University of Puerto Rico
Rio Piedras Campus
Department of Biology

EVOLUTION AND DNA-BINDING SPECIFICITY OF THE SIX CLASS OF TRANSCRIPTION FACTORS

By:
Anthony R. Rivera Barreto

In fulfillment of the
requirements for the degree of
Doctor in Philosophy

December 09th 2022

San Juan, Puerto Rico

Table of Contents

List of Figures

Lift of Tables

List of Abbreviations

TF	Transcription Factor
GRN	Gene Regulatory Networks
CRE	Cis-Regulatory Elements
TFBS	Transcription Factor Binding Site
PWM	Position Weight Matrix
DBD	DNA Binding Domain
HD	Homeodomain
SIX	Sine Homeobox Family
SD	SIX Domain
EMSA	Electrophoretic Mobility Shift Assay
SELEX	Systematic Evolution of Ligands by Exponential Enrichment
ChIP-seq	Chromatin Immunoprecipitation next-generation sequencing

ACCEPTED BY THE FACULTY OF NATURAL SCIENCES
DEPARTMENT OF BIOLOGY
UNIVERSITY OF PUERTO RICO RIO PIEDRAS
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF BIOLOGY

____

José A. Rodríguez-Martínez, Ph.D.
THESIS DIRECTOR

____

Abel Baerga Ortíz, Ph.D.
Riccardo Papa, Ph.D.
DISSERTATION COMMITTEE MEMBER

____

Alfredo Ghezzi Grau, Ph.D.
DISSERTATION COMMITTEE MEMBER

____

Humberto Ortíz-Zuazaga, Ph.D.
DISSERTATION COMMITTEE MEMBER

____

Riccardo Papa, Ph.D.
DISSERTATION COMMITTEE MEMBER

Abstract

Transcription factors (TF) are critical for development and cellular processes and are found in all organisms. How their DNA-binding specificity changes through time has yet to be fully understood. TF DNA-binding specificity is determined by how their DNA-binding domain (DBD) interacts with DNA. TFs are identified by the sequence homology shared with described DBDs, which allows them to be classified into families. It is accepted that similar DBDs have the same DNA-binding specificity and bind to the same sequences. However, changes in a TF can lead to changes in its DNA recognition. TFs members of **the sine oculis homeobox** (SIX) homeodomain family are found from sponges to humans and are considered atypical members of the homeodomain (HD) family. They regulate numerous processes and phenotypic features, from eye development in flies and humans to red color patterning in *Heliconius* butterflies wings to human brain development. How evolutionary related TFs diversify has yet to be fully understood, especially diversification of their DNA-binding specificity. To understand the evolutionary history of this family, we performed phylogenetic inference that placed the first SIX within *Porifera* and the presence of the three canonical SIX (sine oculis, optix, and six4) in *Cnidaria*. In addition, we observe the presence of two major groups that show that optix and six4 are more evolutionary related. To determine changes in DNA-binding specificity, we performed *in vitro* Systematic Evolution of Ligands by Exponential Enrichment (SELEX-seq) using full length SIX TF proteins from *Drosophila melanogaster, Heliconius erato*, and *Homo sapiens*. Our data shows the majority of SIX TFs bind to the canonical binding motif (`5′-TGATAC-3′`), except for six4 members, which seem to prefer (`5′-TGACAC-3′`).

Interestingly, the way they bind to these motifs differs. Both *sine oculis* and *six4* homologs require a `5′-GA` dinucleotide flanking the core motif on the 5’-end. In comparison, optix related

members prefer a shorter flaking region and less dependence on 5′-GA. This is interesting since *optix* is more evolutionarily related to *six4* than to s*ine oculis*. We also found that *Heliconius erato* optix can bind DNA both as a monomer and as a homodimer with a preferred spacing of 2-bp between binding sites. Using the determined DNA-binding specificity of optix, we were able to predict optix binding to cis-regulatory elements (CRE) active during wing development. optix was capable to bind to all the predicted sites, including to its own promoter. Validation of optix binding to these CREs allows to expand the search of optix gene targets and contribute to our understanding of the mechanism of wing development and red color patterns in *Heliconius* butterflies.

Dedication

This work is dedicated to my hard-working and supportive family. I dedicate my thesis to my mom Marta E Barreto Velazquez, for her undying support and to my dad Rafael Rivera Rodriguez. He worked two jobs to ensure that my sibling and I could have a future while guiding me to excel at all I do. To my brother and sister, Christian R Rivera Barreto and Valeria C Rivera Barreto, for allowing me to guide them while learning from them daily. To my grandfather Norberto Barreto Hernández for religiously picking me up from school since elementary school. To my grandmother Hilda Velazquez Mercado for her knowledge and support throughout my life. To my uncles, Norberto Barreto and Edwin Barreto, for being my role models and showing me the importance of learning and cultivating the pleasures of reading.

I also dedicate this work to my friend, my support, my confidant, and my fiancée Elis M De Jesus Figueroa, for her unwavering love, patience, and support. Thanks for motivating me during my lows and hearing my crazy ideas; thanks for being you. Lastly, I would like also to dedicate this work to my great-grandmother Cecilia Mercado, whose unimaginable wisdom guided every step of my life. Even in her absence, her words resonate in every decision. I am lucky to have had this supporting family to end this chapter in my life, and I am grateful for having them for what’s to come.

Acknowledgments

First, I would like to acknowledge Dr. José A Rodríguez-Martínez for giving me the opportunity to join his laboratory. He gave me chance when nobody else was considering it, and his support and knowledge have made me the scientist I am today. If it wasn't for him, I would have not been able to complete my Ph.D., and for that, I am eternally grateful. I would also like to acknowledge my graduate committee, Dr. Abel Baerga Ortíz, Dr. Alfredo Ghezzi, Dr. Humberto Ortíz Zuazuaga, and Dr. Riccardo Papa. I would like to specifically acknowledge Dr. Riccardo Papa for allowing me to work within his EPSCoR grant, which allow me to learn so much both as an academic and personally. If it wasn't for Dr. Papa, I wouldn't have been able to do most of the work done in this thesis. I would also like to give a shoutout to the CSMER, to both Dr. Michelle Borrero and Brenda Santiago, both allowed me to contribute to multiple projects while also supporting me every step of my work.

The JARMlab, past and present members played an enormous role in my work, and for that, I am eternally grateful. I would like to give thanks, especially to Jessica M Rodríguez Rios and Rosalba Velazquez Roig, we make a great team, and the days in the JARMlab wouldn't have been the same without you both. Thanks to Jessica, going from my undergraduate to my colleague to a great friend, you are also part of this work. To Rosalba, thank you for your unwavering support and friendship, and thanks for the unlimited supplies of bananas, sorry for all the sweets. In addition, I would also like to thank my two mentoree Valeria Santa and Luis Daniel Garcia Sanchez, their work and input is deeply appreciated, this work was a team effort and without them it wouldn't have been the same.

I would also like to thank the RISE program for funding my Ph.D., especially Grisselle I. Hernández and Julissa Morales. The RISE program wouldn't be the same without the commitment of both to the program, and both are one of the highlights of being part of the program. I would also like to express my gratitude to BioXFEL and William Bauer for supporting my research. Special gratitude to Dr. Federico Hoffmann, who allowed me to visit his laboratory to finish my thesis work while also providing critical mentorship not only academically but also personally. Thanks to Hoffmann lab for letting me work with everyone while also learning from each member, especially Amanda Black, for being so welcoming and making me feel at home.

I would also like to give thanks to Miguel Urdaneta, whose guidance made me appreciate what meant to be a scientist and, in the one, was a cornerstone to my decision to proceed to graduate school. I will forever cherish my time in the microbiology lab. I would like to give thanks to anyone I may have missed, there are so many people to give thanks to, but that will take multiple chapters.

Lastly, I would like to give thanks to my family, even when they probably didn't understand what I was researching, their support has been invaluable. Special thanks to Elis M De Jesus Figueroa for always being there, my life wouldn't be the same if I hadn't been blessed with your patience and love. Onward together to our new adventures. Finally, thanks to Phoebe (my Yorkie) for knowing when I should stop working to play some catch or just to get some love. All people know that you are mine but don't let Elis know about that.

Biography

Anthony Rafael Rivera Barreto was born July 1, 1991, in San Juan, Puerto Rico. Product of the public educational system, Anthony graduated in 2009 from the Ramón Vila Mayo High School in San Juan, Puerto Rico. That same year he started at the University of Puerto Rico Rio Piedras Campus (UPRRP), pursuing a bachelor's degree in Biological Sciences from the UPRRP College of Natural Sciences. During his bachelor, he worked in the Department of Mathematics and was a student tutor for the General Biology Course. During his senior year, he started doing research at the laboratory of Dr. Gary Toranzos Soria. Working with Dr. Toranzos allowed Anthony to learn about scientific research while working on understanding the *Enterococci* bacteria genus isolated from pristine areas from el Yunque. In August 2015, Anthony was accepted into the Biology Ph.D. program at UPRRP.

In pursuit of new opportunities and expand his research skills, in 2017, he joined Dr. José Rodríguez-Martínez laboratory. In Dr. Rodríguez-Martínez Lab, Anthony worked in an NSF-EPSCoR-funded research dedicated in understanding the transcription factor known as optix. The project grows to consider the evolution and binding specificity of the SIX transcription factors, the family of which optix is a member. He collaborated within the EPSCoR in multiple outreach activities and even contributed to designing and publishing a CURE course. During his graduate studies, Anthony was accepted to participate in the Gene Regulatory Network and development and Molecular Evolution advanced training courses in the Marine Biology Laboratory in Woods Hole, Massachusetts. The research done in Dr. Rodríguez-Martínez Lab allowed Anthony to present his work in multiple venues. As a graduate student, Anthony was also responsible for giving undergraduate laboratory courses while training the next generation of scientists.

Chapter 1: Introduction

Cells are the basic unit of life. They carry on multiple roles and functions encoded within the DNA. The genome is the entire DNA instructions required for proper cell function and determines each cell phenotype. In multicellular organisms, all cells share something in common: the DNA they carry within. The information stored in each cell gives rise to the plethora of phenotypes we observe. How this information is decoded and interpreted is critical for proper cell function. DNA-binding proteins are responsible for numerous cellular activities, one of which is to decode the information stored within the DNA. **Transcription factors** (TF) promote or repress transcription by interacting directly with DNA in a sequence-specific manner. Transcriptional regulation is critical for gene expression because the information stored in the genome is read directly from the DNA[1]. Understanding how TFs read DNA is a nontrivial problem that hinders our ability to understand gene regulation.

1.1. Transcription Factors

Transcription factors (TF) are sequence-specific DNA-binding proteins responsible for numerous developmental processes by regulating gene expression[2]. They are regarded as regulators by controlling cell differentiation and developmental process[3]. They function as regulatory molecules transcribed in the nucleus, made into functional proteins in the cytoplasm, and transported back to the nucleus to regulate gene expression[4]. TFs are encoded by unique genes that account for just a few percentages of the total numbers of protein-coding genes; some estimates predict that of all information stored on eukaryotic genomes, up to 10% are TF genes [4–6]. Transcription factor protein sequences have been highly conserved since the early metazoan, which suggest that gene regulatory networks (GRN) are also conserved since early animals[7]. To regulate gene expression, TFs must directly interpret the genome, and in doing so, they can decode the information stored in the DNA.

The regulation of any gene depends on the information found within the genome. A gene's spatial and temporal expression is tightly controlled by multiple elements that regulate transcription in coordination. These regions are known as cis-regulatory elements (CRE). They are responsible for establishing not only the expression of a gene but also for promoting the genomic environment necessary for transcription to occur or to be stopped[8,9]. There are multiple types of CREs; for example, promoters are regions recognized by the basal transcription machinery. Enhancer elements are responsible for promoting the expression of a gene. Transcription factor binding sites (TFBS) are found within the enhancer regions[4,10,11]. To control the range of enhancers, insulators work by blocking the interaction between enhancer and promoter while also being responsible for preventing the condensation of the chromatin[12]. A gene can also be repressed by regulatory elements known as silencers[11]. **(Figure 1.1B)**

To understand how transcription factors can bind to these elements, it is important to determine the DNA sequences they recognize[7]. Understanding the locations in which a TF binds within the genome provides new opportunities to expand our understanding of gene regulatory networks[6] while providing new opportunities to broaden and understand genetic analysis[7]. A TF binds to DNA by recognizing a 5-12 bases long sequence known as the TFBS or also known as the binding motif [7,13]. These motifs are found throughout the genome thousands of times, which means that the TF must be able to identify its cognate binding site successfully. This is a challenging task when considering that TFs are in a nuclear environment in which the concentration of DNA is high[6].

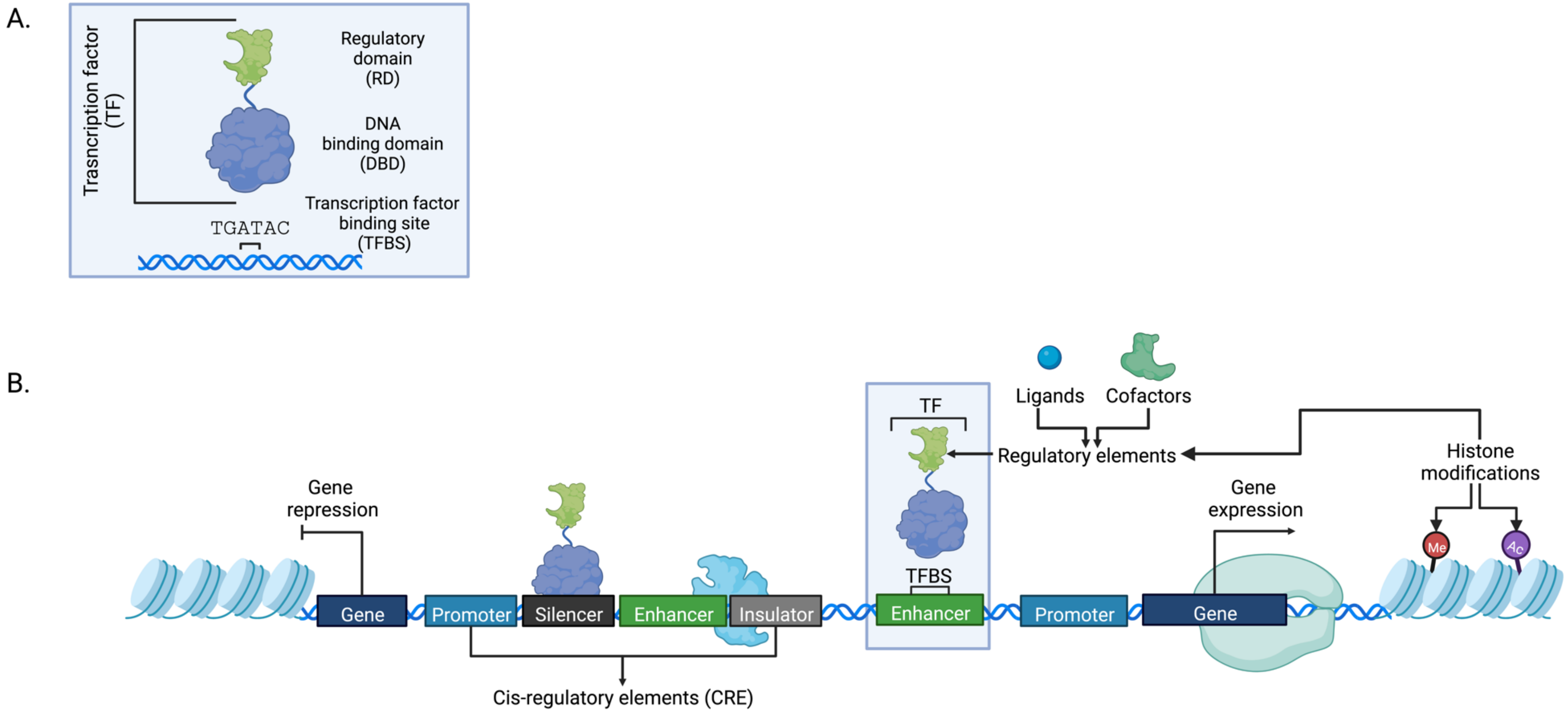


Figure 1.1: Transcription Factors Regulate Gene Expression: (A) Transcription Factors bind to DNA using their DNA Binding Domains (DBD); **(B)** their Regulatory Domain can interact with regulatory elements. This interaction allows for the promotion or repression of transcription by the recruitment or disassembly of the transcription machinery.

1.1.1. DNA:TF interactions

TFs have all the genome for potential transcription factor binding sites to be successful regulators, they must distinguish between functional and non-functional binding sites. Transcription factors have an intrinsic **affinity** towards their preferred sequences; based on the relative affinity observed for a sequence, one can then determine the **specificity** towards the recognized motif. The binding of a TF to DNA depends on the ON rate (k_{on}) in which the TF is bound to DNA and an OFF rate (k_{off}) in which the complex is dissociated[6]. Affinity is defined by the dissociation constant $K_D = \frac{koff}{kon} = \frac{[TF][S]}{[TF \times S\}}$, which takes into consideration the ratio of dissociation and association rates (k_{OFF}/k_{ON}) between the protein-DNA complex[6]. (**Figure 1.2**). Transcription factors are bound with low affinity toward the DNA while scanning the genome. It is not until they recognize their high-affinity sites that they immobilize themselves to exert their function[14]. Thus, a TF binds with high affinity to its target sequence, a trait that results from the interactions with the DNA bases and the backbone of the DNA[5]. This high-affinity sequence is usually represented as the consensus sequence for a TF. Even though most importance is given to high-affinity sites, it must be mentioned that medium- and low-affinity sites also influence gene expression[15].

Simultaneously, a TF must successfully identify a specific DNA sequence, meaning that a TF must be able to distinguish between different sequences crucial for regulating gene expression[16] (**Figure 1.3**). Specificity is defined as the ability of a TF to distinguish between different sequences. The specificity of a TF depends on multiple molecular interactions; is not necessarily restricted to the DNA [14,16]. Protein interactions with cofactors or different arrangements of dimeric formations can alter the specificity of a TF[2,14].

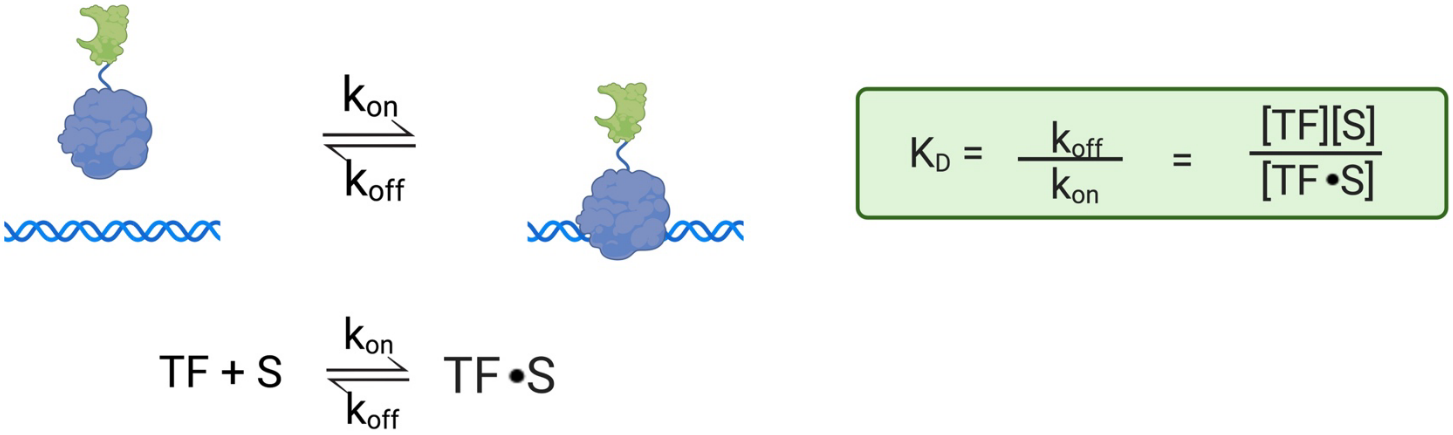

k_{on}
k_{off}
$K_D = \frac{k_{off}}{k_{on}} = \frac{[TF][S]}{[TF \bullet S]}$
$TF + S \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} TF \bullet S$

Figure 1.2: Transcription Factors interact with diverse affinities toward the DNA. The interaction between the transcription factors happens between different rates of interaction. The Off/On rates of interactions define the constant known as $K_{D,}$ which determines how a transcription factor interacts with the DNA. The value of the K_D establishes how strong is the affinity of a TF to a particular sequence. Low K_D (10-^{9}) are observed in sequences to which a TF has a higher affinity, while high K_D (10^{5}) are observed with sequences with low affinity.

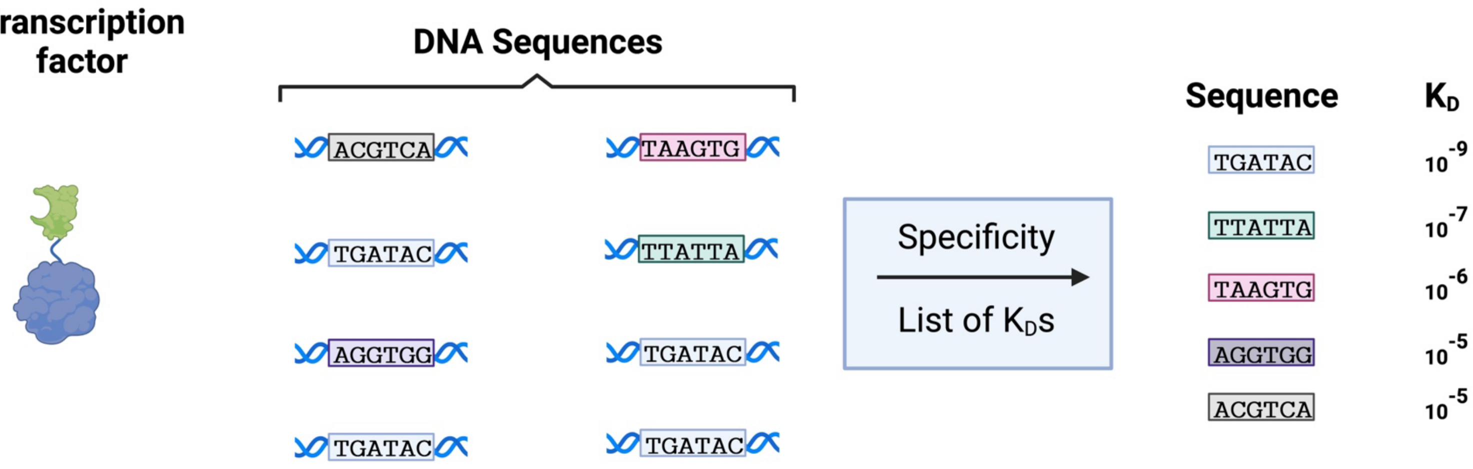


Figure 1.3: Transcription Factor must recognize specific motifs. A TF must be able to differentiate between multiple DNA sequences to bind specifically to its cognate binding site. This characteristic is known as the specificity of a TF, which is usually represented by the difference between affinities (Kd) for different sequences.

