to be only 35% across all the murine OLFRs (Fig. 3.5). Third, any non-coding sequence containing control elements would be lost during some forms of gene duplication, such
as retroposition. Previous studies suggest a *cis*-regulatory element is located within the coding region of the OLFRs (Merriam and Chess, 2007). Since the genetic code is nearly optimal for the ability to host multiple layers of information, we predicted the presence of control elements within the coding region for regulation of expression in addition to merely encoding an amino acid (Itzkovitz and Alon, 2007).

We performed statistical analyses to rule out the possibility that a conserved two-amino-acid motif (dipeptide) explains the high density of YY1-binding sites within OLFRs. We also controlled for the prevalence of each dipeptide combination found in OLFRs when compared to all other genes. The YY1 core motif does not account for an enrichment of any particular dipeptide in the coding region, nor was there any conserved dipeptide motif that could explain the overrepresentation of YY1-binding sites in OLFRs. Thus, we conclude that regulatory information is indeed present within the coding region of this family of genes. We believe the unusual selection pressure driving the rapid evolution of OLFRs has resulted in exactly this outcome. Since few other instances of regulatory information found in coding sequence are known, we expect these findings will impact the field by expanding the search for regulatory elements to encompass the coding regions of genes, as well as flanking regions (Lin and Tam, 2001). Future *in vivo* tests using chromatin immunoprecipitation may experimentally confirm the binding of YY1 to OLFR coding regions. Since the YY1 binding site can be bound by YY2 or REX1, the possibility exists, however, that one of these mammal-specific homologs is responsible for controlling repression. Finally, discovering an animal with only a single paralog or none and examining the pattern of YY1-binding sites within its OLFR repertoire would satisfactorily answer this question.

YY1 is associated with imprinting control regions which are mammal-specific (Kim et al., 2007). Evidence suggests that this novel function is dependent upon longer than usual YY1
binding sites which have increased affinity to the protein (Kim, 2009). Imprinted genes exhibit single-allele expression in a manner similar to olfactory receptor genes. So, transcription factors and olfactory receptors function as more than mere models for gene and species evolution. Our findings serve to tie together the functions of transcription factors and olfactory receptors.

In chapter four, we describe an auxiliary method of microarray analysis. Of the many methods of microarray analysis, most focus on the change in transcription of individual genes, typically generating a list of genes with the highest absolute fold change. Our method incorporates chromosome position, thereby providing additional information from which to extract genes of interest. While this method is supplementary to existing analyses, it is platform independent. Potentially other large expression datasets, including tiling arrays, ChIP-Chip results, and even quantitative next-generation sequencing data, can be analyzed in a position-based manner. DNA is not merely a string of letters but an epigenetically dynamic three-dimensional structure. As such, our position-based methodology can increase the biologically relevant return on experiments by correlating the expression of neighboring genes to better determine regions of co-regulated change. Further, with the plummeting cost of sequencing and the anticipated explosion of individual labs capable of sequencing whole transcriptomes, it is important to provide a means for non-specialists to easily extract biologically relevant information.

This dissertation chronicles the evolutionary history of the YY1 family of proteins. We describe the concurrence of YY1-binding motifs in the coding regions of olfactory receptor genes and the evolutionary consequences of this coupling. Finally, we provide a new method of microarray analysis. With these firm foundations, the study of YY1 and its homologs can move forward and advance the fields of chromatin biology, genomic imprinting, and gene evolution.
References


Dear Chris Faulk,

Thank you for your recent email requesting permission to reuse all or part of your article in a new in a publication, a thesis or as part of your teaching.

As part of your copyright agreement with Oxford University Press you have retained the right, after publication, to use all or part of the article and abstract, in the preparation of derivative works, extension of the article into book-length or in other works, provided that a full acknowledgement is made to the original publication in the journal. As a result, you should not require direct permission from Oxford University Press to reuse your article.

However, if you are required by your new publisher or employer to obtain full written permission prior to reuse, please let us know and we will draw up a letter as soon as possible.

For full details of our publication and rights policy please see the attached link to our website: http://www.oxfordjournals.org/access_purchase/publication_rights.html If you have any other questions or queries, please feel free to contact us.

Yours sincerely

Katherine Randell
Permissions Assistant Academic
Rights & Journals
Oxford University Press
Great Clarendon Street
Oxford OX2 6DP
Direct tel. +44 (0)1865 353672
Direct fax +44 (0)1865 353429
e mail: Katherine.Randell@oup.com
Dear Christopher

Thank you for contacting BioMed Central.

BMC Genomics is an open access journal therefore you are free to use the article as requested. Please ensure you attribute BioMed Central as the original publisher.

If you have any questions please don’t hesitate to contact me.

Best wishes

David Roman

david.roman@biomedcentral.com

www.biomedcentral.com
February 4, 2010

Christopher Faulk
Kim Lab
Systematics, Ecology and Evolution
Louisiana State University
Baton Rouge LA 70803

Dear Mr. Faulk,

In response to your request regarding permission to reproduce the following material:

Selection: Article reprint
Title Reference: http://cshprotocols.cshlp.org/cgi/content/full/2009/9/pdb.prot5280
Chapter Title: Position-Based Clustering of Microarray Expression Data
Chapter Authors: Faulk and Kim
Copyright date: 2009

To appear in:
New work: PhD dissertation
Publisher: Louisiana State University
Author:Christopher Faulk
Date of Publication: 2010

I confirm that we are willing to grant non-exclusive rights in the English language, subject to the following conditions:

1. The material to be used has appeared in our publication without credit or acknowledgment to another source.
2. Full acknowledgment shall be made to the source and correct reference given. Copyright holder is Cold Spring Harbor Laboratory Press.
3. Permission is limited to the proposed use, on a non-exclusive, one-time-only basis, with distribution rights in the English language. This permission relates to use in print and electronic formats only as described in request. Rights do not apply to any other use.

By Carol C. Brown
Date: February 4, 2010
For Cold Spring Harbor Laboratory Press, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA
VITA

Christopher Don Faulk is the son of Donald and Aleta Faulk. He was born in 1978 and raised in Houma, Louisiana. Christopher graduated with a Bachelor of Arts degree in English from Louisiana State University in 2001. He worked as a research associate at Pennington Biomedical Research Center in Baton Rouge, Louisiana, under the direction of Dr. Leslie Kozak. Subsequently he began his doctoral research in the spring of 2005 in the Department of Biological Sciences at Louisiana State University and was mentored by Dr. Joomyeong Kim. Mr. Faulk will graduate with the degree of Doctor of Philosophy in May 2010.