Binding sites are not unidirectional, as the recognition sites of an individual TF can have different orientations and or spacings when considering the formation of protein complexes[17]. These DNA-binding differences allow a TF to attain a wide range of functions to coordinate gene expression[18]. Furthermore, characterizing the specificity of a TF allows for the prediction of target genes that can provide insights into the connectivity between gene expression and gene regulatory networks[2].

1.1.2. Experimental methods to determine transcription factor specificity

To determine the DNA-binding specificity of a transcription factor, one can use an array of techniques that assess this in *in vivo* or *in vitro* environments. Using *in vivo* techniques takes into consideration the biological events in the cellular context. These assays consider the accessibility of the DNA entangled inside the chromatin[19]. The information obtained from these assays can help differentiate between different cellular environment that leads to other cell types. One example used to evaluate genome-wide TF binding is chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq)[20]. In this technique, genomic regions bound by a TF are first isolated by crosslinking and fixating proteins bound to DNA. After shearing chromatin, DNA-bound proteins are then selected by immunoprecipitation, utilizing an antibody specific towards a protein of interest. Bound proteins are then separated from the DNA, which is then isolated and characterized using high-throughput sequencing[2]. The data obtained from this type of assay provides insight into the framework of GRN by indicating which factors are more likely to regulate which genes[6].

However, the resolution obtained of the binding locations is insufficient to identify TF binding sites precisely, with regions being 100 or more base pairs long [6]. This is important since TFBS motifs are short base pairs; one can expect multiple potential binding sites inside the isolated sequences, with the possibility of multiple binding sites across all the genome[2,20]. This leads to thousands of binding events for a TF when it is expected for a TF to regulate just a handful of genes[20]. Therefore, the enriched regions isolated from ChIP-seq experiments, also known as peaks, must be analyzed by algorithms to search for enriched motifs within the resulting peaks[2]. This means that this technique is not a direct readout but a reflection of several processes and effects[21]. In addition, ChIP may not identify the correct binding motif of a TF due to other external factors such as chromatin structure, the presence of nucleosomes or chromatin modifications, and the requirement for partner proteins like cofactors[18,22]. Finally, the lack of specific antibodies for each protein to be studied limits the usage of this technique. However, despite these limitations, this technique has allowed the annotation of regulatory regions to understand TF-DNA interactions in an *in vivo* environment.

Another experimental approach to determine the binding specificity of a TF is by doing *in vitro* assays. There are multiple *in vitro* assays [6]. One example is the systematic evolution of ligands by exponential enrichment (SELEX). In this technique, a TF can interact with double-stranded DNA oligomers that contain a randomized region whose length allows a greater sequence combination (**Figure 1.4**)[23,24]. For example, when using a 20-bp randomized region, one can have 10^{12} sequence combinations[25] (**Figure 1.4A**). This vast array of sequences is what is known as a DNA library. The randomized section is also flanked by regions crucial for downstream procedures, such as PCR amplification, barcoding and sequencing adapters (**Figure 1.4B**)[23,24].

In a SELEX experiment (**Figure 1.4C**), a TF interacts with a DNA library and then is separated by electrophoresis on native gel or immunoprecipitated. TF bound sequences are purified and amplified using polymerase chain reaction (PCR). The enriched sequences are then re-used for subsequent selection rounds. The selected sequences from each round are then barcoded and prepared for next-generation sequencing (Figure 4B)[24]. In this process, it is also essential to characterize the starting library, as it can have biases in the composition of the sequences[24]. The results are then analyzed to determine the DNA binding specificity of a TF[26]. This type of *in vitro* methodology is helpful since only nanogram quantities of the protein are needed achieve a successful SELEX[23].

However, obtaining soluble proteins is a limitation of the procedure since TFs are challenging to produce, thus limiting the number of TFs that can be analyzed[2]. But still, numerous specificities have been identified using different *in vitro* assays. Yet, it has become challenging to predict in vivo genomic bindings using only the data obtained from in vitro assays[24]. Some even consider that studying the specificity via *in vitro* technologies alone cannot provide the whole idea of how TF do their regulatory work[20]. Moreover, others expect that *in vivo* binding experiments are more precise since they considers the organization of the chromatin[24]. Thus, a combination of both, *in vivo*, and *in vitro* approaches, provides insights into characteristics that play a role in the TF-DNA interactome[2].

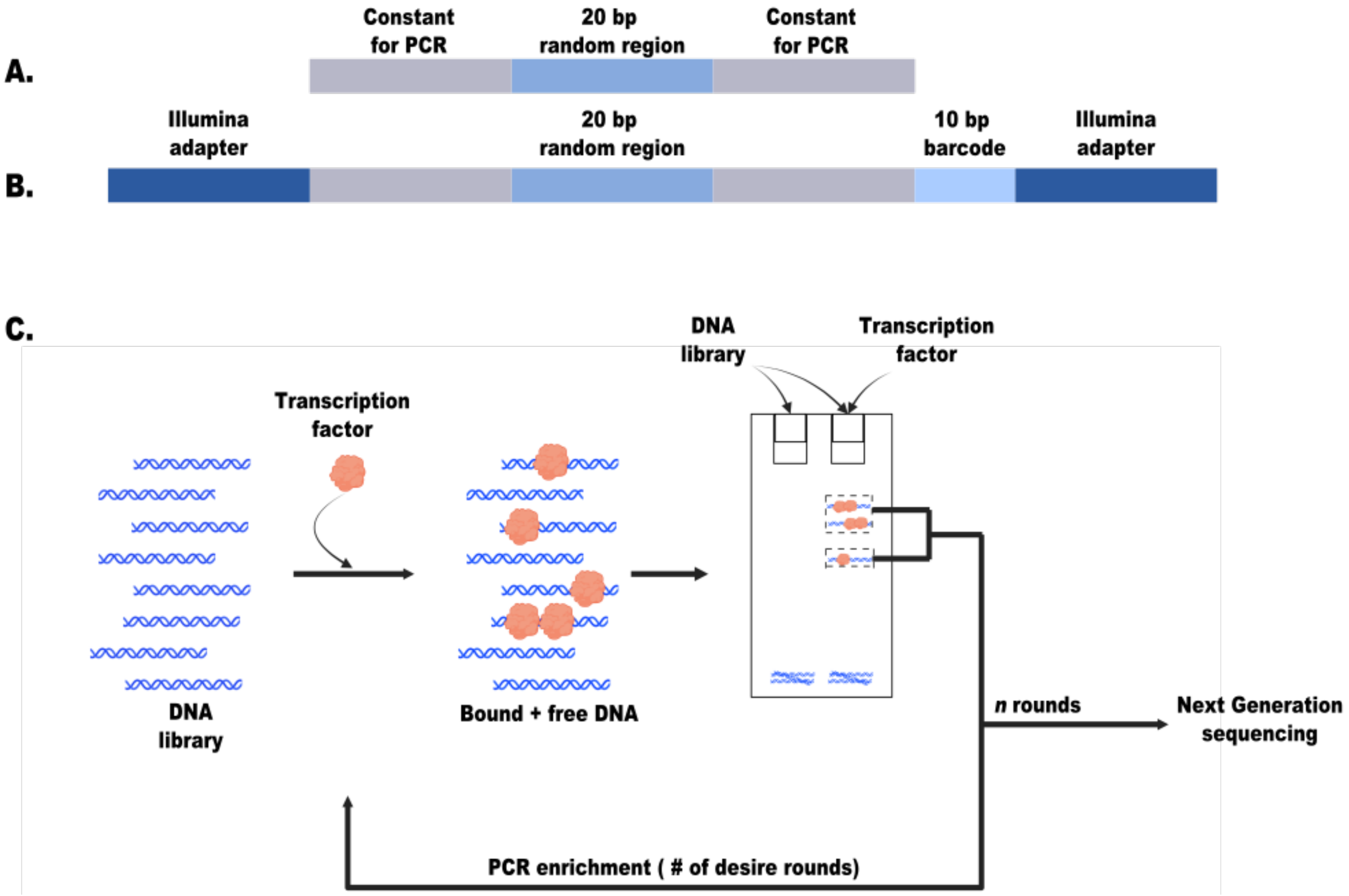
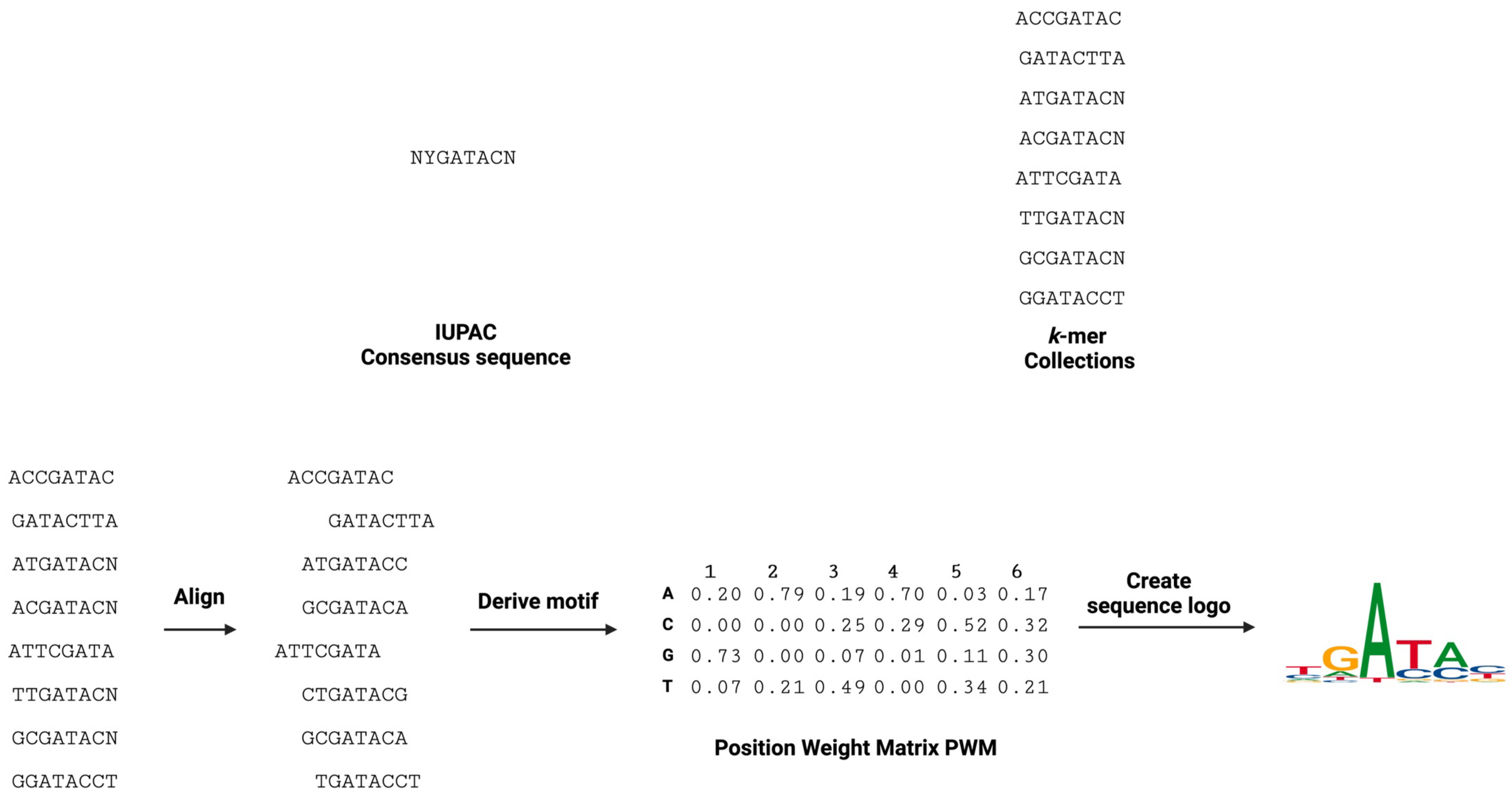


Figure 1.4: Transcription Factor specificity can be determined *in vitro* by SELEX-seq. (A) The DNA library used can have different randomized region lengths; a 20bp randomized region allows for 10^{12} possible combinations. (B) Schematic of the DNA library after selection and identification (barcode) and preparation of Illumina Sequencing, hence the adapters. (C) SELEX-seq methodology representation of how the procedure is done.

1.1.3. Representation of binding specificities

In vivo and *in vitro* techniques have enabled the capability to develop DNA-binding specificity models that are used for the representation of binding specificity (**Figure 1.5**). The specificity is represented as motif or position weight matrix (PWM)[18,27], which has been used since 1982[28]. It is crucial to establish that the usage of PWMs is a generalization of the concept of consensus sequences[28,29]. A PWM model assumes that each position in the binding site is independent, meaning that the presence of a nucleotide in one position does not affect the TF preference for another nucleotide at another position[6,18,27]. The PWM functions as an approximation of the true specificity, with the need to address how good of an approximation it is. This is based on the realization that some PWM models can have limitations and may not capture the true specificity of a TF[28]. The approximation considers that affinity is dependent on the genomic environment, DNA shape, or interactions with other cooperative proteins[7]. However, new models can consider these variables and improve the accuracy of generating a PWM that depends on the TFs being studied[7].

Additionally, some models use *k-mer* (DNA sequence of a determine *k* length) based representations that consider other models, and some can outperform PWM based models[18]. Using both methodologies has allowed scientists to interpret and explain the binding of TFs, and the subsequent development of several databases that contain both *in vivo* and *in vitro* binding specificities for many TFs[7,18].



	1	2	3	4	5	6
A	0.20	0.79	0.19	0.70	0.03	0.17
C	0.00	0.00	0.25	0.29	0.52	0.32
G	0.73	0.00	0.07	0.01	0.11	0.30
T	0.07	0.21	0.49	0.00	0.34	0.21

Figure 1.5: Representations of binding specificity. Groups of sequences bound by a transcription factor (TF) can be used to create a consensus sequence, represented using IUPAC notation. The group of k-mers themselves can be used to denote sequences bound by the TF. Bound sequences are aligned to create a motif, which indicates the probability of each nucleotide at every position within the binding site.

1.1.4. DNA binding and regulatory domains

TFs have conserved protein domains that affect how they interact with DNA and other regulatory elements. A TF has a regulatory domain responsible for interacting with regulatory molecules and recruiting the transcription machinery components. They can function by recruiting cofactors, participating in chromatin binding, or promoting nucleosome remodeling, while some can read histone modifications (Figure 1B). In contrast, the DNA-binding domain (DBD) is responsible for directing the TF to a specific DNA sequence (Figure 1B)[7].

1.1.4.1. Regulatory domains

Regulatory domains can be classified into activating or repressing domains. Both types of domains can interact with regulatory ligands promoting the assembly or disassembly of the transcription machinery. They can also interact with chromatin modification components that allow for the remodeling of DNA[5,7]. However, not all TF have effector domains and are almost entirely comprised of a single DBD. In these TFs the steric mechanism blocks other proteins from binding to the same binding sites[7].

1.1.4.2. DNA-binding domains

The DNA-binding domain specifies the genes to be regulated by the TF. There are multiple types of DBDs with a wide range of structures[5]. In eukaryotes, there are around 100 types of DBDs[7]. Most DBDs are identified by the homology found in these domains, and they are used to classify TFs into families[4,5,7,30]. For example, in mammals, the three largest families of DBDs are the C2H2-zinc fingers, homeodomains (HD), and helix-loop-helix (HLH)[5]. The DNA-binding specificity is usually conserved between members of the same family. However, similar DBDs can have distinct binding profiles toward DNA, where different family members recognize different core binding sites[31–33]. Even though you can have high homology in their DBDs, there isn't a universal code that can predict DNA binding specificities[20]. However, classifying TFs into families allows for an organized manner that helps to understand similarities between family members while highlighting the differences that make them unique.

1.1.5. Homeodomains

The HDs are the second most abundant family of TFs found in Metazoan[34]. HDs are expressed in all tissues found in plants and animals; for example, in animals, they are expressed from the early stages of development [35]. Among these are the HOX proteins, which are known for coordinating the development of the anterior-posterior axis of body development in Metazoan [5]. Within the HDs, the HOX proteins are just one among the 12 subfamilies that are members of this family.

Homeodomain transcription factors were first discovered in *Drosophila,* where the homeobox was discovered when studying homeotic genes [35,36]. Molecular studies on the DNA organization of the homeotic locus found a 180 bp element shared between *Antennapedia*, *Ultrabithorax*, and the *Fushi tarazu* genes, which raised the possibility of a common structural component[37]. The first homeodomain gene was cloned from *Xenopus laevis* in 1984 and named (*AC1*) [35], which is found active during neurulation[38]. It has been shown that the HD encoded in these regions allows for the sequence's specific recognition of genes that have precise spatial and temporal patterns[36].

The HD is composed of 60 amino acids encoded in the homeobox sequences. The three-dimensional structure characteristic of the homeodomains is a three-helix bundle preceded by an unstructured N-terminal arm[34–36]. The first two helices are connected by a hexapeptide and organized in an antiparallel order with specific contact with the DNA backbone (**Figure 1.6**)[39].

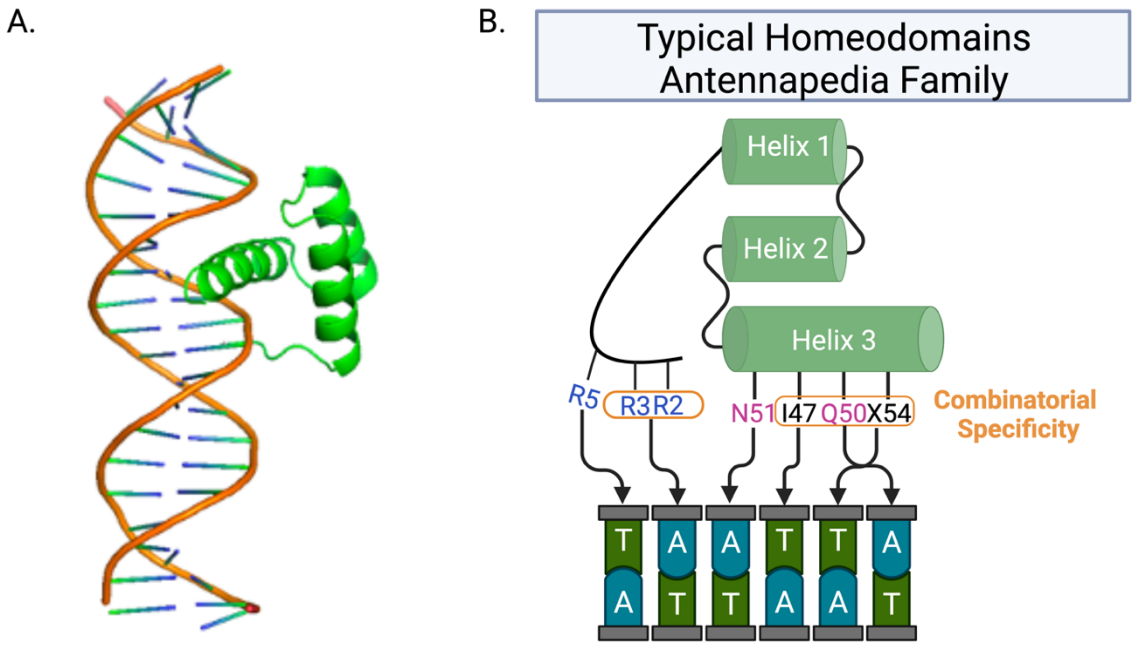


Figure 1.6: The homeodomain interacts directly with the DNA. (A) Antennapedia PDB:9ANT TF configuration when interacting with DNA, notice the recognition helix is found within the DNA's major groove. (**B**) Binding schematic of key DNA-binding residues interacting with DNA. Each key DNA-binding amino acid interacts with specific DNA bases that lead to the canonical TAATTA core motif associated with the HDs.

The recognition helix and the N-terminal arm confer the DNA-binding specificity of the homeodomain[39]. The third helix, also known as the recognition helix, interacts directly with the major groove of the DNA[34]. The specificity of this family is defined by the amino acid residues found at positions 2, 3, and 5-8 in the N-terminal arm and residues 47, 50, 51, 54, and 55 in the recognition helix (**Figure 1.6B**). These residues have been highly conserved since the Metazoan common ancestor, but they alone are insufficient to explain the high diversity of DNA-binding preferences observed in this family[40]. However, even though a high-sequence conservation is observed, duplication and divergence events can provide diversity to the DNA-binding specificity and the roles found through this family.[36,40]

The canonical motif usually associated with HDs is (TAATTA). Unsurprisingly changes in key residues provide diversity in the specificity observed in the family[34]. For example, position 50 has been shown to modify the DNA-binding specificity. The presence of lysine (K50) has been described to promote changes to the core motif, preferring a GG before the ATTA core (GGATTA), demonstrating that single residue changes allow for different specificity[36]

Sequence composition in the amino acids found within the N-terminal arm and the recognition helix are responsible for the varying specificity observed within the HDs. For example, the sine homeobox (SIX) is a K50 subfamily of HDs in which the recognition helix has a (I47N, Q50K, M54Q, and K55R) composition that changes the TAATTA core motif to a TGATAC core motif (**Figure 1.7**) [34]. In addition to the recognition helix, residues in the N-terminal arm can impact the DNA-binding specificity. The N-terminus interacts with the minor groove of the DNA and is characterized by the presence of positive residues (K and R). Residues 2, 3, and 5 play a key role in interacting with the first two nucleotides of the HD core motif (TAATTA)[34]. The N-terminus and the recognition helix composition are used to classify HDs into two categories, typical and atypical homeodomains, where atypical HDs have significant modifications than consensus typical HDs[34]. The SIX group is considered an atypical homeodomain, with characteristically amino acids in the N-terminal and recognition helix contributing to how this group interacts with the DNA. For example, the SIX has a negatively charged N-terminus arm that contrasts with the typical HDs, suggesting that this region may not interact with the minor groove of the DNA[35]. However, even when some differences are observable, the HDs have undergone little change through evolution; but they can recognize a diversity of DNA targets which contribute to the diversification found in the family[40]. The differences in DNA-binding specificity between

HDs provides another layer in which a sequence combination of residues influences the HD's contacts with the DNA[34].

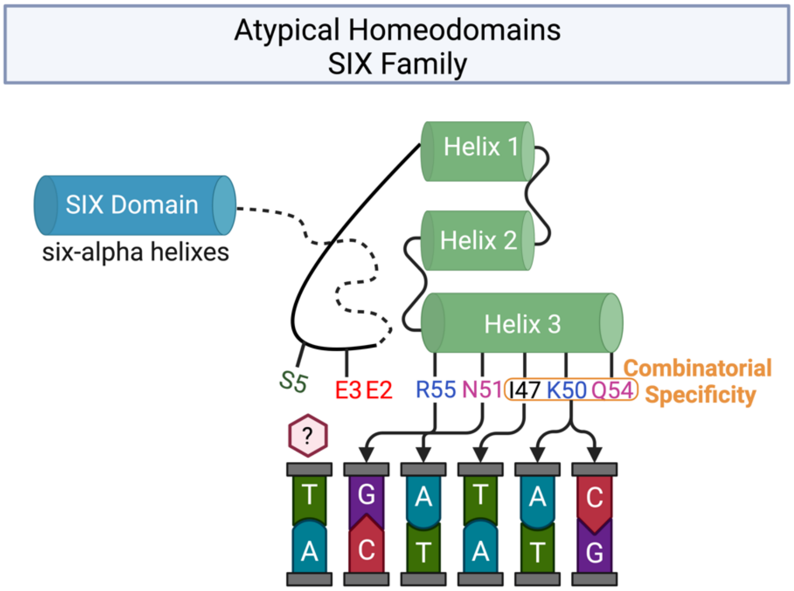


Figure 1.7: The SIX are atypical HDs. Negative residues within the N-terminal arm alter the mechanism by which the SIX interacts with the DNA. A combination of amino acids within the recognition helix demonstrates a variant of the canonical HD binding motif, showing a preference for a `TGATAC` motif. Which amino acid is responsible for interacting with the Thymine at position one hasn't been determined; however, the presence of an exclusive protein domain (SIX) could play a role in this interaction.

Interactions between other proteins can also alter the specificity. For example, HDs can bind to DNA as monomers or higher protein complexes like homodimers or heterodimers. The specificity of these complexes can change when compared to monomeric binding[36]. This shows that the specificity of TFs is complex phenomena and not limited to one specificity model. Understanding how homeodomains interact with DNA can explain how these transcription factors achieved their developmental roles in the organism. The HD family shows that there is great diversity in the developmental process they regulate and how they achieve them. However, there are multiple subfamilies of HDs some showing structural and binding specificity differences. This diversity shows that it is crucial to consider how TF change, allowing them to participate in new roles. Understanding the different HDs can enable us to understand how genomic roles have been achieved and established, with great interest in those HDs that are considered atypical.

1.2. The sine homeobox (SIX) family of transcription factors:

The SIX are a highly conserved TF family found throughout the animal kingdom. They carry on multiple developmental roles and are members of the homeodomain family. However, how they carry their DNA-binding roles has been elusive since they are considered atypical HDs, meaning they must adopt novel interactions to bind to DNA successfully. This family also carries a conserved protein domain exclusive to the family known as the SIX domain (SD). The SIX family is N-terminal to the HD and provides a novel way to understand how atypical TFs work and allows us to understand how developmental roles have diversified through evolutionary time.

1.2.1. Background

The SIX Family is a multigenic family that plays multiple roles in numerous developmental processes. The first role associated with them was when Milani[41] in 1941 noticed some strains of *Drosophila melanogaster* without eyes (**Figure 1.8**). This phenotype was not associated with any gene until 1994 when Cheyette[42] showed that an HD gene was crucial for forming the morphogenic farrow. In the absence of the HD gene, it would lead to apoptosis of the morphogenic farrow. This established that *so* or the *sine oculis* allele described by Milani was responsible and was able to determine the sequence of this TF. *Sine oculis* was the first SIX gene described, and its discovery was invaluable in determining the identity of other SIX members. In 1995, Oliver et al. described two murine SIX genes[43]. The genes were identified as *SIX1* and *SIX2* based on their homology towards *sine oculis*. The genes were first identified in the limbs of mice, particularly the digits, with both genes having distinct expression patterns in developing fingers. This article also proposed the name SIX, which originates based on being homeobox related to *sine oculis* or Sine oculis Related Homeobox (SIX)[43]. That same year Oliver et al. described another SIX member, in this case, *SIX3*[44]. Sequence analysis identified it as a SIX gene, and its name after the fact that it was the third SIX to be discovered. In this discovery, the SIX domain was noticed and described as flanking the HD's N-terminal. It was also seen that SIX3 exhibits a lower degree of amino acid homology than the other described SIX proteins[44]. In addition, it was shown that it could also be used as a molecular marker for the development of the anterior neural plate, with expression noticed in the brain and eyes. Subsequent studies also show that the overexpression of SIX3 in fishes could lead to the formation of ectopic eyes[45].

Another SIX was discovered in 1996 and was first recognized as AREC3 since it was isolated as a regulatory factor of K-ATPase[46]. When analyzing the cDNA of this factor, it was found to be homologous to the *sine oculis* from *Drosophila melanogaster*. The presence of an activator domain C-terminal to the HD was also noticed. Based on the homology to other SIX genes, this gene was renamed *SIX4/AREC3* as some speculation persisted about its role in regulating K-ATPase. However, different roles were observed in muscle and kidney development[46]. Further studies wanted to clarify the *SIX4* role in the development, and since the other SIX genes had been associated with nervous or eye development, a new approach was made. It showed that *SIX4* could be found in neural cell fate, suggesting that *SIX4* is involved in neural cell fate decision while raising doubts about the role of a K-ATPase[47].

Further studies about this role led to the discovery of another SIX gene named *SIX5,* which was observed to be expressed in muscle differentiation and other nervous developmental roles[48]. The remaining murine SIX to be discovered was first described in chicken and named *Optx2* (*optic*

SIX gene 2)[49]. It wasn't until this discovery that another SIX gene was found to be expressed selectively in the eye. This discovery led to the identification of the ortholog of *Optx2* in *Drosophila* and named *optix*[49]. This led to discovery of the murine homolog of *Optx2* which was then named *SIX6*[50]. All these discoveries completed the entire catalog of SIX murine genes, which are also found in other mammals. Another SIX gene has been described but is exclusively found in ray-finned fishes and is associated with *SIX3/SIX6* and is known as *SIX7*[51–53]. Its function is still under research, but it has been shown to play a role in zebrafish opsins. Seo et al.[54] completed the SIX catalog in *Drosophila*—showing that the SIX in fruit flies is composed of three distinct genes: known today as *sine oculis*, *optix,* and *six4*.

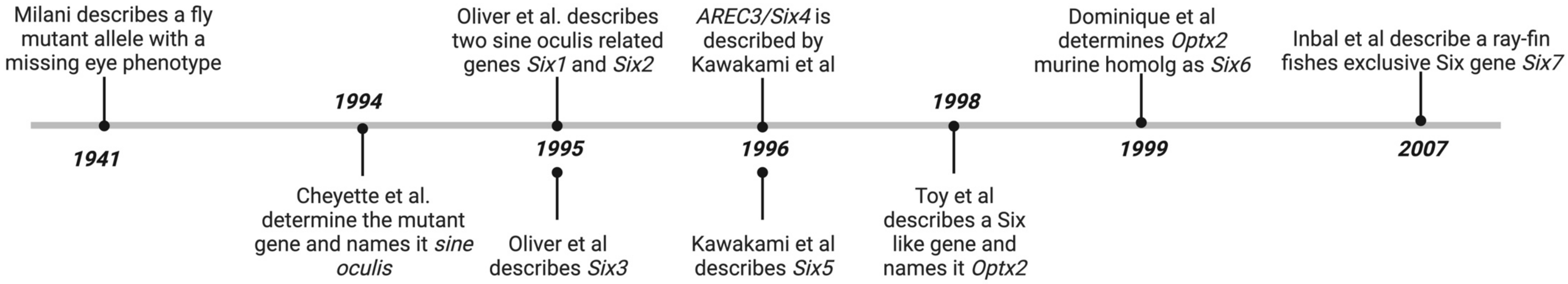
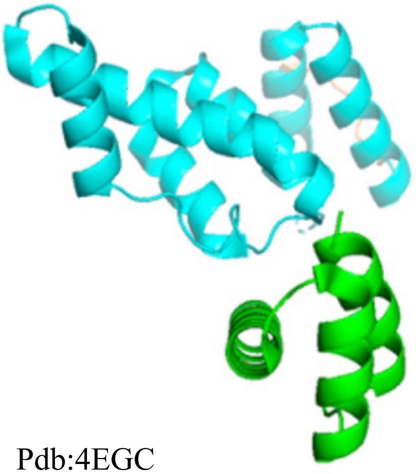

Milani describes a fly mutant allele with a missing eye phenotype
1941
1994
Cheyette et al. determine the mutant gene and names it *sine oculis*
Oliver et al. describes two sine oculis related genes *Six1* and *Six2*
1995
Oliver et al describes *Six3*
AREC3/Six4 is described by Kawakami et al
1996
Kawakami et al describes *Six5*
1998
Toy et al describes a Six like gene and names it *Optx2*
Dominique et al determines *Optx2* murine homolg as *Six6*
1999
Inbal et al describe a ray-fin fishes exclusive Six gene *Six7*
2007

Figure 1.8: The SIX transcription factor family discovery timeline. The SIX proteins have been studied for around 84 years, with the first descriptions by Milani. Since then, 3 SIX proteins have been described in *Drosophila* and 6 SIX proteins in Humans. Most SIX proteins were described in Murine and opened the door for discovering all SIX proteins through the Metazoan tree.

1.2.2. Structure

All the SIX proteins have specific characteristics that make them unique, there are conserved amino acid motifs with each protein domain that are useful for cataloging each gene. The SIX family are members of the homeobox class of transcription factors[34,35] and are atypical members with significant changes when compared to typical homeodomains. The substitution of positive amino residues (lysine and arginine) for negative residues like glutamate on the N-terminal arm of the HD, complemented with the substitution of glutamine for a lysine at position 50 in the recognition helix contributes to their atypical nature [34]. In addition, the presence of the SIX domain N-terminal to the HD is unique to this family, which contributes to the functional roles of the whole protein.

The SIX domain is an exclusive domain found in the SIX family and was found to be N-terminal to the HD[44]. Structurally the SD is arranged into six alpha-helixes (**Figure 1.9**), and the composition of this domain is highly divergent, specifically to the protein to which its associated[35,55]. The SD domain is 116 amino residues long; however, in the optix/SIX3/SIX6, there is a 4-residue insertion that is exclusive to this subgroup[56], highlighting the sequence divergence of this subgroup. Other traits can be associated with each subgroup, such as position 12 being subgroup-specific (valine to sine oculis/SIX1/SIX2, alanine to six4/SIX4/SIX5, and threonine optix/SIX3/SIX6)[56]. The SD has been associated with playing a role in protein-protein interactions with the binding partner of so/SIX1 cofactor Eyes absent (Eya)[55–57]. The formation of this protein complex is essential since it plays a role in preventing the degradation of SIX1, and complex formation is found to be within the first alpha-helix in the SD[55]. The so/SIX1: Eya interaction is also responsible for multiple developmental processes, and loss of the complex formation can lead to syndromes like Branchio-oto-renal syndrome (BOR)[57]. Studies about this syndrome have shown that six mutations lead to a reduction in DNA-binding[55,57]. It was noticed that some of these mutations were found to reside within the SD domain, which previously wasn't considered to be responsible for DNA binding[56]. However, three mutations were shown to diminish DNA binding and reside within an unstructured region of the last alpha-helix of the SD and N-terminal of the HD[57]. This characteristic linker region has been proposed to be involved directly with DNA-binding[57].



Pdb:4EGC

Figure 1.9: Structure of SIX1 SD and HD. Structure of the SD (blue) and HD (green) of the SIX1 protein. The SD domain is structured as 6 –α– helixes with a linker (not observable) between the last alpha helix and the HD.

The homeodomain of the SIX family doesn't share the basic residues found in the N-terminal of the HD **(Figure 1.10 & Figure 1.12**) [35]. This region canonically interacts with the minor groove of the DNA and has direct contact with bases within the binding motif[34]; the acidic nature of these residues means that the SIX HD must undergo novel configurations to bind to DNA properly[55]. The acidic nature of the N-terminal region is conserved on the SIX family, and it's also a helpful tool for identifying each subgroup. This trait was first described by Seo[54] , who noticed a tetrapeptide's presence, with sine oculis/SIX1/SIX2 sharing an **ETSY** motif. At the same time, six4/SIX4/SIX5 has an **ETVY; limits** to the sequencing didn't allow notice of the first two amino acid residues that composed the HD. The N-terminal is composed of a unique hexapeptide **(Figure 1.10**) that is subgroup-specific; for sine oculis/SIX1/SIX2, the known hexapeptide is **GEETSY**, while in six4/SIX4/SIX5 Is **GEETVY** in the case of the optix/SIX3/SIX6 the hexapeptide is more divergent having a **GEQKTH**[34,56]**.**

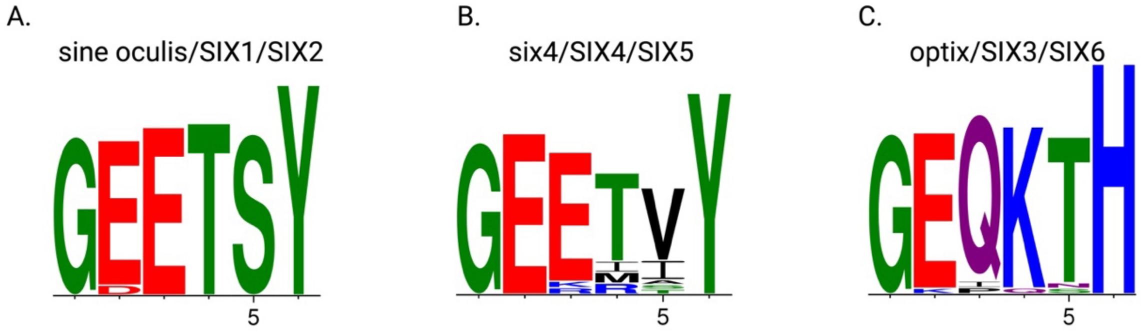


Figure 1.10: Conservation of the SIX hexapeptide by protein subgroup. Sequence alignment of 86 SIX sequences show the conservation of each protein subgroup hexapeptide. It can be noticed how each protein subgroup has a highly conserved motif based on the specific SIX protein subgroup. (A) corresponding to the sine oculis/SIX1/SIX2, the GEETSY motif (B) has a consensus motif of GEETVY with the last three positions showing high divergence, the remaining SIX subgroup optix/SIX3/SIX6 (C) has a corresponding hexapeptide motif of GEQKTH.

This motif can be found on early evolving SIX genes and has been conserved in evolutionary time, highlighting that although it is canonically atypical, it's functional, even when it's possible these residues do not form stabilizing DNA contacts[57]. However, some researchers believe that the SD can stabilize the HD binding capabilities. For example, for SIX4, it has been shown that the presence of the SIX SD is required for proper DNA binding[46] has been hypothesized that the SD linker between the SD and the HD stabilizes the HD allowing it to have direct contact with the DNA. For example, the linker region found in SIX1 (**AVGKYRVRRK**) is highly positive, having a net charge of +6[57], this motif which is highly conserved in all SIX members.

Another essential region found within the HD; is the recognition helix. The third alpha interacts with the major groove of the DNA, having base-specific contacts made in three positions[40]. The SIX have some changes that highlight their differences toward other homeodomains, one the presence of lysine within the recognition helix at position 50 of the HD[34,35,40,56]. Categorizing the SIX as a K50 HD, like *Bicoid,* another highly studied HD[58]. Other residues are also divergent and contribute to how this family interacts with the DNA, especially when considering the nucleotides being recognized by the TF (**Figure 1.7**). These changes will then lead to modification into the specificity and affinity attribute to the HD, changes that lead to new motifs being identified. A multiple sequence alignment for SIX proteins in *Drosophila melanogaster* can be observed in (**Figure 1.11**) and (**Figure 1.12**) for both the SD and HD, respectively.

1.2.3. Specificity

The DNA sequences identified by a TF are based on the dynamics between amino acidic residues and the DNA. The HDs are known to bind to a canonical `TAATTA` core sequence. However, about half of the HDs prefer other sequences[40]. In comparison, the SIX transcription factors bind to a `TGATAC` core motif[34,35,56]. How members from the same family can identify diverse binding motifs is based on the amino acids responsible for DNA-binding. Understanding these dynamics is helpful in comprehending the motifs recognized by the SIX. In the first position (T), eighty-nine percent of HDs identified thymine in this position, a trait associated with the ARG5 (N5) located on the N-terminal of the HDs and interacts with the minor groove of the DNA[34,36]. In the case of the SIX TF, there is no ARG5 in this position, and it still recognizes the same nucleotide. For the SIX, this position is highly variable and subgroup-specific; the sine oculis/SIX1/SIX2 subgroup has a Serine (S) while the six4/SIX4/SIX5 have valine (V), whereas the optix/SIX3/SIX6 keep a threonine (T)[56].

The second position of the motif (A) is a position that is highly diverse but with most HDs (83%) recognizing this nucleotide. This ability is usually recognized in the ARG2 and ARG3 found in the N-terminal of the HD[34–36]. The SIX genes are different in these positions, where both positions have glutamate (E), except for the optix/SIX3/SIX6 subgroup, which has glutamine (Q) at position 3[34,56]. These differences lead to the recognition of a guanine (G), and it has been observed that atypical homeodomains with an ARG55 show a preference for this nucleotide[34]; all the SIX proteins have an ARG55[56]. For the third position, the presence of an ASN51 shows a preference for adenine (A); the SIX have this residue conserved in all its members[34–36,56]. The fourth position is the most diverse since all nucleotides are observed; however, either ILE or VAL at position 47 shows a preference towards thymine (T); interestingly, the SIX recognize the same nucleotide while having an arginine[34]. It is believed that the SIX can recognize (T) based on the presence of a K50, a position also conserved by Bicoid (HD), which can interact with both positions 4 and 5 of the core motif[34]. There is a mix of observable bases for the fifth position in the motif; however, a (T) is usually observed[34]. Multiple research projects have hypothesized[35,36,40] that there are dynamics found within the amino acid residues from positions 47,50, and 54. Depending on the amino acids found in these positions are the bases to be recognized. In the case of Berger et al.[40], they propose that a combination of N47, K50, and Q54 shows a preference toward the motif `TGATAC` the same motif recognized by the SIX family. This combination also shows a tendency towards a `TGACAC`.

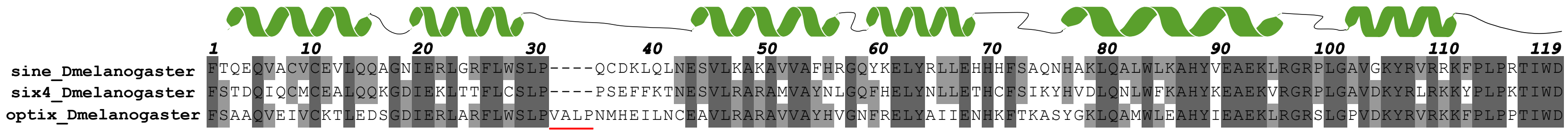
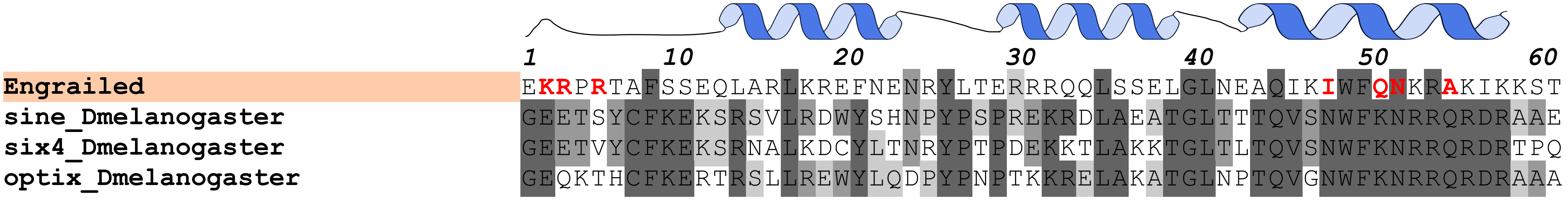

1 10 20 30 40 50 60 70 80 90 100 110 119
sine_Dmelanogaster FTQEQVACVCEVLQQAGNIERLGRFLWSLP----QCDKLQLNESVLKAKAVVAFHRGQYKELYRLLEHHHFSAQNHAKLQALWLKAHYVEAEKLRGRPLGAVGKYRVRRKFPLPRTIWD
six4_Dmelanogaster FSTDQIQCMCEALQQKGDIEKLTTFLCSLP----PSEFFKTNESVLRARAMVAYNLGQFHELYNLLETHCFSIKYHVDLQNLWFKAHYKEAEKVRGRPLGAVDKYRLRKKYPLPKTIWD
optix_Dmelanogaster FSAAQVEIVCKTLEDSGDIERLARFLWSLPVALPNMHEILNCEAVLRARAVVAYHVGNFRELYAIIENHKFTKASYGKLQAMWLEAHYIEAEKLRGRSLGPVDKYRVRKKFPLPPTIWD

Figure 1.11: *Drosophila melanogaster* sequences alignment of the SIX (SD) domain. Sequences alignment for SIX domain of the three SIX proteins found in D. melanogaster. Notice the four amino acid insertion (red underline) only found in optix. The approximate structure of each alpha helix location (green) and unstructured linker region is shown.



```
Engrailed             EKRPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKKST
sine_Dmelanogaster    GEETSYCFKEKSRSVLRDWYSHNPYPSPREKRDLAEATGLTTTQVSNWFKNRRQRDRAAE
six4_Dmelanogaster    GEETVYCFKEKSRNALKDCYLTNRYPTPDEKKTLAKKTGLTLTQVSNWFKNRRQRDRTPQ
optix_Dmelanogaster   GEQKTHCFKERTRSLLREWYLQDPYPNPTKKRELAKATGLNPTQVGNWFKNRRQRDRAAA
```

Figure 1.12: Homeodomain sequence alignment of the SIX TF in *D. melanogaster*. Sequence alignment of the HD of SIX proteins and a typical HD (Engrailed). Residues responsible for DNA contact are shown in red.

The final position (A) has observed diversity and is believed to depend on the amino acid residues found at positions 47,50, and 54[34,40]. Position 50 seems to be one of the key residues in differentiating distinct DNA sequences[35,36;] based on its flexibility, it can interact with multiple bases while also influencing other amino acid residues like R54, which interacts with the DNA backbone[34,40]. It is interesting how the dynamics found within the HDs modulate the protein-DNA interaction, showing how combinatorial effects play a role in the sequence to be identified by the HD.

The SIX TFs also have another characteristic that is particular and is the presence of an extended motif. For example, it has been reported that the SIX can bind to sequences like TCAGGTTC, TGATAC, GGGTATCA, or GTAANYNGANAYC/G for sine oculis [34,40,46,47,56,59]. But in 2005, Pauli et al. reported that 5' of the core motif there is an extension of the core motifs, and mutations within this 6bp extension can prohibit DNA-binding[60]. The motifs recovered by some authors show a lack of this extension or the absence of the core motif. Some SIX transcription factors seem to rely on interactions dependent on the flanking region of the core motif. This characteristic was demonstrated for sine oculis, but it has been shown that Six1, Six4, and Six5 can bind to the MEF3 promoter, which has a consensus sequence of GAAACCTGA[46,47,56,59,61]. The sequence observed is like the flanking region identified by Pauli et al., this could mean that it could be possible that Six4/Six5 can also recognize the flanking region described by Pauli et al. and that the need for the flanking region is not limited to sine oculis. In the case of Six6 and probably Six3, an extension C-terminal of the HD plays a critical role in DNA binding. The extension was shown to be vital for Six6 to bind two different promoters. The research done by Hu et al. shows that Six6 can bind to both the MEF3 promoter and the Pitx2 HD consensus motif (GGATTA)[62], while Six2 can only bind to the MEF3 promoter. By doing multiples construct in which the C-terminal was exchanged between both factors, they demonstrated that when Six2 carried Six6 extension, it was now capable of recognizing the Pitx2 motif. The substitution of the Six6 extension with the one of Six2 showed that Six6 could only recognize the MEF3 sequence. It was then demonstrated that the C-terminal of Six6 is critical for the specificity of some sequences identified by Six6 and not Six2 and that by substitution, Six2 could recognize Six6 sequences when having the Six6 HD extension. The presence of a critical HD extension in the optix/SIX3/SIX6 is not only observed in Six6 since Weasner and Kumar demonstrated the presence of extension exclusively found in *Drosophila's* optix[63]. In this case, the absence of the C-terminal extension allows optix to carry on repressive roles in tissue differentiation, and in its absence, eye formation could be promoted. It shouldn't surprise us that the extension in the optix/SIX3/SIX6 subgroup was conserved and exclusively found within this subgroup. The diversity of the binding capabilities of this family shows how members of the same family can diversify their functions. How these capabilities have diversified and changed through time can provide insights into how the functions of these TFs have changed and have been coopted to new phenotypes.

1.2.4. SIX proteins roles in development

The regulation of each species genome has been optimized for millions, if not billions of years. Today's phenomes result from multiple evolutionary events that led to the regulation we observe. Transcription factors were recruited into these developmental pathways and are coopted into new networks as time passes. Understanding the history of a TF provides insight into how regulatory mechanisms have originated and specialized in the numerous developmental processes we recognize today. How the SIX family has evolved and diversified can give us information on their effects on multiple phenotypes, some could have originated since the first animals.

The evolution of animals is an ongoing process that started billions of years ago. Their last common ancestor had a sister group in *Porifera* (sponges). There is some discussion if this is true; however, recent publications state that sponges are the most ancestral multicellular organism and sister to all animals[64]. Sponges are organisms with no real tissues or systems as we know them. Surprisingly, even though they are the simplest organism, SIX proteins exist within their genome. For example, the highly studied sponge known as *Amphimedon queenslandica* has a SIX-like protein, and observation is seen in other sponges like *Spongilla lacustris*[65]. This SIX-like gene can't be placed within any of the three major subgroups since its sequence is highly diversified. However, it still corresponds to this family[66]. How a SIX protein works within sponges is interesting since the family is related to multiple roles, but their first associated phenotype was the development of the eyes.

New reports have shown that sponges are not as simple as previously believed, as sensory cells have been reported[65,66]. For example, *A. queenslandica* has a pigmented ring that helps steer the cilia of the sponge larva by using a blue light-sensitive cryptochrome[66,67]. It has been hypothesized that these cells function as the precursors of future neural machinery[65]. In addition, it is also interesting that partners known to interact with SIX proteins are also found during this period. The retinal determination gene network is responsible for developing sensory organs and has proteins like Pax, Six, Eye absent, and Dachsund. New reports have shown that some members of this network can be found in sponges; for example, Hoshiyama et al. (2007) identified members of the *Pax* gene family in sponges, comb jellies, and jellyfishes[68], but were unable to determine if there was any interaction between Pax and Six. Fortunato et al. (2014) later demonstrated the presence of Eye absent (Eya), an essential co-factor of sine oculis[66]; however, not all sponges analyzed showed the presence of Eya, which they believe suggest that some sponge lineages lost this cofactor. Their work also mentions that some interaction between Pax and SIX has been observed, but current techniques cannot test this. However, even though it is challenging to determine interaction in sponges, it is currently known that both carry synergistic interactions[61]. This establishes that during this period, both proteins are present in sponges[66,68].

Also, Brodbeck and Englert (2004) reported that the network established by Pax, Six, and Eya was coopted in nephrogenesis[69]. Suggesting, as the authors state, that since this interaction is so evolutionary conserved is not surprising that evolution has decided to redeploy the network into new developmental process while also bringing new components into the combination [69]. All these studies show that the phenotypes associated with the SIX family appear to have originated in some way since sponges and shared with the last common ancestor of animals. It is also interesting how some of the roles associated with the SIX TFs, can be traced back to the sensory cells observed in sponges. And how what seems to be protein interactions in sponges are now highly sophisticated gene regulatory networks that can be evaluated in multiple developmental processes in multiple organisms.

Another organism that carries SIX proteins and that are highly divergent proteins are found in the nematode *C. elegans*. In the genome of *C. elegans,* there are 4 described SIX proteins not named with common nomenclature but are known as ceh-32, ceh-33, ceh-34, and ceh-35 (unc-39)[63 70–72]. The functions of these proteins are not fully understood, with most of the research being done in ceh-32, ceh-34, and ceh-35. The first (ceh-32) has been located during embryogenesis in the hypodermal and neural precursor cells of the head with abnormal head morphology in RNAi experiments[72]. In work done by Dozier et al. (2001), it was established that ceh-32 is SIX3-like, and works in coordination with vab-3, an ortholog of Pax-6[72], which we have mentioned previously is a known SIX co-factor [66,69,70,73]. For ceh-34, it is essential in embryonic and early larval development, while Amin et al. (2009) showed that for proper specification of non-muscle coelomocytes in the mesoderm[71]. In their work, they also identified Eya and that it works in synergy with ceh-34 and interaction reported previously in this work with multiple developmental processes. The function of ceh-35 (unc-39) is not completely determined, but Amin et al. (2009) mentions that it is essential for cell migration in the mesoderm. While Yanowithz et al. (2004) demonstrated that mutation in this protein can lead to defects in mesodermal and ectodermal cells that are precursors to both muscle and neural cells, they also mention that this gene is like the six4/SIX4/SIX5 subgroup of SIX genes[74]. The remaining SIX-like gene ceh-33 is an understudy, and no phenotype has been associated with it[70]. It should be mentioned that the SIX proteins found in *C elegans* are highly divergent and categorizing these genes into the SIX subgroups is not an easy endeavor; it relies mainly on some sequence similarity and overlapping functions with other SIX homologs.

The three canonical SIX proteins (sine oculis, six4, and optix) can be traced before the divergence of Bilateria. Genera like *Cnidaria* (jellyfish) and *Ctenophore* (comb jellies) have three SIX members. Tracing the origin of these genes before the Urbilaterian, the hypothetical last common ancestor of Bilateria and Cnidaria separated but after a series of duplication events that predate Sponges from the other animals around 980MYA[75].

How the functions of these TFs have changed through evolutionary history is an important approach to understand how TF are coopted. Especially when considering the multiple detrimental phenotypes associated with mutations in this family. The following sections will describe the functions observed in the SIX proteins.

1.2.4.1. sine oculis/SIX1/SIX2

The sine oculis (so) protein is the founding member of the SIX Family. It was in 1941 that Milani described the first mutants associated with it[41]. Afterward, much research was done based on the observations done by Milani and the phenotype that led to naming the factor sine oculis. Although some SIX proteins are observed in sponges, they are too divergent to be considered part of any SIX subgroups. Therefore, it is believed that the SIX proteins observed in *Drosophila* and other organisms must have originated from duplication events of a single ancestral SIX protein, which probably is the one found in sponges. This single ancestral SIX protein diverge into three SIX proteins before the separation of Bilateria, that are identified in Cnidaria [75,76].

In the case of Stierwald et al. (2004), they identified the expression patterns of all SIX genes in two different jellyfish with and without eyes[75]. It was found that expression of *sine oculis* was observed in the muscle layer of both jellyfish and like the expression observed in non-neural mice Six1 and Six2. For jellyfish with eyes, it was also shown that *sine oculis* was expressed during the regenerating process of the eye while showing reduced expression after regeneration, highlighting that the role is concentrated on the reformation of the eye and not its maintenance. In the case of jellyfish without eyes, they could localize the expression of *sine oculis* to the tentacle bulbs and other types of muscles. In another jellyfish, the work done by Hroudova et al. (2012) also observed a relationship with sensory organs like the eyes while also being involved in other myogenesis roles[76].

The conservation of sine oculis roles in sensory organs can also be found in planarians, whose eyespots are one of the most uncomplicated prototypes of eyes. The works of Pineda et al. (1999) showed that sine oculis is responsible for the maintenance of photoreceptor cells and eye regeneration in these animals[77]. Their work also showed that inhibiting the expression of *sine* *oculis* using dsRNA injections resulted in the absence of eye formation. The injection was necessary to be reapplied after three weeks. If not, the eye will regenerate as normally, demonstrating that planaria sine oculis is responsible for eye development and tissue regeneration if needed.

In the case of sine oculis in *Drosophila*, Kumar's work (2009) encompasses most of the research done about sine oculis in flies[56]. Kumar (2009) has shown that sine oculis is present in the entire visual system during embryo development, with expression observed in the anterior cells of the developing morphogenic farrow of the imaginal disc. This makes sine oculis a member of the eye specification for retinal determination by interacting with known cofactors like Eyeless, which is responsible in initiating *sine oculis* expression. The interaction that sine oculis does with other cofactors is critical for its role in development, a characteristic that seems to have been conserved through time[66,78]. A trait also conserved in the products of the gene duplication event that happened to *sine oculis* that gave rise to both SIX1 and SIX2.

The paralogs of sine oculis are the product of a duplication event sometime in the metazoan evolutionary history and are named Six1 and Six2; both proteins were first reported by Guillermo et al. (1995) in the connective tissue of the digits of mice[43]. The expression was seen in the limbs with different expression patterns, with Six1 proceeding from posterior to anterior of the limb, while Six2 did so in an anterior to posterior. Further research by Lacleaf et al. (2003) showed the effect of mice deficient of Six1[79] resulted in mice dying at birth with noticeable skeletal defects and muscle hypoplasia. In addition, they report that mice neonates were born lacking kidneys and thymus. The authors mention that these phenotypes are also seen in mutants of Eya, which is a critical cofactor in the SIX Family. The relationship of both Six1 and Six2 compared to SIX1 and SIX2 in nephron progenitors was done by O'Brien et al. (2016), highlighting how differential regulation contributes to the species differences[80]. In the case of Six1/SIX1, they noticed that Six1 expression was required in early metanephric development and stopped on the first round of branching. In contrast, SIX1 expression extends beyond this point while overlapping with SIX2. The localization of Six2/SIX2 was observed similarly in both mouse, and human kidneys, with transcriptional targets shared between them, except for a SIX1 target of SIX2 in the fetal kidney. They also demonstrated that both orthologs have autoregulatory activity with cross-regulation only seen in humans. Jinshu et al. (2022) demonstrated that their roles are not interchangeable; Six1 cannot rescue Six2 deficient kidneys, proposing that different protein complexes promote the selectivity of both transcription factors[81]. The importance of Six2 in renal development has been highlighted in the works mentioned by Logan et al. (2021), with renal hypoplasia observed in Six2KO[82]. While also noting that a chromosomal deletion of SIX2 is responsible for frontonasal dysplasia syndrome, leading to the abnormal development of the head and face. It has also been reported that Six1, in coordination with Six4, is responsible for male sex determination and the development of the gonads by having essential roles in precursor cells, with smaller gonads in double knock-out embryos[83]. The BOR syndrome has been associated with SIX1 mutations in which it is unable to bind to its cofactor EYA1[55,78]. Structural analysis done by Patrick et al. (2013) determined that a single alpha helix is responsible for binding with EYA1 and that a single mutation in this helix is enough to disrupt the interaction[55]. While the same research group also noticed that mutations on the sixth alpha-helix of the SD could affect the DNA binding of SIX1, demonstrating that some DNA binding capability is observed within the SD[57].

Recently much research has been done on both SIX1 and SIX2, especially their role in cancer development. Christensen et al. (2008) mentions that SIX1 is responsible for the progression of the cell cycle, with overexpression linked to an acceleration to enter the S phase, which leads to an accumulation of SIX1 and possibly allows the activation or repression of other genes[84]. It is also mentioned that SIX1 overexpression is observed in around 5% of breast cancer, with 90% of metastatic lesions also demonstrating SIX1 overexpression. SIX1 has also been mentioned to be responsible for numerous other types of cancers[82]. SIX2 role in kidney development has been highlighted in this work. Hence it shouldn't be surprising that an increased level of expressions of SIX2 can be associated with numerous nephrotic cancers, for example, a pediatric cancer known as Wilms tumor[82].

The developmental roles of sine oculis and its paralogs are numerous, and some can appear to be highly conserved since the beginning of animals. Most works establish that the critical component of their developmental roles is the dynamics of how these proteins are recruited. With the recruitment of cofactors responsible for their interactions with multiple transcriptional targets, SIX1/SIX2 interaction with each cofactor is believed to occur within both the 5' and 3' flanking regions of both the SD and HD[81]. They are the most variable regions across all the SIX proteins and could explain the uniqueness of the work done by each factor.

1.2.4.2. optix/SIX3/SIX6/SIX7

The diversity associated with the SIX proteins allows us to recognize the other major SIX subgroup is optix, the original member, and its paralogs SIX3 and SIX6. In addition, reed fishes have an extra SIX protein associated with this subgroup which is believed to have originated from an independent duplication event exclusive to this group[51]. Regarding optix, Stierwald et al. (2004) demonstrated that optix is expressed in the tentacle bulbs of jellyfish, with a strong presence in the eyecup and nearby corneal tissues[75]. The same expression pattern was observed in another jellyfish research by the same group. At the same time, Hroudova et al. (2012) identified that optix could be involved in forming neural structures and maintaining and releasing progenitor cells responsible for regeneration[76].

In *Drosophila*, much work has been done regarding its role in eye development. Some have shown that ectopic expression of optix leads to the formation of ectopic eyes in flies[85], a characteristic observed previously by Oliver et al. (1996) with optix paralog Six3[45]. The role of optix in eye development is restricted to cells responsible for the development of the morphogenetic furrow while also being responsible for differentiating photoreceptors[56]. The retinal developmental network has both optix and sine oculis as components, where both are regulated by Eyeless[56]. A particular trait only observed in *Drosophila's* optix is the presence of C-terminal segments that are believed to have repressive features. The work done by Weasnar and Kumar (2009) found within the C-terminal the presence of five-segment, with only two regions being conserved exclusively in *Drosophila*. Their functional work demonstrated that one of these exclusive regions is necessary for optix to function as a repressor[63]. This is an example of how orthologs can differ between species and carry species-specific functions determined by these differences. The expression of optix in the eye is a dynamic interaction with numerous other factors. The work of Dominguez et al. (2015) located optix in the anterior dorsal strip of the imaginal eye disc, being responsible for establishing the anterior borders with another factor engrailed[86]. They believe that optix and engrailed might have mutually repressive interactions between them, where overexpression of optix leads to engrailed suppression and vice versa. Further work demonstrated additional roles of optix, not in the eyes but in the wing[87]. The work done by Al-Khatib et al. (2017) confirmed that optix is required to develop the anterior region of the wing while also establishing the haltere, a balancing organ in two-winged flies, which is regulated by optix in coordination with decapentaplegic by suppressing the development of sensory bristles. The role of optix within the wing is interesting since most phenotypes associated with it are based on its role in eye development. Surprisingly some extinct midges' fossils show that their wings have an organ that resembles the surface of a compound eye[88]. This probably suggests the presence of a light-sensing organ within the wing that could have also been regulated by optix, which was then coopted in other pathways even when the organ phenotype was lost.

The role of optix in wings is especially interesting within the genus of *Heliconius* butterflies. The work done by Reed et al. (2011) demonstrates that within this genus, mimicry is used to alert its prey about attacking or eating the butterfly. This mimicry is developed with diverse color patterns conserved within different species from the same genus in the same geographical space[89]. For example, *Heliconius erato* and *Heliconius melpomene* shared red color pigmentation on their wings. This red color pigmentation was found to be regulated by optix[87,89], which aminoacidic sequence was described to be synonymous and highly invariable between the *Heliconius*[89].

Further works demonstrated that optix knockouts resulted in the loss of the red color pigmentation in the cells where the knockout was done, with observed changes in the wing scales compared to the wild type[90,91]. How optix was coopted from eye development in *Drosophila*, to being the master regulator of red color pigmentation in the *Heliconius* butterflies demonstrates how TFs can be coopted into new roles specific to new species. A duplication event of optix led to the origin of both SIX3 and SIX6, which have multiple new roles in diverse phenotypic pathways. These two paralogs are major members responsible for developing the retina showing multiple expression points in the developing eye, with first localization in the optic vesicle and stalk with subsequent expansion of neural functions[56].

The SIX3 protein has been associated with numerous alter activities associated with protein mutations. Like optix, Six3 has been observed to be responsible for the development of ectopic retinal tissue and material; Six6 can induce retinal but not lense tissue [56]. Six3 is responsible for promoting the expression of rhodopsin, a visual pigment found in light receptor proteins with a reddish-purplish associated color. Holoprosencephaly is a severe malformation in which the brain cannot develop both hemispheres properly. This syndrome has been associated with mutations within SIX3[56,82,84,92–94]. Some protein interactions have been associated with SIX3, with SIX3 being the only family member that does not interact with Eya while also interacting with the GEMIN, an inhibitor of the cell cycle[84,95]. This interaction has been conserved within the C-terminal of both SIX3 and SIX6[95]. The work done by Turcu et al. (2019) demonstrated this interaction within the C-terminal and GEMIN, using electrophoretic mobility shifts assays they demonstrated that SIX3 could bind to both DNA and GEMIN, working antagonistically by regulating proliferation and differentiation of cells[95]. For example, the interaction with GEMIN is required to repress the ability of cdt1 to initiate cell replication[84,95]. Interaction with sonic hedgehog has also been implicated with schizencephaly, while the same cooperation can also be observed in repressive roles by suppressing signaling produced Wnt[82]. Further research has highlighted that, as with other SIX members, this factor has also been associated with multiple types of cancers[82,84,95]. Surprisingly the role of Six3 in cancer has been linked to suppressive functions, in which overexpression of Six3 appears to play a tumor-suppressive role with better survival outcomes in different types of cancers[96].

Six6 is quite different from Six3, resulting in size differences and possibly functional differences. For example, Turcu et al. described these differences highlighting that the N-terminal regions of Six6 is only eight amino acids long, lacking 78 residues found in Six3, which are in the majority Glycine-rich regions. The C-terminal, shown to be critical for GEMIN binding, is highly variable between proteins, with only the first eight and last fifteen amino acids being conserved between both paralogs[95]. This region has also been associated with the specificity of Six6, where modifications reduce the specificity of the protein by being unable to recognize some promoters[62]. The remaining domains, SD and HD, are highly conserved, with the HD having only one amino acid difference between it and Six3. Six6 is found on multiple tissues, high expression is observed within the retina, and overexpression of Six6 in *Xenopus laevis* leads to bigger eyes[56,84]. Besides the eyes, Six6 has been restricted to the forebrain with expression in the developing hypothalamus and pituitary[82]. Given the highlighted role of Six6 on the eyes, it isn't surprising that most disease-associated phenotypes are related to the eye. For example, glaucoma has been linked with SIX6 mutations; the disease is the second most common cause of blindness worldwide[97]. The mutations have been predicted to hinder the binding specificity of SIX6. Lastly, one member of this subgroup is now exclusively found within the ray-finned fishes and is named Six7. This factor is predominantly expressed in the cone photoreceptors, where deficiency in zebrafish leads to loss of expression of green opsins[51]. The green opsin is controlled by the *rh2 green opsin* gene, which shows reduced expression in Six7 deficiencies. However, it has been shown that this feature can be rescued by the overexpression of Six6b in zebrafish, highlighting a redundancy or compensatory role between both factors[52].

The optix/SIX3/SIX6/SIX7 group are diverse proteins with multiple cooption events throughout their evolutionary history. Acquiring roles that go from developing the sensory organs in jellyfishes to mastering the red wing color pigmentation in the *Heliconius* butterflies. The paralogs have become critical components of the nervous system. One paralog has demonstrated to have retained the developmental roles in the eye, to which diseases have been described in human eyes. While the other has become central in regulating proper brain development. These diverse functions, which have been shown to be regulated by the SIX family, highlight their importance in cell differentiation and specialization spanning different tissues.

1.2.4.3. six4/SIX4/SIX5

The remaining subgroup within the SIX family proteins comprises six4 and its paralogs named SIX4 and SIX5. Like previous subgroups, it is first found in *Cnidaria* and other jellyfish. Work on jellyfish done by Stierwald et al. (2004) noticed that expression is also found in the tentacles, and it is absent in the eyecup. Interestingly they also tried to isolate six4 in a jellyfish without eyes (*Podocoryn sp.*) and were unable to do so, suggesting that six4 may have been lost in this type of jellyfish[75]. They were also able to detect a high expression within the gonads. The research done by Hroudova et al. (2012) showed an association with regulating Na+/K+ATPase transporters[76], like previous roles associated with six4 during its discovery. The role of six4 is interesting since it was also detected in the statocyst, a balancing sensory receptor found in some aquatic animals, in this case on jellyfish[76]. six4 has also been detected to interact with the Eya4 cofactor, like some SIX proteins. They also detected six4 within the gonads, especially during the release and maturation of oocytes. In *C. elegans*, the homolog to six4 (unc-39) has been linked to neural disruption migration[84].

In the case of *Drosophila,* it seems to follow *Cnidaria's* functions, especially regarding the development of gonads, emphasizing this role to be conserved between animals. For example, Kirby et al. found six4 expression in the developing head, especially the precursors of muscle tissue (myoblast)[98]. Furthermore, they successfully detected six4 in the gonads, noticing that defects in six4 leads to defects in the gonads. Other roles were also described mainly within *Drosophila's* mesoderm. The work done by Clark et al. (2006) demonstrated that six4 was required to develop lateral and ventral muscles while also being a pattering mediator[99]. They also mentioned that defective six4 lead to failure in cell fate determination and affecting the maintenance and or survival of cells. Misexpression experiments in the mesoderm suggested that six4 contributes to patterning roles in coordination with Eya. Concerning eye development, most SIX factors have been associated with eye development. However, no role has been associated with six4 in the *Drosophila* eye system[56].

When talking about paralogs, we find that some functions have been conserved while others have become novel, with some controversy about the actual role between SIX4 and SIX5. What role, if any, SIX4 has is under scrutiny. Mice in which Six4 is knocked out don't demonstrate any visual defect, whereas Six5 mutants can develop cataracts or reduce the viability of the spermatogonia in mices[82]. However, some functional cooperation can be seen with other SIX members, like SIX1. CRISPR/CAS9 knockouts done to both Six1 and Six4 in pigs demonstrated that kidney disruption was severe when knocking out both[100]. Mice also deficient in both factors died after birth and exhibited defects in multiple organs; the same was observed in double knockout pigs[100]. In humans, some studies have demonstrated that SIX4 can be linked to different types of cancer and could be a marker for some types of them[101]. SIX4 appears to work in coordination with its peers by compensating other family members[82], forming protein complexes between them or other mechanisms that are still not fully understood.

SIX5 seems to have retained most of the roles associated with six4. Contrastingly Six5 is in the developing retina with later expression in the lens, with deficient mice demonstrating ocular cataracts[56,84]. This feature is interesting since no eye phenotype had been associated with SIX5 ancestral gene six4 in flies, showing a possible cooption into the regulatory networks in vertebrates. BOR syndrome has also been observed in SIX5 mutants with some mutations associated with the inability to bind with Eya1, like SIX1[84]. Myotonic dystrophy has also been associated with SIX5, patients with this disease show muscle wasting, defects in heart conduction[102], and fertility problems[82]. These diseases are not directly linked to mutations within SIX5 and are more based on changes in three-nucleotide expansion in the 3'UTR in chromosome 19, which can lead to the disruption of SIX5[82]. Little is known about mutations directly related to SIX5 and the possible consequences in the development of other diseases like cancer[56,82].

This subgroup highlights how evolution allows paralogs to gain new functions in evolutionary time. SIX5 still maintains the ancestral roles found in six4 while being responsible for new roles not previously linked to the subgroup. SIX4 roles are the most trivial but critical since it seems that it has developed a more cooperative role within the family by interacting with other members like SIX1 to carry on their roles. How these interaction works are still open to discussion.

Since their origins, the SIX family have used as mediators in numerous developmental features as phenotypes diverge. They have been coopted into these new pathways and even acquired new roles exclusive to some animals, like optix in *Heliconius*. What mechanism evolution uses to allow these factors to gain new functions is an interesting concept to consider since changes to their sequences can lead to the breakdown of the roles they accomplish. How gene regulatory networks recruit new TF to their pathways?

1.3. Evolution of gene regulatory networks

The regulation of gene expression between a TF and its targets is the central mechanism by which cells read their regulatory directives to develop and differentiate to a designated phenotype. As more complex phenotypes emerge from new species, it is believed that changes in regulatory elements are preferential since they wouldn't require large-scale rewiring of already established GRN. These changes can occur over a shorter period, while changes to protein-coding sequence consider to be evolutionary disfavor. In the case of TF, mutations within the protein can have a catastrophic effect that can lead to the collapse of the regulatory network. It should be then expected for the TF repertoire to be conserved; however, it isn't the case. For example, optix now has two paralogs that conduct multiple developmental roles; SIX4 has a compensatory role, while SIX5 is responsible for other functions. How can a TF change and mutate while simultaneously being responsible for carrying on the ancestral role? Do mutations alter the factor specificity, and could that explain functional changes? Evolution provides an opportunity for these changes to happen while also selecting the preferred roles for each factor. This mechanism is known as a gene duplication event.

Ohno (1970) describes a gene duplication event as the process of producing more of the same[103]. As mentioned in his work, for a duplication event to be successful, the vital function of the gene must be retained. This is because natural selection has assigned a role that prohibits mutations or changes. Mutations that occur in a TF can alter its binding mechanism, changes that can enhance or reduce affinity but without changing its core function. When duplication happens, one of the copies becomes redundant. Since it doesn't suffer from selective pressure, it allows for the accumulation of new mutations, which can alter the function of the new gene. For transcription factors, this could mean changes to their specificity or interactions with other regulatory elements.

During the process of gene duplication, the copies are under selective pressure to establish themselves or be deleted**.** There are multiple pathways for duplicates to be successfully established. The first is for the ancestral roles of the gene to be split between the new duplicates, in which each new copy will have different temporal and spatial moments so that no overlap will happen. If both copies retain the same genetic function with no spatial-temporal differences, this causes redundancy. Having redundant TF is usually unfavored by evolution, and eventually, one copy will establish itself while the remaining could be deleted. To prevent redundancy, one copy retains the ancestral function while the remaining gene can acquire mutations. During this process, this copy can then be neo-functionalized and attain a unique spectrum of roles, allowing the genome to use it in new processes[104]. All these processes start with the duplication event, where the genome ends with new genetic material that can change while preserving the organism's survivability but also introducing new components that can contribute to the already established GRN or be used to create new regulatory pathways.

The process of gene duplication can alter how TF can recognize DNA, with one copy that can retain the ancestral specificity, while the other TF can identify new sequences not capable before. This is observable in numerous examples that demonstrate how TF can evolve. For instance, it has been shown that within the HOX genes, the ancestral *Drosophila Labial* gene duplicated to give rise to both HOXA1 and HOXB1. Both paralogs are highly similar, with the main difference being the presence of a conserved six amino acid motif found in HOXA1, which allows it to retain the ancestral roles of labial and recognize similar genomic targets[105]. HOXB1, on the other hand, diverges in functions and now recognizes new genomic targets not previously identified for labial [105]. The mechanism by which a transcription factor can regulate expression can also be modulated by the protein-protein interactions they achieve.

The T-box transcription factor family has as founding member named Brachyury, which can be found in non-metazoan organism[106]. Work on Brachyury has demonstrated that this family has retained the ancestral binding specificity even in its paralogs. What is interesting is that it was shown that Brachyury could not interact with cofactors linked with other Tbox's[106]. These novel interactions must have originated during the evolutionary process of the family and function as mechanisms of selectivity and regulation used by the family to carry on with their roles. The possibility of recognizing secondary motifs can also be how a TF can undergo new biological functions.

When studying the Tbrain (Tbr) TFs, it was discovered the presence of a secondary binding motif. The work done by Cheatle et al. (2014) when studying three orthologs of Tbr found that when studying sea stars, sea urchins, and mice, only the sea star recognizes a secondary motif [33]. This secondary motif was demonstrated to correspond with changes in Tbr expression during development. Another example can be seen within the forkhead TF family. It has been shown that this family can recognize a broad spectrum of motifs besides the canonically recognized motifs. In some cases, some members can even recognize three binding motifs[31]. This capacity seems to develop independently between members since closely related factors cannot recognize more than one binding motif. It was also shown within the same family that a bispecific member could recognize two different motifs. Surprisingly the way the factor recognized either one depends on the DNA shape[32]. Numerous articles have also demonstrated other examples found in other transcription factor families[15,18,107–111].

Gene duplication is known to be an essential mechanism for TFs, to change and adapt to any role the genome has need for it. We have seen how different families of TFs have changed and, in the process, can achieve new functions, either through the new recruitment mechanism or by the changes in specificity. When taking into consideration the SIX family, the diversity in functions observed may be based on differences between members. The SIX family has gone through three duplication events that have led to 6 new copies, *sine oculis* (*SIX1* and *SIX2*), *optix* (*SIX3* and *SIX6*), and *six4* (*SIX4* and *SIX5*), and even an independent one that leads to *Six7*. It should not be surprising that modifications in their DNA-binding specificity could explain their functioning. One can propose that this diversity in function is given by how the SIX TF interact with DNA. How this family of transcription factors evolved can show how similar or dissimilar each member is. These differences can be translated into changes in binding specificity, which can explain how this family is involved in many developmental processes.

1.4. Thesis aims

Aim 1. Determine the evolutionary history of the SIX Transcription Factors

The SIX transcription factor family evolution is an ongoing debate with multiple phylogenetics analysis demonstrating contrasting topologies. Much of the work done has been realized using only the proteins domains found in the SIX TF (Six domain and/or homeodomain) which leads to the different results based on the configuration of the data. To determine a more precise phylogenetic analysis I will used the most recent available tools for phylogenetic inference using the full-length sequence of SIX proteins. The sequences used will come from various Metazoan genus to be able to comprehend the most about the history of the SIX family.

Aim 2. *In vitro* determination of the intrinsic DNA-binding preference of the SIX transcription factor

The specificity of a transcription factor is a conserved feature that allows for the preservation of numerous gene regulatory networks. However, genomic duplications allow for changes in specificity to the produce copies. It has also been observed that orthologs can have exclusive specificities based on the species-specific modifications that results in divergent specificities. By using Systematic Evolution of Ligands by Exponential Enrichment followed by massively parallel DNA sequencing (SELEX-seq), I will determine the DNA-binding properties of SIX orthologs and paralogs. I will compare the DNA-binding preferences of the SIX transcription factors found in *Drosophila melanogaster*, *Heliconius erato* and *Homo sapiens*.

Aim 3. DNA-binding properties of *Heliconius erato* optix

The *Heliconius sp.* butterflies use an array of different colors to discourage predators. The red color pigmentation found on the wings is regulated by the transcription factor known as optix. The optix transcription factor is a known regulator of eye development in fruit flies being coopted into the Heliconius wings. Little is known about the DNA-binding preferences of optix. By using Systematic Evolution of Ligands by Exponential Enrichment followed by massively parallel DNA sequencing (SELEX-seq), I will determine the DNA-binding properties of *Heliconius erato* optix. The results of SELEX-seq will allow me to predict and validate optix binding in previously described cis-regulatory elements specific for optix.

Chapter 2: The Evolutionary History of the SIX Family of Transcription Factors.

2.1.Introduction

The instructions by which genes are regulated are stored within each species' genome, which dictates what is necessary for an organism to succeed. Extant genomes are the result of billions of years of evolution. Protein-coding genes are found within each species' DNA and are also the result of evolution and have changed since the first protein-coding gene originated. Understanding how these genes have diverged is an essential question in biology. Gene regulatory networks (GRN) are the regulatory mechanism between transcription factors (TF) and the genes they orchestrally regulate to produce a specific phenotype. Little is known about the evolution of these pathways and their components[112]. For example, natural selection usually coopts genes into new functions and, in the process, develops new features[113]. Gene duplication is the mechanism by which new genes are made, on which the paralogs that originate develop distinct functions that allow for diverse roles[103,105,112–114]. Transcription factors are critical components in gene regulation, and gene duplication events can provide pathways that enable them to function in new roles[18]. Understanding how TFs have evolved can provide information on how they have been used for phenotypic innovation that led to the establishment of life on Earth.

The SIX are members of the second most common class of transcription factors, the homeodomains (HD), in which they are considered atypical[34]. This is based on crucial amino acids responsible for the affinity and specificity to bind and read DNA. Within the SIX coding sequence, there is a unique protein domain named the SIX domain (SD), which is known to participate in protein-protein interactions[55,56,66,78]. The SIX family has three members in *Drosophila*: sine oculis, optix, and six4. They are involved in numerous developmental processes, from eye development to the red color pigmentation in the *Heliconius* butterflies' wings, to kidney or brain development, and much more[56,65,67,75,76,79,100]. Like many other protein-coding genes, this family has undergone a gene duplication event that has led to six paralogs (SIX1-6)[56].

Being critical components in multiple developmental processes, it is essential to understand this TF family's evolutionary story. Multiple phylogenetic analyses have been done on the SIX family, in which numerous topologies have been proposed[54,68,72,75–77,115–117]. Since their discovery, these analyses have been done as more sequences are obtained. However, much of the research has been conducted using only the protein domains (SD and HD). Both domains have been used in different configurations that have provided conflicting topologies. For example, most phylogenies propose that optix is evolutionarily related to sine oculis, with six4 being the more evolutionary divergent. While others have suggested the opposite topology, in which optix is more evolutionary related to six4. These differences are observed depending on the information used for the analysis, the usage of the protein domains, and the species taken into consideration. Using the sequence of protein domains is practical; however, with limitations. The full-length sequences carry essential information about how a protein is regulated and possibly provide information to understand its evolutionary history. For example, some reports have demonstrated that non-domain regions in the SIX proteins are crucial in determining the specificity of some SIX proteins and the genes they regulate[62,63]. The full-length aminoacidic sequence of the SIX proteins has not been consistently used to determine their phylogeny, which can provide a more precise resolution to this family's history and origins.

This work aims to reconstruct the evolutionary history of the SIX family. By using the whole aminoacidic sequence of SIX proteins to fill the current gaps in the family. While also to pinpoint the moment in which the genes were established.

2.2.Materials and methods

2.2.1. Sequence Data

We searched and retrieved SIX proteins sequences from sponges to humans. Some sequences were obtained from the Ensembl database, using the paralogs feature to identify protein paralogs[118], others were obtained from the Uniprot database[119]. Candidates were also obtained from TBLAST searches using default settings[120,121], using as input the SIX protein sequences found in *Drosophila* (UniprotID: Q27350). Some gene fragments were found within these databases. We used fragments with at least two protein domains (SD and HD) to determine if the sequence would be considered. The longest sequences were selected if multiple isoforms were found for a protein. These decisions responded to our objective to attain a broader coverage of sequences. A total of 86 sequences of SIX-related proteins were used for downstream analysis.

2.2.2. Phylogenetic analysis

Protein sequences were visualized using MEGA[122]. For the multiple sequence alignment, we used the MUSCLE alignment algorithm found within MEGA[123]. The sequence alignments were then exported for phylogenetic inference. Maximum Likelihood (ML) analyses were run using IQtree 1.6[124] using the web implementation from the IQtree web server[125] (last accessed in September 2021). To determine the best-fitting model, we used the ModelFinder[126] feature in IQTree. The JTT+R10+F was determined to be the best-fit model. Support for the nodes was evaluated using the UFboot2[127] feature from IQtree, with 1,000 pseudoreplicates. The phylogenetic inferences obtained were then visualized using the FigTree V1.4.4 software (http://tree.bio.ed.ac.uk/software/figtree/). Tree selection was based on the support of the nodes and observable tree topology (**Supplementary image 2)**. Selected tree topologies were tested using already implemented methods in IQtree. We used the Dendroscope software[128], to view large trees and export them for better imagining. The exported tree was then edited using Inkscape (https://inkscape.org/) for a publication-quality phylogenetic tree.

2.3. Results

The presence of the SIX proteins were able to be traced from sponges to humans (**Figure 2.1)**. This places at least a SIX protein since early metazoan, with at least one SIX protein found in the sponge *Amphimedon queensladinca,* while two more SIX-like proteins were found in other sponge species[66]. The three canonical SIX proteins (Sine oculis, optix, and Six4) were observed in *Ctenophore* and *Cnidaria* (**Figure 2.1A**). These proteins can be traced to around 810MYA, predating Bilateria. The SIX family composition remains constant in other animals, with the majority having the three canonical proteins. However, some divergence is observable in some phyla. The nematode *Caenorhabditis elegans* (*C. elegans*) has 4 SIX-like proteins, while in the *Trichoplax sp*., only two SIX-like proteins are found.

In the echinoderm *S. purpuratus,* the three canonical SIX proteins are found. This group is the most related phylum to Chordates. The SIX are conserved within the Chordate phylum, and it is within them that the genomic duplication of the SIX happened (**Figure 2.1B**). In lampreys, there are multiple copies of SIX proteins, with at least five of the SIX paralogs. Fishes have multiple copies of some SIX proteins while having only one copy of the Six4 protein (*Elephant shark*). This is also observable in *Xenopus*. The six paralogs of the SIX family are entirely found in both the *Spotted Gar* and the *Coelacanths*, with an extra SIX protein found in the *Reed fishes*. In mammals, all the paralogs are observable from SIX1 to SIX6. Within these sequences, it is noticeable that the SIX protein family has diverged in multiple aspects since forming the three canonical proteins to the presence of the SIX paralogs.

Sequence alignment was done to 86 SIX proteins (**Supplementary Image 1**). Both protein domains (SD and HD) have been highly conserved since *Porifera*. Protein-specific differences within the HD differentiate the proteins. The sequence of the N-terminal arm hexapeptide [34–36] is specific to each canonical SIX protein. A trait conserved in *Cnidaria* and *Ctenophora*, where sine oculis is characterized by the GEETSY motif, optix (GETQKH), and six4 (GEETVY). Some protein sequences show a divergent hexapeptide motif (GEETNY). Both domains are highly conserved; however, the SD is the more variable. The SD domain conservation is found on the six alpha helix and the unorganized linker region between the SD and HD. The SD domain also has a characteristic four amino acid insertion that is only observable in optix/SIX3/6 proteins. This insertion is observable since Cnidaria and Ctenophora and is not lost in subsequent orthologs and paralogs. In the case of Six7 proteins, the two sequences in our analysis have an insertion of 6 amino acids long. This insertion is not observable in sponges or other SIX protein members and is solely found in optix members.

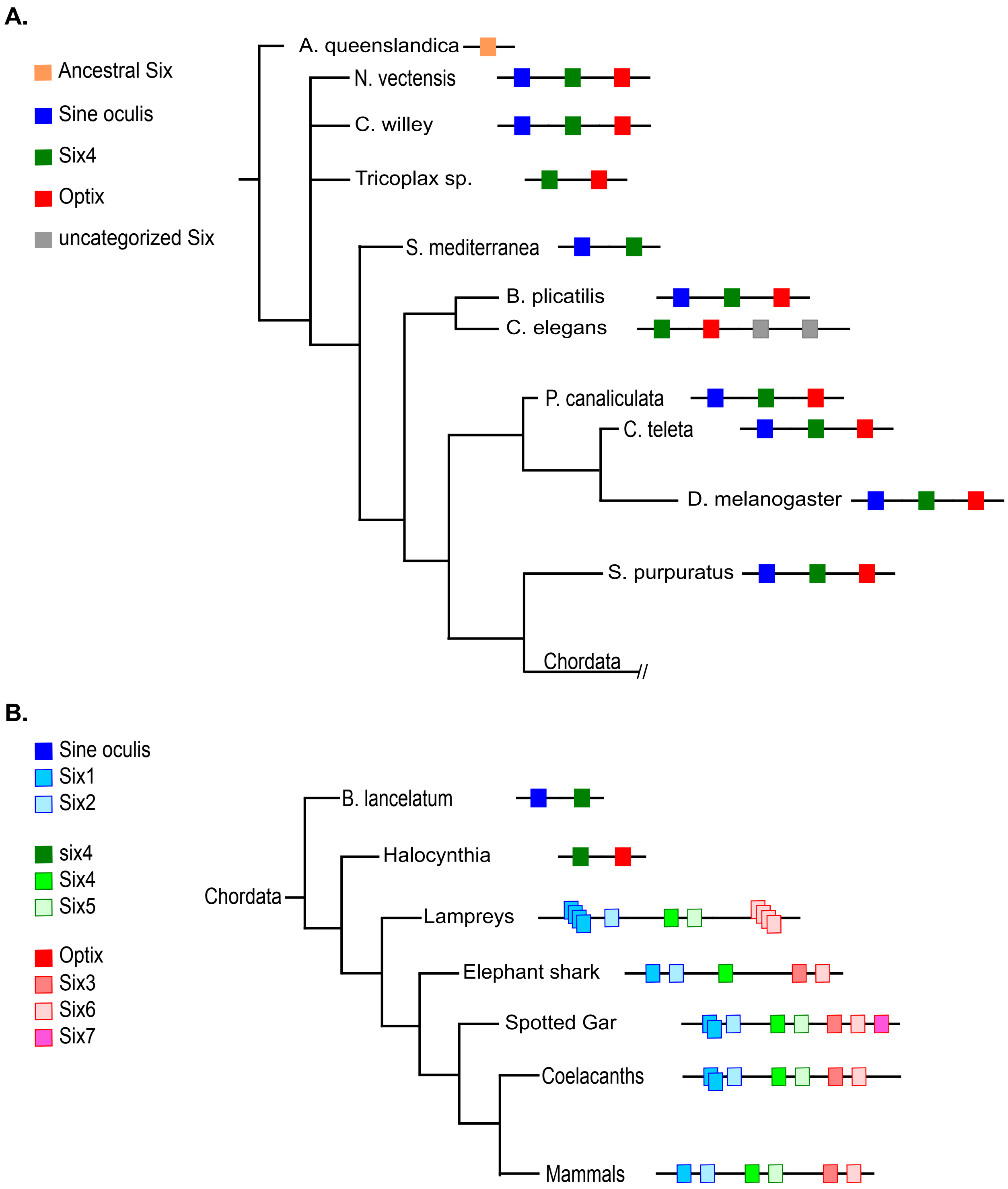


Figure 2.1. The SIX proteins are found from sponges to humans. Fig. 2.1A. At least an ancestral SIX protein is found in *A.queenslandica*. The three SIX proteins (Sine oculis, Six4, and optix) are observable from Radiata (*N. vectensis* & *C. willey*) and Bilateria, with some protein loss and gain in two species (*Tricoplax, C. elegans*). Fig. 2.1B. The gene duplication event occurred between the Tunicata (*Halocynthia*) and Vertebrates while an independent occurred in Reedfishes (Spotted gar).

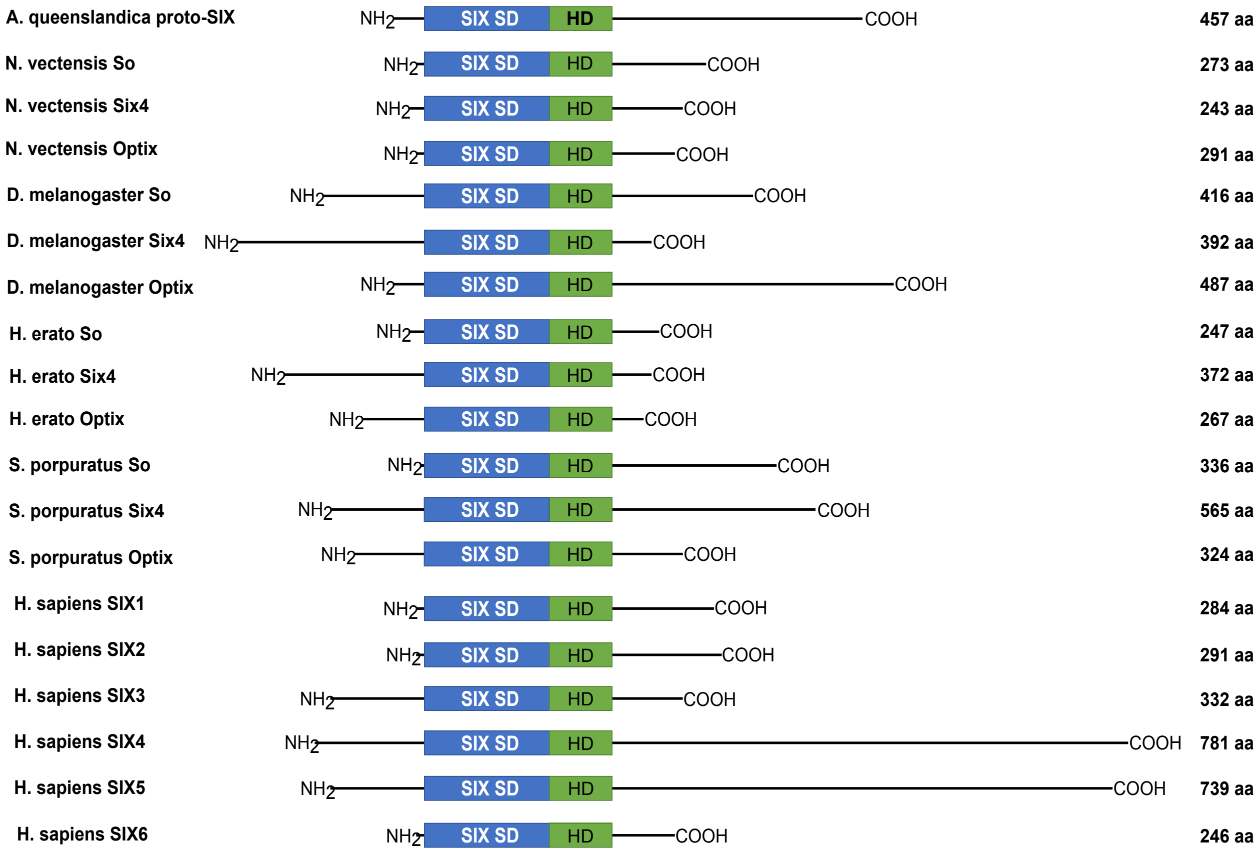


Figure 2.2. SIX protein domains (SD and HD) are the most conserved regions. High divergence is observed surrounding the protein domains, with adjacent regions variable.

The SD has low divergence in two significant areas, the first 30 amino acids and then from positions 80 to 120 of the SD. Some residues seem group-specific between the sponge SIX and sine oculis members while sharing the same N-terminal arm hexapeptide. Some amino acid residues are group specific in the SD, and others are conserved between the sponge SIX and sine oculis. An example is residue (G103); this residue is different in optix and six4 with a (D103) residue-specific between them. The opposite can be said at position 107; both optix and sine oculis have a (V107), while six4 has I and L. The isoleucine is conserved from sponges and jellyfish with subsequent evolutionary selection for the leucine.

These DNA-binding positions, V102, R106, and R108, are conserved since sponges as no divergence was observed. The homeodomain (HD) remains highly conserved, with the recognition helix being preserved since sponges with no noticeable modifications. The protein domains haven't changed in size, with the SD being from 114 to 119 amino acids depending on the SIX members; the latter is based on the optix/SIX3/6/7 insertion. The HD is canonically defined by 60 amino acids, and the SIX follows that definition and structure. The unstructured regions found outside each domain are highly variable, with no clear consensus on their similarity. The amino and carboxylic regions found outside are highly variable. **Figure 2.2** shows the size distribution of some SIX proteins. The variability in size found both upstream and downstream of each domain is divergent, even between orthologs. For example, *Drosophila's* optix is the bigger protein of the set, while *Heliconius* is the opposite.

The phylogenetic inference determined for the SIX protein family shows the presence of four subgroups (**Figure 2.3**). These subgroups are composed of the three canonical SIX proteins (sine oculis, optix, and six4), while the remaining subgroup are proteins that are uncategorized to any of the three canonical subgroups. The uncategorized is composed of SIX proteins found in *Porifera* and two SIX proteins from *C. elegans*. They are labeled as uncategorized since they are not within the principal subgroups. The remaining SIX proteins in *C. elegans* are within the optix and Six4 subgroups. The last common ancestor between the SIX proteins originated from this uncategorized subgroup. This common ancestor gave origin to the sine oculis protein, while another divergence step must have occurred for the origin of optix and six4. From sine oculis, the genomic duplication event led to the formation of its paralogs Six1 and Six2. Since the subgroup is composed of all sine oculis members, we wanted to study the HD hexapeptide's conservation based on its characteristic for each protein subgroup (**Figure 2.3 and Supplementary Image 3**). Based on the sequences of our analysis, it was found that within the sine oculis/SIX1/2 subgroup, there is a high grade of conservation of the hexapeptide with slight divergence observed. The hexapeptide GEETSY is also observed in SIX proteins in sponges, but the candidates are part of their uncategorizable group.

The optix and six4 proteins are identified in Ctenophora and Cnidaria, and this suggests both proteins are more evolutionary-related between them than Sine oculis. The optix clade has all the orthologs and paralogs of optix with the unique reed fish duplication that led to the Six7 protein that seems to have originated from a Six3 common ancestor. The proteins identified from *Ctenophora* and *Cnidaria* are all arranged within the optix subgroup, establishing that optix was present in these phyla. The hexapeptide identified within the optix subgroup has a consensus sequence of GEQKTH, with some divergence observed in residues 2 to 5. The last subgroup is composed of six4 proteins which can be traced to the comb jellies and jellyfish as optix. The *Coeloplana* six4 protein branches outside the main Six4 subgroup, but this may result from using a fragment of the candidate protein.

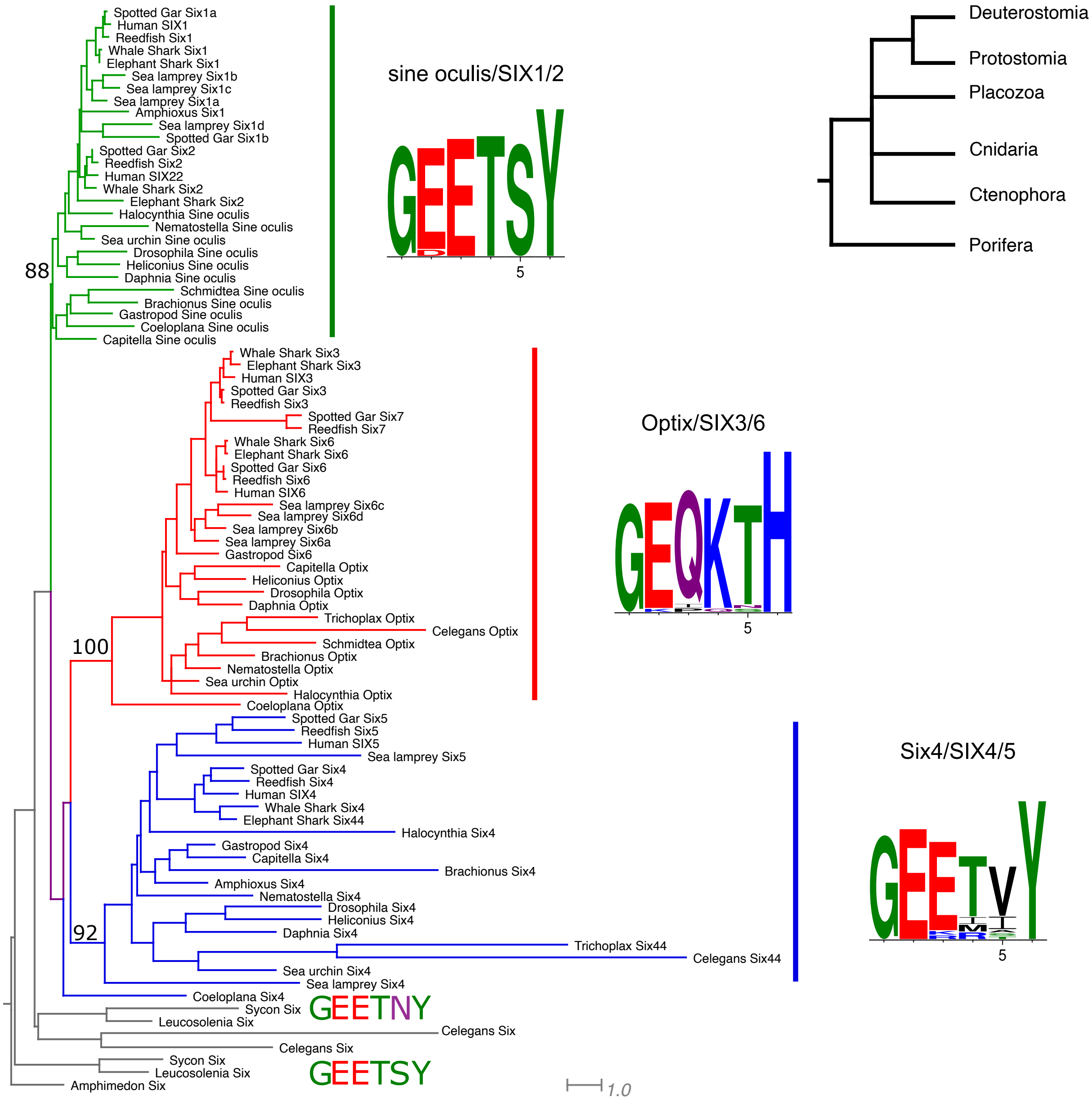


Figure 2.3. Maximum Likelihood phylogram describing the evolution of the SIX proteins. Numbers on the nodes correspond to maximum likelihood bootstrap support values from IQtree. The SIX family is distributed into three main subgroups with an uncategorizable ungroup zone. Both optix/SIX3/6 and Six4/SIX4/5 share a common ancestor that originates from the last common ancestor with the Sine oculis/SIX1/2 subgroup. Hexapeptide sequence is conserved within each subgroup with some low divergence observable.

2.4. Discussion

The SIX are an essential TF family with roles established since at least the simplest of animals, for example sea sponges. Finding this protein family in Porifera traces the SIX origins to hundreds of millions of years ago. The search for SIX proteins in sponges shows that at least in *A. queenslandica,* there is only one SIX protein, with other sponges demonstrating at least 2 SIX-like proteins. The divergence in the number of protein copies is exciting and suggests that the SIX proteins were constantly changing within sponges without proper establishment. This is observable when considering that all SIX-like proteins in our phylogenetic analysis are in the uncategorized subgroup (**Figure 3**), not fitting within the canonical SIX proteins. The proteins observed in sponges can be considered proto-SIX proteins and are the ancestors of all the current SIX genes. Interestingly the proto-SIX found in *A. queenslandica* shares the N-terminal arm hexapeptide characteristic of sine oculis (GEETSY). The number of SIX proteins in sponges during our analysis was restricted to only three sponge species, limiting the number of copies considered. However, there is a need for sequencing sponge genomes that could provide resolution to understand the protein dynamics in Porifera. An increase in available genomes could give a better understanding of the origins of the SIX proteins. Even with this limitation, we successfully observed the presence of these proto-SIX proteins that have been reported to be responsible for sensory processes in sponges, particularly in sensing light[65], a trait shared with almost all its descendants.

The three canonical SIX proteins are identified in jellyfish and comb jellies. This places all three proteins predate the origin of Bilateria. This establishes that the three SIX proteins have been present in Metazoan for almost 810MYA. However, in *C. elegans*, we could not identify a sine oculis protein, even when one candidate does carry the expected hexapeptide. Some reports have demonstrated that this protein functions during head morphogenesis [72]. Still, two SIX-like proteins are found within the uncategorized. This could mean that these two proteins are highly divergent, and their placement reflects this. Surprisingly when considering the hexapeptide sequence, it is observed that some highly divergent proteins share the same hexapeptide of sine oculis. But even though they share this characteristic, they are highly divergent to be considered members of sine oculis. In contrast, the remaining uncategorized SIX proteins have a hexapeptide of GEETNY. This was also observable in some proto-SIX proteins in sponges, possibly because of a transitionary phase with these proteins. These observations also demonstrate that even when some proteins can share the characteristic hexapeptide on the HD, this alone is not enough to differentiate each protein class. It shouldn't be surprising that these proteins differences are based on the regions outside the domains.

Why do some species have missing SIX proteins? This is an exciting question that could result from the current genomic assemblies available that have been unable to identify all the SIX proteins in some species. However, it is also possible that they may have been deleted or lost based on an independent evolutionary process. For example, in humans, there are both SIX4 and SIX5, with spotty representation in some species. Surprisingly, some reports on humans [82] have shown that the deletion of SIX4 has no apparent effect on the phenotype; the same cannot be stated for SIX5. It is plausible that some redundancy has occurred during the evolutionary process of both proteins and some species seem to have lost one copy of either SIX4 or SIX5.

Sequence conservation is observable within protein domains but high divergence is observable in regions outside the domains. The HD is the most conserved region, with the recognition helix being conserved since sponges. The SD has some divergence while retaining most of its conservation within its final alpha helix and its linker to the HD. Some changes are species-specific, but at least key binding residues have remained constant since Porifera.

optix/SIX3/SIX6/SIX7 SD is different from the other SIX proteins; the SD domain has a unique insertion that is highly variable with no consensus sequence. This insertion was also observable in SIX7, with the main difference is having a more extended insertion.

The evolutionary history of the SIX hasn't been completely established, with multiple tree topologies being reported [54,68,72,75–77,115–117]. The phylogenetic inferences have been made using, in the majority, only the protein domains to establish the phylogeny of the SIX family. This has resulted in the different placement of both optix/SIX3/6/7 and Six4/SIX4/5, where it is determined that optix/SIX3/6/7 is the distinct subgroup with six4/SIX4/5 and sine oculis/SIX1/2 sharing a common ancestor. We believe these results are the result of only using the protein domains. We can provide a complete phylogenetic analysis using the whole protein sequence of SIX proteins found in numerous phyla. Here we demonstrated that the optix/SIX3/6/7 and six4/SIX4/5 related proteins are closer than sine oculis/SIX1/2. During the evolutionary process of each protein, sine oculis diverged independently, while optix and six4 share a common ancestor. This highly contrasts with previous phylogenies done to the family. Our work used all the protein sequences for each member and not the domains, which provided a greater resolution for the placement of each subgroup. The sine oculis/SIX1/2 subgroup seems to be the less divergent SIX, than optix/SIX3/6/7 and six4/SIX4/5.

Since early Metazoans, the optix/SIX3/6/7 subgroup has been present, with expression reported in jellyfish, especially within the developing eye cup[75]. In addition, a *C. elegans* SIX candidate protein can be placed within this subgroup. Within the paralogs of optix, one can find Six7, which originates from an independent duplication event within the Six3 protein. This makes Six7 more evolutionary related to Six3 than to Six6. It was also interesting to notice differences between optix members from Arthropoda. For example, *Drosophila's* optix doesn't group with its more relatable species in insects with *Heliconius* optix, being more like the optix found in the Brachiopoda crustacean *Daphnia magna*. This is interesting but not completely surprising since previous reports mention that Drosophila's optix is different from other Arthropods, with an exclusive domain within the C-terminal after the HD [63]. These types of modifications highlight the differences that are possible within orthologs.

In the case of Six4/SIX4/5, this is the most divergent group of all the canonical SIX proteins. This subgroup carries the remaining *C. elegans* gene, which has been reported to play essential roles in motility and differentiation[74]. *Ctenophore* Six4 protein was located near but not within the Six4/SIX4/5 subgroup. This is based on the sequence used in our analysis and its fragment status. However, we performed topologies tests to confirm that this gene was a six4. Our analysis show no changes in the topologies observed, which means that the placement of this gene is within the Six4/SIX4/5 subgroup. Enhanced sequencing could fill in the gaps and officially place the gene within its subgroup.

2.5. Conclusion

This work determined the evolutionary history of the SIX transcription factor family. This family has been conserved since the early presence of Metazoan, placing them in Porifera. However, some sponges demonstrated the presence of more SIX proteins that were possibly lost during evolution. The need for new factors led to the formation of the SIX protein that we currently recognize, and further evolutionary time led to their duplication. The current analysis demonstrated that phylogenies using the whole-length protein sequence provide a more precise understanding of protein evolution. Doing so allows us to provide a more detailed evolutionary history and predict the controversial placements reported. The story of the SIX family is diverse and complex and filled with fascinating examples of how genes evolve. Their conservation highlights their vital roles in development and their critical place in the numerous developmental processes. What makes them different in a functional context is an exciting endeavor to provide insight into their functional evolution.

Chapter 3: DNA binding Specificity of the SIX Transcription Factors.

3.2 Introduction

Understanding the dynamics of DNA-binding specificity by transcription factors (TF) provides insight and valuable information into the mechanism of gene regulation. The DNA binding domain (DBD) found within TFs is responsible for interacting directly with the DNA and contributes significantly to their specificity [4,5,7,30]. The homeodomains (HD) are the second most abundant class of DBDs [34]. The HDs have a conserved binding motif associated with the class, with divergent members having different specificities[34]. These alterations can be correlated to changes in key amino acid residues that are important for DNA interaction[34,35]. There are multiple pathways in which a TF can change its specificity. One possibility is species-specific modifications that regulate specific developmental roles in different organisms[33,105]. Another possibility is an evolutionary event like gene duplication that can lead to changes in specificity[103]. In this process, one copy retains the ancestral role and specificity, while the other can diversify and, in the process, alter its specificity[104].

Changes in specificity can originate from changes in the amino acid residues involved in the protein-DNA interaction. DBDs are classified according to the homology shared between members[34,129,130], and protein sequence conservation leads to a conserved binding motif between members. The HDs are described to share a conserved DNA binding motif of `TAATTA`. This motif has been found in HD families like the HOX and Antennapedia. However, some families can demonstrate divergent binding motifs and are usually recognized as atypical[34].

The SIX are an atypical family member of the HDs. The SIX HD has a highly negative N-terminal arm that contrasts with the highly positive nature of typical HDs[34]. They also have an exclusive protein domain known as the SIX (SD) domain found N-terminal of the HD[43]. The SIX family specificity is characterized by a `TGATAC`, different from the canonical `TAATTA`[26,34,56]. The divergence in specificity can be the result of numerous factors, from the effect of amino acid modifications within the HD or the presence of the SD.

The SIX are involved in multiple developmental processes. The family has been found from sponges to humans[56,65,75]. The SIX are recognized by the three canonical members (sine oculis, optix, and six4)[56]. In addition, a duplication event led to at least six new members (SIX1 to SIX6) in animals, from Lampreys to Humans, all originating from the canonical SIX members. Their roles are diverse, which highlights their rich evolutionary history [42,54,75–77,90,115]. Roles that span from the eye development in flies to the red wing color pigmentation in the *Heliconius* butterflies and the lateralization of the human brain. This family contributes to multiple developmental pathways which have become critical for the survivability of an organism[66,69].

How the SIX can achieve their roles isn't fully understood, especially when considering changes in their specificity. Little is known about the specificity of all SIX members and if binding differences can differentiate between members. Less is known about the evolutionary changes between orthologs (sine oculis, six4 and optix) and their paralogs SIX1, SIX2, SIX4, SIX5, SIX3 and SIX6. This work will by evaluating changes in specificity between the SIX members, taking into consideration the binding specificity between orthologs and paralogs. This analysis can demonstrate if changes in specificity have occurred in the SIX family while also validating if genomic duplication events lead to binding specificity divergence.

3.2 Materials and methods

3.2.1. SIX genes

To analyze the specificity between orthologs and paralogs of SIX transcription factors, we decide to use the SIX genes found in *Drosophila melanogaster* (Dmel), *Heliconius erato (*Heli*),* and *Homo sapiens* (Homo). The sequences from *Drosophila melanogaster* were obtained from plasmid repositories. The *sine oculis* and *six4* clones were obtained from the Drosophila Genomics Resource Center. The clone for *sine oculis* was GM13131 (DGRC Stock 1280637; https://dgrc.bio.indiana.edu//stock/1280637; RRID: DGRC_1280637), while *six4* clone is FI01103 (DGRC Stock 1623344; https://dgrc.bio.indiana.edu//stock/1623344; RRID: DGRC_1623344). The *optix* gene was obtained from the DNASU plasmid repository (DmCD00766251).

The *Heliconius* SIX genes were synthesized using the Gene Wiz Standard Gene Synthesis service (https://www.genewiz.com/). The sequences were obtained from multiple databases. The *sine oculis* and *six4* were obtained using the Lepidopteran genome database (http://lepbase.org/). The protein sequence of sine oculis and six4 from *Drosophila melanogaster were* used as queries. The optix sequence (ID: L7X1S6) was obtained from the Universal Protein Resource (Uniprot https://www.uniprot.org/).

The Human SIX genes for *SIX2*, *SIX3*, *SIX4*, and *SIX6* were obtained from the DNASU plasmid repository. Each clone can be identified by the following clone IDs, HsCD00618350 (*SIX2*), HsCD00297116 (*SIX3*), HsCD00733174 (*SIX4*), and HsCD00079817 (*SIX6*). The *SIX1* gene was obtained from the Addgene plasmid repository (MSCV-Six1 was a gift from Heide Ford (Addgene plasmid # 49263; http://n2t.net/addgene:49263; RRID: Addgene_49263)). The SIX5 sequence (ID: Q8N196) was retrieved from the UNIPROT database. SIX5 was synthesized using Genewiz as used for the *Heliconius* SIX genes.

3.2.2. Cloning of SIX genes

The SIX genes in our catalog were cloned in two types of expression vectors, one for a bacterial expression system and the second for the expression of proteins in a cell-free system. The vector utilized for bacterial expression was the pET32a (+) vector from Millipore Sigma (Cat. 69015). This plasmid has a Thioredoxin solubility tag, a S•tag, and a 6xHis-tag for protein purification. The cell-free protein expression system was from CellFree Science (https://www.cfsciences.com/eg/). The manufacturer recommends the expression vector pEU, which is optimized for the system. The plasmid selected was the pEU-E01-His-TEV-MCS-N1.

To insert our genes into these expression vectors, we used Gibson Assembly[131]. Gibson cloning requires the linearized plasmid vector and for each gene insert to have a 15bp flanking region on both ends (5' and 3'), which must overlap with the vector. All primers used can be found in **Table 3.1 and Table 3.2**. The DNA primers were purchased from Integrated DNA Technologies (IDT, Coralville, Iowa, USA) as standard desalted. The linearized vector and all the SIX inserts were produced using a high-fidelity polymerase from New England BioLabs (NEB M0530L). We used the Phusion High-Fidelity (HF) DNA Polymerase master mix. It comes with a High-Fidelity buffer (Cat. M0531) and a GC buffer (Cat. M0532), used for PCR reactions with templates rich in GC content. To determine the annealing temperature for all reactions, we used the NEB Tm Calculator (https://tmcalculator.neb.com/#!/main).

All reactions, minus *SIX5* and Dmel *optix*, were done using the HF buffer with successful results. The inserts from both *optix* and *SIX5* were produced using the GC buffer. To isolate and clean our inserts, we used the QIAGEN QIAquick PCR Purification Kit (Cat. No. / ID:28106). All

SIX inserts were isolated using this kit except for *SIX5*. To successfully isolate the *SIX5* insert, the PCR product was separated by electrophoresis using a 1% Agarose gel. The insert was visualized using Ethidium bromide and a UV transilluminator. The DNA band corresponding to *SIX5* was cut from agarose and purified using the QIAGEN QIAquix Gel Extraction Kit (Cat. No. / ID: 28707.) The pET32a (+) and the pEU-E01-His-TEV-MCS-N1 vector were linearized and purified using the same procedure.

To develop our plasmid constructs, we used the Gibson Assembly Cloning Kit from NEB (Cat. E5510). The reaction was done as the manufacturer suggested using the NEB Ligation Calculator (https://nebiocalculator.neb.com/#!/ligation). The vector mass was set to 30ng, and the inserts were added in a 5:1 ratio, as calculated in the NEB Calculator. The products of these reactions were then used to transform competent *E. coli* bacteria. We used the NEB 5 alpha Competent *E. coli* (Cat. C2987H), and 5 µL of each ligation reaction was used in each transformation process using the High-Efficiency Transformation Protocol suggested by the manufacturer. Luria-Bertani (LB) plates were used to validate the transformation procedure; these had ampicillin as the selective agent. Plates were left at 37^{o}C to incubate overnight, and colony growth was evaluated the next day. Bacterial colonies were isolated from these plates and used as templates for Colony PCR. The procedure confirms that the gene of interest is found within the vector. We used NEB Taq 2X Polymerase Master Mix (M0270L) and specific primers for each vector. In the case of pET32, we used the standard T7 promoter and terminator primer sequences. For inserts within the pEU-E01-His-TEV-MCS-N1, we used custom-made primers based on the SP6 promoter and a reverse primer named MCS-Rv (**Table 3.3**) to cover all the insert regions.

Primer annealing temperatures were determined using the NEB Tm Calculator, and reactions were done according to manufacturer protocol. PCR reaction mixes were run by electrophoresis in an agarose 1.5% agarose gel, and the presence of the insert was evaluated based on the expected weight for each amplicon. Bacterial colonies confirmed to have the expected SIX factor were left overnight for no more than 16 hours. The plasmids were then purified using the Qiagen QIAprep Spin Miniprep Kit (Cat. No. / ID: 27106). Purified plasmids were then sent for Sanger Sequencing utilizing a selection of standard primers (T7 promoter and terminator) and custom-made primers for some genes of interest (**Table 3.3**). Some primers didn't entirely cover some gene sequences. Hence, we made specific primers to cover all the coding sequence, allowing us to evaluate the complete gene sequence. All primers used for sequencing can be found in **Table 3.3**.

Table 3.1: Primers for gibson cloning using the pET32 vector

Primer Name	Primer Sequence
Fw_Linear_pET32(+):	TAACAAAGCCCGAAAGGAAGCTGAG
Rv_Linear_pET32(+):	CTTGTCGTCGTCGTCGGTACCCAGA
SO_Dmel_pET32_Fw	GACGACGACGACAAGATGTTACAGCATCCC
SO_Dmel_pET32_Rv	GTGGTGGTGGTGGTGTAAGTGCTGGTACTCGCC
SIX4_Dmel_pET32_Fw	GACGACGACGACAAGATGTTTGACAAGAAT
SIX4_Dmel_pET32_Rv	GTGGTGGTGGTGGTGTTGCCCATTGAAAAT
optix_Dmel_pET32_Fw	GACGACGACGACAAGATGGCCGTTGGACCG
optix_Dmel_pET32_Rv	GTGGTGGTGGTGGTGTGTGATCTCGGGAGCACTG
So_Herato_pET32_Fw	GACGACGACGACAAGATGCTGGGTGGTCCG
So_Herato_pET32_Rv	GTGGTGGTGGTGGTGGGTATGATGGTATTGCAGATGC
optix_Herato_pET32_Fw	GACGACGACGACAAGATGCGCGGTAGCTGG
optix_Herato_pET32_Rv	GTGGTGGTGGTGGTGCTCTTCGTCAACGTT
SIX4_Hmel_pET32_Fw	GACGACGACGACAAGATGGAAAGTTGCAGTGATCGT
SIX4_Hmel_pPET32_Rv	GTGGTGGTGGTGGTGCGGCGGATTACCATGCA
SIX1_Huma_pET32_Fw	GACGACGACGACAAGATGTCGATGCTGCCG
SIX1_Huma_pET32_Rv	GTGGTGGTGGTGGTGGGACCCCAAGTCCACCAGA
SIX2_Huma_pET32_Fw	GACGACGACGACAAGATGTCCATGCTGCCCACC
SIX2_Huma_pET32_Rv	GTGGTGGTGGTGGTGGGAGCCCAGGTCCACGA
SIX3_Huma_pET32_Fw	GACGACGACGACAAGATGGTATTCCGCTCCCC
SIX3_Huma_pET32_Rv	GTGGTGGTGGTGGTGTACATCACATTCCGAGTCGC
SIX4_Huma_pET32_Fw	GACGACGACGACAAGATGGAAAGCGCCTCG
SIX4_Huma_pET32_Rv	GTGGTGGTGGTGGTGTAAGTCTTGCATATCTTCATCC
SIX5_Huma_pET32_Fw	GACGACGACGACAAGATGGCGACGCTGCCG
SIX5_Huma_pET32_Rv	GTGGTGGTGGTGGTGCAGCTCCAGTGGTTCCTCA
SIX6_Huma_pET32_Fw	GACGACGACGACAAGATGTTCCAGCTGCCCA
SIX6_Huma_pET32_Rv	GTGGTGGTGGTGGTGGATGTCGCACTCGCT

Table 3.2: Primers for gibson cloning in the pEU-E01-His-TEV-MCS-N1 vector

Primer Name	Primer Sequence
Fw_Linear_pEU-E01-His-TEV-MCS-N1	GATATCTCGAGGATCCCGGGTAC
Rv_Linear_pEU-E01-His-TEV-MCS-N1	GCCCTGAAAATACAGGTTTTCG
SO_Dmel_pEU-E01-His-TEV-MCS-N1 _Fw	AACCTGTATTTTCAGGGCATGTTACAGCATCCC
SO_Dmel_pEU-E01-His-TEV-MCS-N1 _Rv	GTAAATTCTATACAACTATAAGTGCTGGTACTCGCCC
SIX4_Dmel_pEU-E01-His-TEV-MCS-N1 _Fw	AACCTGTATTTTCAGGGCATGTTTGACAAGAATTTGGAC
SIX4_Dmel_pEU-E01-His-TEV-MCS-N1 _Rv	GTAAATTCTATACAACTATTGCCCATTGAAAATCGT
optix_Dmel_pEU-E01-His-TEV-MCS-N1 _Fw	AACCTGTATTTTCAGGGCATGGCCGTTGGACCG
optix_Dmel_pEU-E01-His-TEV-MCS-N1 _Rv	GTAAATTCTATACAACTATGTGATCTCGGGAGCAC
So_Herato_pEU-E01-His-TEV-MCS-N1 _Fw	AACCTGTATTTTCAGGGCATGCTGGGTGGTCCG
So_Herato_pEU-E01-His-TEV-MCS-N1 _Rv	GTAAATTCTATACAACTAGGTATGATGGTATTGCAGATG
optix_Herato_pEU-E01-His-TEV-MCS-N1 _Fw	CTGTATTTTCAGGGCATGCGCGGTAGCTGG
optix_Herato_pEU-E01-His-TEV-MCS-N1 _Rv	GTAAATTCTATACAACTATTACTCTTCGTCAAC
SIX4_Hmel_pEU-E01-His-TEV-MCS-N1 _Fw	AACCTGTATTTTCAGGGCATGGAAAGTTGCAGTGATC
SIX4_Hmel_ppEU-E01-His-TEV-MCS-N1 _Rv	GTAAATTCTATACAACTACGGCGGATTACCATGC
SIX1_Huma_pEU-E01-His-TEV-MCS-N1 _Fw	AACCTGTATTTTCAGGGCATGTCGATGCTGCCG
SIX1_Huma_pEU-E01-His-TEV-MCS-N1 _Rv	GTAAATTCTATACAACTAGGACCCCAAGTCCAC
SIX2_Huma_pEU-E01-His-TEV-MCS-N1 _Fw	AACCTGTATTTTCAGGGCATGTCCATGCTGCCCAC
SIX2_Huma_pEU-E01-His-TEV-MCS-N1 _Rv	GTAAATTCTATACAACTAGGAGCCCAGGTCCAC
SIX3_Huma_pEU-E01-His-TEV-MCS-N1 _Fw	AACCTGTATTTTCAGGGCATGGTATTCCGCTCCCC
SIX3_Huma_pEU-E01-His-TEV-MCS-N1 _Rv	GTAAATTCTATACAACTATACATCACATTCCGAGTCG

SIX4_Huma_pEU-E01-His-TEV-MCS-N1 _Fw	AACCTGTATTTTCAGGGCATGGAAAGCGCCTCG
SIX4_Huma_pEU-E01-His-TEV-MCS-N1 _Rv	GTAAATTCTATACAACTATAAGTCTTGCATATCTTCATCC
SIX5_Huma_pEU-E01-His-TEV-MCS-N1 _Fw	AACCTGTATTTTCAGGGCATGGCGACGCTGC
SIX5_Huma_pEU-E01-His-TEV-MCS-N1 _Rv	GTAAATTCTATACAACTACAGCTCCAGTGGTTCC
SIX6_Huma_pEU-E01-His-TEV-MCS-N1 _Fw	AACCTGTATTTTCAGGGCATGTTCCAGCTGCCC
SIX6_Huma_pEU-E01-His-TEV-MCS-N1 _Rv	GTAAATTCTATACAACTAGATGTCGCACTCGCT

Table 3.3: Sequencing primers

Primer Name	Primer Sequence
T7 Promoter	TAATACGACTCACTATAGGG
T7 Terminator	GCTAGTTATTGCTCAGCGG
pET32_sequencing primer_Fw	ATGGACAGCCCAGATCTG
pEU-E01-His-TEV-MCS-N1_Fw_Seq_E01	CATTCAATCACTCTTTCCACTAACC
pEU-E01-His-TEV-MCS-N1_Rv pEU	CCTGATATAGGAAGGCCGG
So_Dmel_pET32_Fw2	ATTCCACCGGGGACAATAC
Six4_Dmel_pET32_Fw2	TATCCAATCTCACAGCCG
optix_Dmel_pET32_Rv2	GAAGCTCTGCAGGTAACAG
SIX4_Hmel_pET32_Rv2	GCCTTAAACCACAGCGTCTG
SIX4_Huma_pET32_Fw2	AGCTCTACAGCATCCTCG
SIX4_Huma_pET32_Fw3	TGGCATCACCAACCTCAG
SIX4_Huma_pET32_Rv2	ATGTCGCGGTTTAGCTCAG
SIX5_Huma_pET32_Fw2	AGATCTGTACCTCCGTGC
SIX5_Huma_pET32_Fw3	AGCCGTTCTGCTGAATGG
SIX5_Huma_pET32_Rv2	CACCCTCTGGCACACTAATC

Sequence validation was done on all constructs built with Gibson Cloning. The gene insertion of our SIX genes was successful in pEU-E01-His-TEV-MCS-N1. We decided to express our proteins under this system since all constructs were sequence-verified with the cell-free expression vector.

3.2.3. Cell-free expression of SIX proteins

The SIX proteins were expressed using a cell-free expression system bought from CellFree Sciences. The kit we used was the WEPRO7240; it comes with all the necessary reagents and products for the cell-free reaction, including the pEU-E01-His-TEV-MCS-N1 vector used in our cloning. To express our proteins, we used the guidelines and protocols established by the manufacturer. We also used a positive control that comes with the kit, used to express the GFP protein. If the procedure is done successfully, the control reaction must show the green fluorescence light emitted by GFP.

The first step consisted in producing the mRNA of each of our genes. This was done using 300 ng of each SIX plasmid in their respective reaction tubes. Each reaction tube has the pEU-E01-His-TEV-SIX-N1 construct, SP6 RNA Polymerase (Promega P1085), RNase Inhibitors (Promega N2611), NTP mix (NEB N0450L), transcription buffer LM (CellFree Sciences, Inc.) and water, for a reaction volume of 20uL. After the transcription mix was prepared, it was incubated at 37°C for 6 hours. This was done on a thermocycler to control the incubation temperature and time. After incubation, wheat germ extract and creatine kinase were added to the reaction tubes. This mixture was then added to mini dialysis cups with a 10K MWCO pore size (Thermo Scientific 88401). These dialysis cups were placed in a deep-well plate filled with a mixture of reaction buffers provided by the manufacturer. The reaction buffer is a mixture of four buffers named 40x SUB-AMIX SGC (S1-S4). Each buffer has the reagents required to produce our proteins. The components of the buffers are proprietary. The reaction buffers were used to fill the deep-well plate wells. Each cup was left in its respective well, not letting it fully submerged in the buffer, only allowing the necessary depth for the membrane to be in direct contact with the buffer. The deep-well plate was then wrapped in plastic foil and incubated for 20 hours at 15°C.

After incubation, the content of each dialysis cup is extracted and placed in 1.5 mL collection tubes. This mixture has our protein of interest. The GFP positive control is evaluated to confirm that the reaction was done effectively. Western blot is done to validate that the proteins were produced. Since the proteins have a His-tag, they are visualized using an Anti-HIS-HRP antibody [1 μL:10000 μL] (NOVUS NB100-63173) and ECL substrates (Bio-Rad 1705061). Proteins are confirmed based on their molecular weight.

3.2.4. Electrophoretic mobility shift assay

DNA-binding is critical for a TF to be functional. Because of this, we must evaluate that the produced proteins can successfully bind DNA. This is done by electrophoretic mobility shift assay (EMSA), in which we run our protein sample on a native gel in combination with a labeled DNA probe. This is done by identifying a DNA sequence that can be recognized by the TF, in this case, a SIX TF. We use the **Wnt1** enhancer sequence recognized by Six3[132]. We designed a 60bp-long DNA probe labeled with an infrared fluorescent dye. We also designed an alternative DNA probe (soAE regulatory element) based on a sequence obtained from a previous publication that used it for *Drosophila's melanogaster* sine oculis[60]. To prepare the ssDNA probe for EMSA, a fluorescent dye must be incorporated, this is done using an already marked ssDNA primer complementary to the constant regions of the ssDNA probe. This is done by producing dsDNA in a PCR reaction. **Table 3.4.** has the sequences for the DNA sequences used to prepare our probes with the corresponding fluorescent dye. The DNA oligos were purchased from Integrated DNA Technologies (IDT, Coralville, Iowa, USA) as standard desalted, except where noted otherwise. PCR purification columns are then used to purify the resulting dsDNA probe.

The EMSA is done using a mix consisting of our protein of interest and the label dsDNA probe. A specific binding buffer for our SIX proteins was determined and named the SIX binding buffer, done initially in a 10X concentration (250 mM HEPES, 750 mM NaCl, 10 mM EDTA, 100 mM $MgCl_2$). The SIX binding buffer is then diluted to a 5X working concentration, to which Glycerol is added to a final concentration of 50 %. A 1X SIX binding buffer dilution is done and used for control reactions or protein dilutions.

The reaction mix is made of (5.1 µL of SIX binding buffer [5X], 4 µL of pdI-DC [1000 ng/µL], 0.5 µL of Bovine Serum Albumin [1 mg/mL], 1 µL of Tween-20 [1%], 0.2 µL of Dithiothreitol [1 M], 0.75 µL of the label DNA probe [1000 nM] and 5 µL of our Cell-Free expressed SIX protein for a 15 µL final volume. A control reaction is prepared on which no protein is added, the protein volume is substituted by 1X SIX binding buffer. Each reaction is left at room temperature for an hour of incubation. A 6% acrylamide native gel is prepared for a 22 cm-high crystal chamber. The gel is then pre-ran with the (0.5X) TBE buffer for 15 minutes at 160V. After the gel is pre-ran, the voltage is reduced to 60V, and samples are added to each well. The gel is run for 2 hours and then visualized in an Azure Sapphire™ Biomolecular Imager (Azure Biosystems Inc., Dublin, California, USA) with laser excitation wavelength 658 nm and filter Red 710BP40.

Table 3.4. DNA oligos for EMSA.

DNA oligo name	Oligo Sequences
Wnt1 Promoter	CTTTACTCTCTCCCCAAGGGACATCTAATGATAAGCACAGGACACTTCTGCCCAGGCGAG
Wnt1 Promoter Scrambled	CTTTACTCTCTCCCCAAGGGAATTCAGTCCCGAAAGTAAAGACACTTCTGCCCAGGCGAG
SoAE Regulatory Element	CTTTACTCTCTCCCCAAGGGCTGGTAATTCGATATCATTGGACACTTCTGCCCAGGCGAG
SoAE Regulatory Element Scrambled	CTTTACTCTCTCCCCAAGGGATTGTTTAAAGTAGTTCCCGGACACTTCTGCCCAGGCGAG
IR700_rPCR_v2 (HPLC purify)	/5IRD700/CTCGCCTGGGCAGAAGTGTC

3.2.5. Systematic evolution of ligands by exponential enrichment (SELEX)

The DNA binding preferences of the SIX proteins were determined by employing the Systematic Evolution of Ligand and Exponential Enrichment (SELEX) assay. This is an *in vitro* technique used to determine the DNA targets of any DNA-binding protein. To determine the SIX proteins' binding preference, we used a SELEX variant called SELEX-seq[25]. To start, we use a 60bp long DNA-Library, with a 20bp randomized region flanked by constant regions used to enrich selected sequences and add Illumina compatible barcodes to each sample (See **Table 3.5.** for all DNA oligos used during the SELEX-seq). The DNA oligos were purchased from Integrated DNA Technologies (IDT, Coralville, Iowa, USA) as standard desalted except where noted otherwise. The 20bp randomized region gives 10^{12} unique sequences, which in theory, can provide coverage for each possible 20mer. The ssDNA library is made into a dsDNA library by PCR, this is done using the rPCR primer (from a 100 μM stock), which makes de dsDNA biotinylated. This PCR reaction is done in two independent PCR reactions. PCR purification columns are used to then purify the dsDNA libraries, and concentration is determined by UV/Vis Nanodrop (expecting a working range from 0.5-2 μM).

The binding reaction is done at the same concentrations as the EMSA procedure but in a final volume of 20 μL. For each protein, there is a total of three binding reactions, the dsDNA library (200 nM) + protein of interest, the dsDNA probe + protein of interest, and one only with the dsDNA probe. Since the library cannot be visualized, we used the dsDNA probe as a guide when running and visualizing the gel. Samples as previously described for EMSA, and each dsDNA library reaction are run beside their homologous dsDNA probe reaction. After the samples are run, they are visualized with the Azure Sapphire™ Biomolecular Imager (Azure Biosystems

Inc., Dublin, California, USA) with laser excitation wavelength 658 nm and filter Red 710BP40. This step allows us to determine if DNA-binding is observable in those lanes with dsDNA probe + protein. The produced image is then printed on a 1:1 scale, allowing the gel and the image to overlap fully. The gel with its back crystal cover is placed over the scaled image. This enables us to locate the samples on the gel using the controls seen in the picture. Using the dsDNA probe + protein lane as a guide, the corresponding lane of dsDNA library + SIX protein. All bound regions and unbound (free DNA) regions are cut from this lane. The cut gels are placed in EB buffer and incubated at room temperature with constant shaking overnight, allowing the DNA oligos to be released into the buffer. The bound sequences are purified using Dynabeads Streptavidin magnetic beads (Thermo Fisher 11205D). Using a magnetic well plate to isolate biotinylated oligos and magnetically selected oligos. DNA oligos corresponding to the bound fraction are washed and enriched (15 cycles) using fPCR and rPCR primers (both from a 10 µM stock) in a PCR reaction. The products are then purified, and their concentration is measured by UV/Vis Nanodrop, and this concentration is used to determine the required volume for another round of selection. This method shows what is considered the first round of SELEX-seq. This process is done two more times for a total of three rounds. Always using the bound fraction purified from the previous step to carry on to subsequent rounds. If any bound sequence resulted from a possible dimeric binding, we treated each bound shift independently during the SELEX-seq. In the case of multiple dimeric shifts in the same reaction, the bound sequences are mixed equimolar, maintaining each bound shift's representation.

The products of each round, including bound and unbound oligos, were then uniquely barcoded to be sequenced by Illumina sequencing. To barcode, each sample, 5 µL of every DNA sample, including DNA from all three rounds, negative controls (blanks), and the starting library, are run in a PCR reaction independently for each sample. This is done by using the rPCR + barcode and the fPCR + adapter (both from a 5µM stock). The sequencing was done using the services of Novogene. Samples were prepared and sent as recommended by the company with the corresponding sequencing primer. After sequencing, de novo motif analysis is done to determine the enriched sequences in our datasets.

Table 3.5. DNA oligos for SELEX-seq

SELEX-seq DNA oligo	Oligo Sequences
20 library (20N)	CTGATCCTACCATCCGTGCT(N)$_{20}$CACAGCTTCGTACCGAGCGG
Fw_primer (fPCR)	CTGATCCTACCATCCGTGCT
Rv_primer (rPCR) *Adds biotin	CCGCTCGGTACGAAGCTG
Fw_primer + Illumina adapter (fPCR + Ad)	AATGATACGGCGACCACCGAGATCTGCTCTTCCGATCTCTGATCCTACCATCCGTGCT
Rv_primer + Barcode (6p) +Illumina adapter (rPCR+barcode)	CAAGCAGAAGACGGCATACGAGATXXXXXXTCTTCCGATCCCGCTCGGTACGAAGCTGTG
Sequencing primer (HPLC purify)	GCTCTTCCGATCTCTGATCCTACCATCCGTGCT

3.2.6. De novo motif analysis

The analysis process of SELEX-seq data was done using the bioinformatical pipeline described by Nitta et al.[26]. Detection of sequences selected by our SIX proteins was done with the Autoseed algorithm, developed by Nitta and colleagues. This algorithm works by identifying all subsequences that are enriched more than any other relatable sequence. The program scans the datasets from SELEX-seq and detects motifs that are enriched based on a baseline, in this case, the starting library. We used two input files for our analysis: the library file that functions as the background and our SIX DNA bound sequences from our SELEX-seq rounds. We ran our analysis natively on our computers. The code that we used to determine the enrichment of our sequences is the following:

- **Autoseed running command structure**

```
./totalautoseed -20N <background file>.txt <SelexRound_SIXprotein>.txt 1 8 10
0.35 - 50 40 > <output file name>.txt; cp Kmer_summary8to10.svg <output file
name>.svg
```

- **Autoseed running example for Round 3 Human SIX1**

```
./totalautoseed -20N R50_333.txt R50_SIX1.txt 1 8 10 0.35 - 50 40
>SIX1_R3_097_R50_333.txt; cp Kmer_summary8to10.svg SIX1_R3_097_R50_333.svg
```

This command was run for each of our proteins and was used to determine the sequences that were bound and enriched for every SIX protein. Following analysis, we use Autoseed to determine the binding matrices for each enriched seed we have identified. This is done to represent our analysis in a DNA Logo format to provide information regarding the binding of our SIX proteins. To determine the binding matrices for our proteins, we used the following code:

- **Autoseed command to determine binding matrices based on a sequence seed**

```
./spacek40 --f -dinuc -nocall -m=1 -20N -q R47_666.txt R47_142.txt NYGATACN
200 | grep "One Hit" > DmelSo_R47_666_R47_142_NYGATACN.pfm


#Crea el Logo del motivo a base del pfm
./spacek40 --logo DmelSo_NYGATACN.pfm DmelSo_R47_666_R47_142_NYGATACN.svg
```

Subsequently, the position frequency matrix that was determined by Autoseed was used in an R pipeline to better represent the information stored within the matrices.

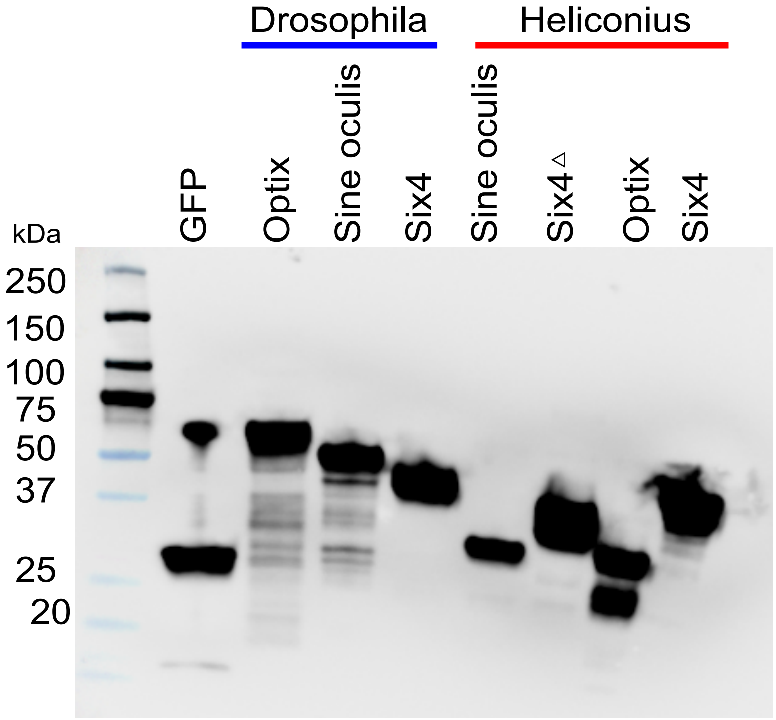
3.2.7. Specificity matrix comparison

Specificity matrices were obtained using the Autoseed pipeline and curated to only the extended 12bp DNA-binding motif[26]. Matrices were compared using the RSAT web server http://rsat.eead.csic.es/plants/compare-matrices_form.cgi **(Supplementary image 5)**. Correlation values were then visualized in RStudio [Rversion & Rstudio] and heatmaps were generated with correlation values.

3.3. Results

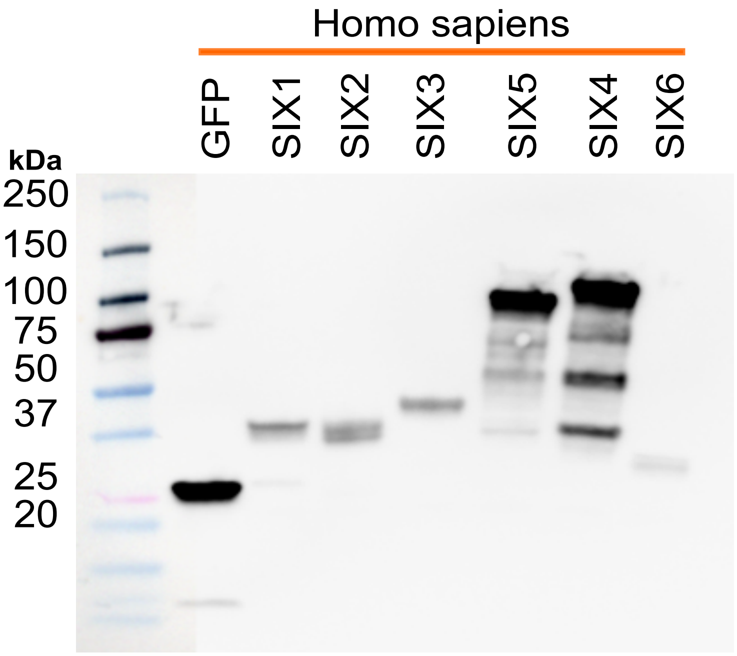
3.3.1. Protein expression and SELEX-seq

The SIX proteins (Sine oculis, optix, and Six4) from *Drosophila melanogaster* and *Heliconius erato* were successfully expressed (**Figure 3.1**). The human paralogs (SIX1-SIX6) were also successfully expressed (**Figure 3.2**). The binding of each SIX transcription factor was evaluated using Electrophoretic Mobility Shift Assay (EMSA). A total of 8 SIX proteins were able to bind to the Wnt1 enhancer sequence (**Figure 3.3**). The remaining SIX proteins, Sine oculis from *Drosophila* and *Heliconius*, Six4 from *Heliconius*, and optix from *Drosophila*, didn't bind to the Wnt1 promoter.



Protein	Weight (kDa)
GFP	27
Dmel/sine	48
Dmel/six4	46
Dmel/optix	55
Heli/sine	28
Heli/six4∆	34
Heli/optix	32
Heli/six4	46

Figure 3.1. Anti-his Western blot of SIX proteins from *Drosophila* and *Heliconius*. All SIX proteins from *Drosophila melanogaster* and *Heliconius erato* were expressed. Six4∆ is a non-HD isoform of the full-length Six4 in *Heliconius*. GFP is a positive control for this blot.



Protein	Weight (kDa)
GFP	27
SIX1	34
SIX2	34
SIX3	38
SIX4	83
SIX5	78
SIX6	30

Figure 3.2. Anti-his Western blot of SIX proteins from *Homo sapiens.* All SIX proteins from Human paralogs were expressed. GFP is a positive control for this blot.

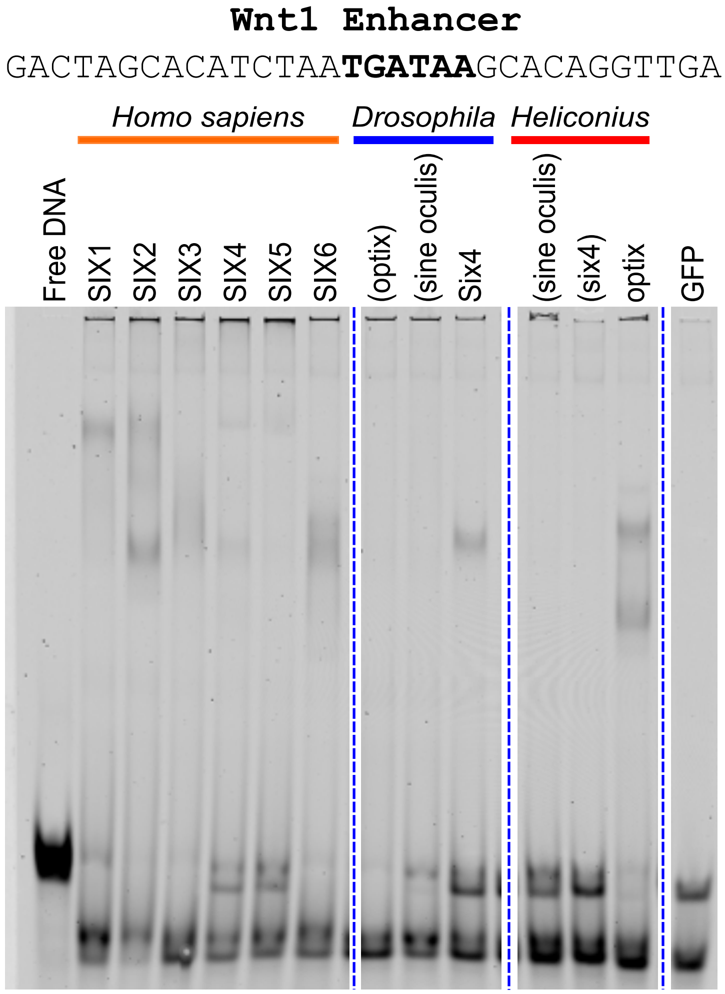


Figure 3.3. EMSA of SIX proteins binding to the Wnt1 enhancer. 8 of 12 SIX proteins bind to the Wnt1 promoter. No binding was observable in the remaining SIX proteins illustrated with parenthesis. GFP functions as a negative control for this assay.

The SIX proteins that didn't bind to the Wnt1 promoter were tested for binding using the soAE (sine oculis autoregulatory element)[60]. To which *D. melanogaster* sine oculis and optix were able to bind to the regulatory element successfully, this was also observable in *Heliconius* sine oculis and six4 (**Figure 3.4**). SELEX seq was done using the results from these EMSA assays and was successfully completed for every SIX protein. All samples were run using the dsDNA probes determined by our EMSA assays. **Figure 3.5** shows a SELEX-seq native gel of *D.melanogaster* SIX proteins at Round 3. **Supplementary image 4** has all SELEX gels from Round 3.

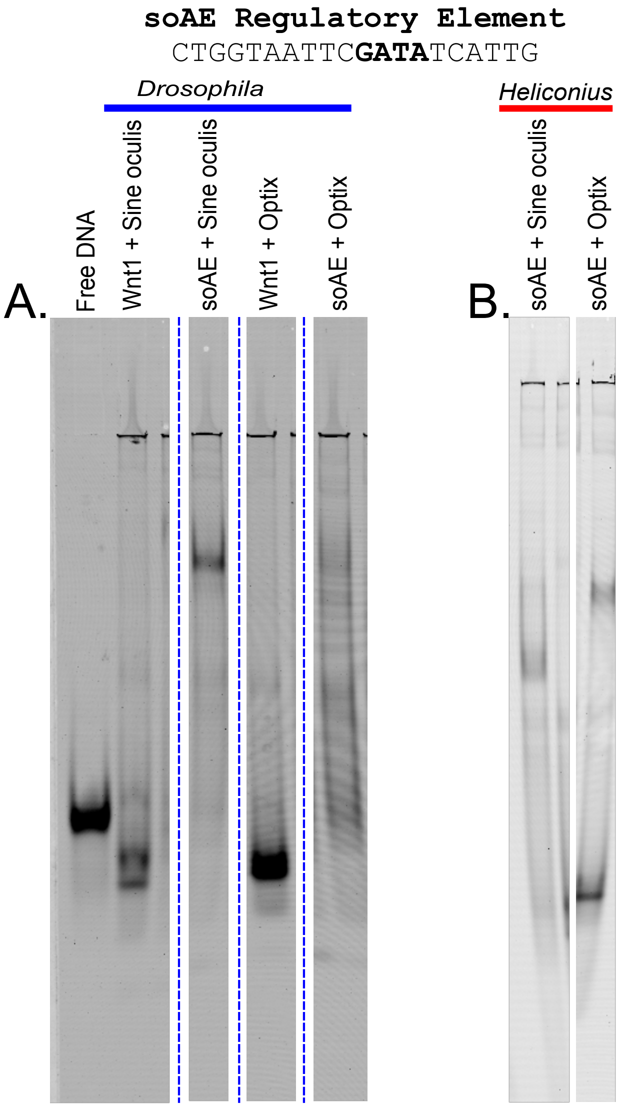


Figure 3.4 EMSA of SIX proteins binding to the soAE regulatory element. The binding of *D. melanogaster* sine oculis and optix; *Heliconius* sine oculis, and six4 were observable when using the sine oculis regulatory element.

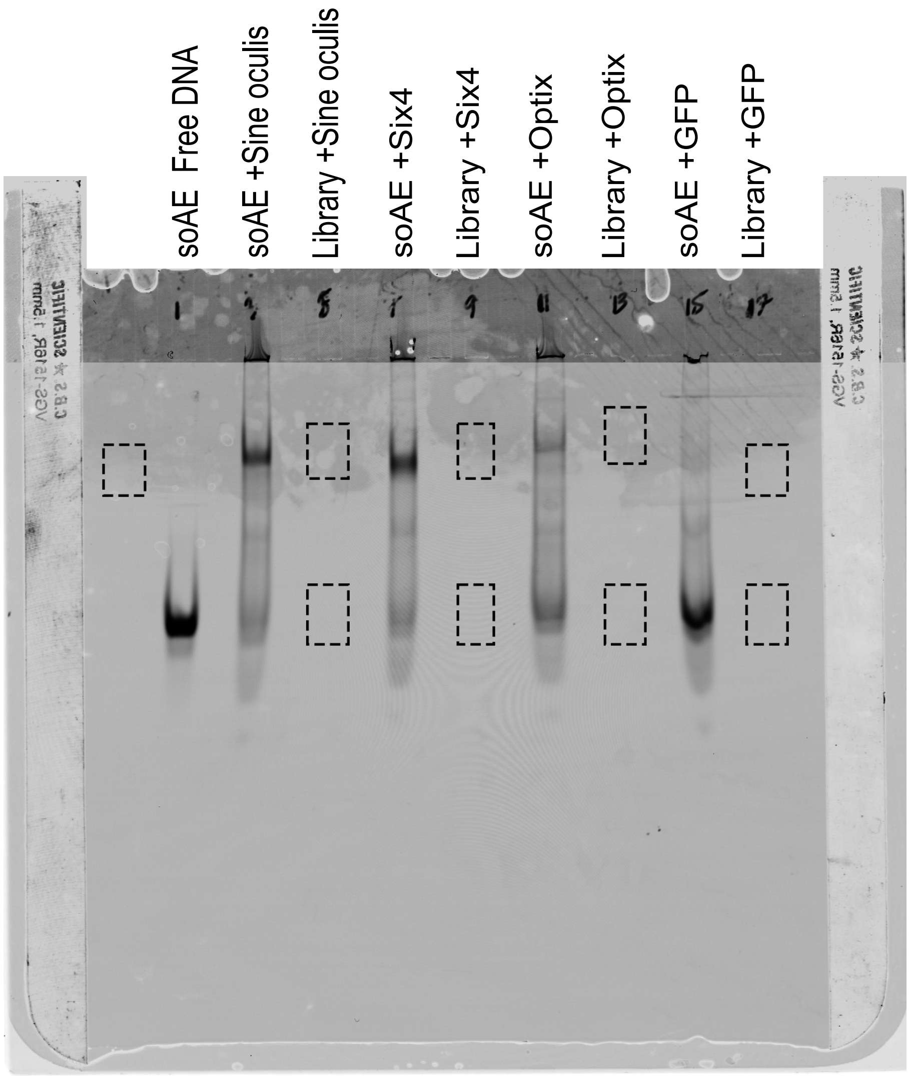


Figure 3.5. EMSA gel of Round 3 of SELEX-seq for SIX proteins on Drosophila melanogaster. Cut regions are marked by a dashed square in the dsDNA library-SIX protein lane. Unrelated cut regions are used as a negative control for subsequent steps.

3.3.2. SELEX-seq analysis

We did SELEX-seq to the SIX proteins found in *Drosophila melanogaster*, *Heliconius erato*, and *Homo sapiens*. SELEX-seq was done as described SELEX seq was done as described by Slattery et al.[133] and Riley et al. [25]., using a synthetic library with a 20bp randomized region that offers 10^{12} unique sequences or all possible 20-mers. Selected sequences were enriched by three selection rounds; each round was sequenced and analyzed bioinformatically. **Figure 3.6** shows the PWM Logo of each SIX protein at Round 3. The results from this analysis show the following:

- **5' Flank region:** All SIX proteins in our analysis showed a 5' flanking region before the SIX core motif of `TGATAC`. The enrichment of this region depends on the SIX protein subfamily. The sine oculis/SIX1/2 subfamily demonstrates an enrichment in the region with a consensus sequence of `GNAANN` in both *Drosophila* and *Heliconius*. This flanking region is observable in both SIX1/2 but with a lower enrichment, showing differences between arthropods and humans. The flanking region is also observable in six4/SIX4/5 with a similar consensus sequence to sine oculis/SIX1/2. This sequence is also conserved between arthropod orthologs and human paralogs. In optix/SIX3/6, this region shows a lower enrichment than the other proteins but has some similarities; the same was observable with optix paralogs.

- **Flanking region length**: The core motif is conserved between proteins with slight divergence observable in six4/SIX4/5 that demonstrate a `YGACAC` from the canonical `YGATAC`. When comparing the 5'Flanking region, it is noticeable that both sine oculis/SIX1/2 and six4/SIX4/5 have a flanking region 6bp long before the core motif. However, in optix/SIX3/6, it can be observed that it is 1bp shorter than in the other proteins. In this subgroup, the enriched region flanking 5' the core motif is only 5bp long. These observations are also seen in the human paralogs, highlighting a conserved trait between optix/SIX3/6 members.

- **Binding specificity of the SIX proteins**: Binding correlations were done to our proteins to evaluate their similarities and differences (**Figure 3.7**). Using the binding matrixes from our bioinformatical analysis, we determined their correlation based on the 5' Flank and the core motif (12-11bp) and ran it on the RSAT web server. **Figure 3.7** shows the heatmap correlation of this analysis. The results demonstrate different binding profiles for some proteins. For example, one profile is observably composed of the optix/SIX3/6 proteins, revealing similar binding specificities. Another group has SIX1/2 not associated with their Arthropod orthologs. Six4 from Heliconius has its own profile, not included in other profiles. The remaining profiles are those of sine oculis from *Heliconius* and *Drosophila* and all Six4 proteins.

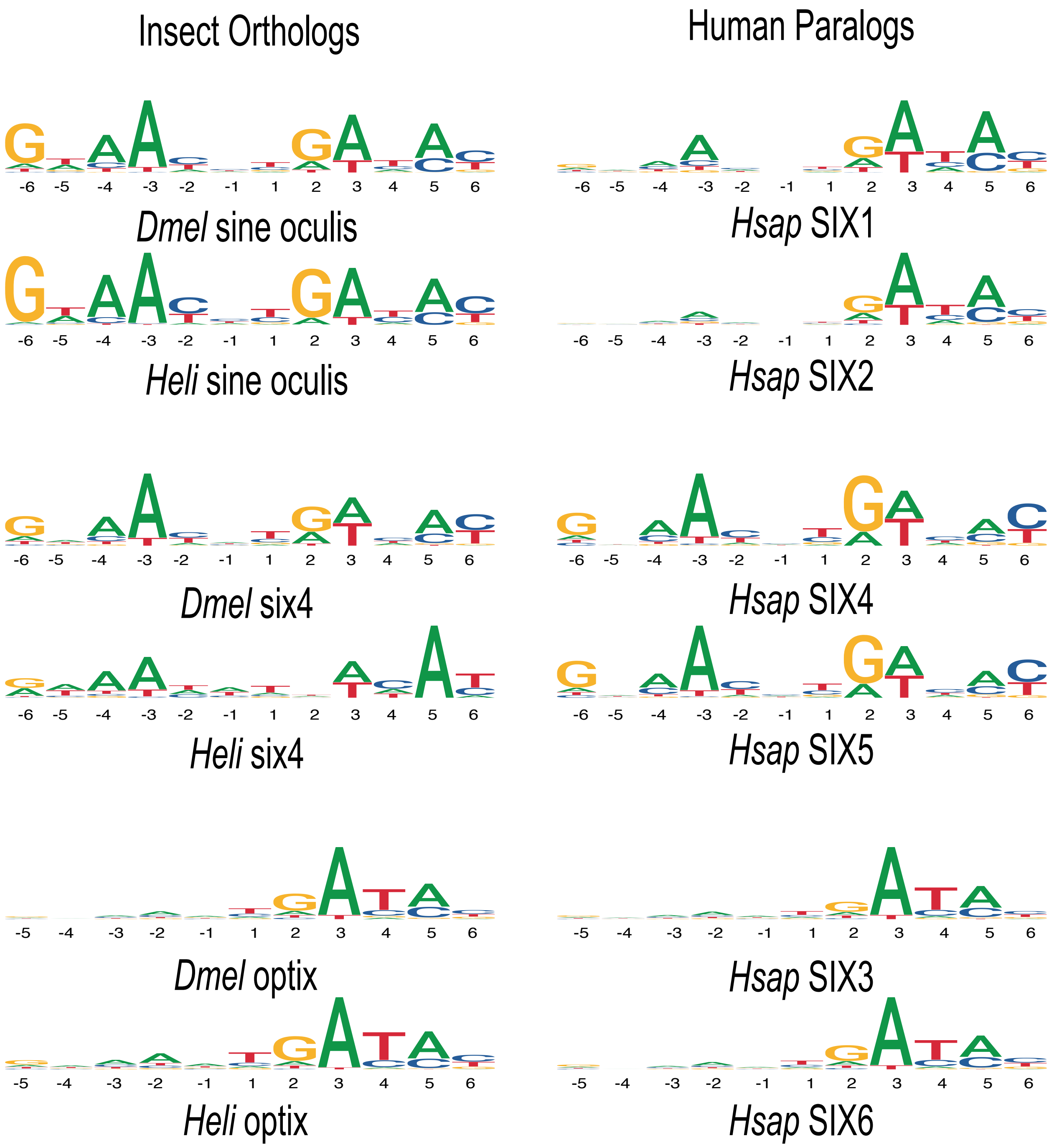


Figure 3.6. Informational PWMs of SIX proteins: Binding informational PWM derived from our SELEX-seq data. Comparing orthologs between paralogs demonstrates a similar `YGATAC` core motif with a 5’ flanking region on all samples. Flanking between Sine oculis orthologs and paralogs is similar but with lower enrichment on the latter. Six4-like proteins share the same binding specificity profile, while optix demonstrates a 1bp shorter extended motif.

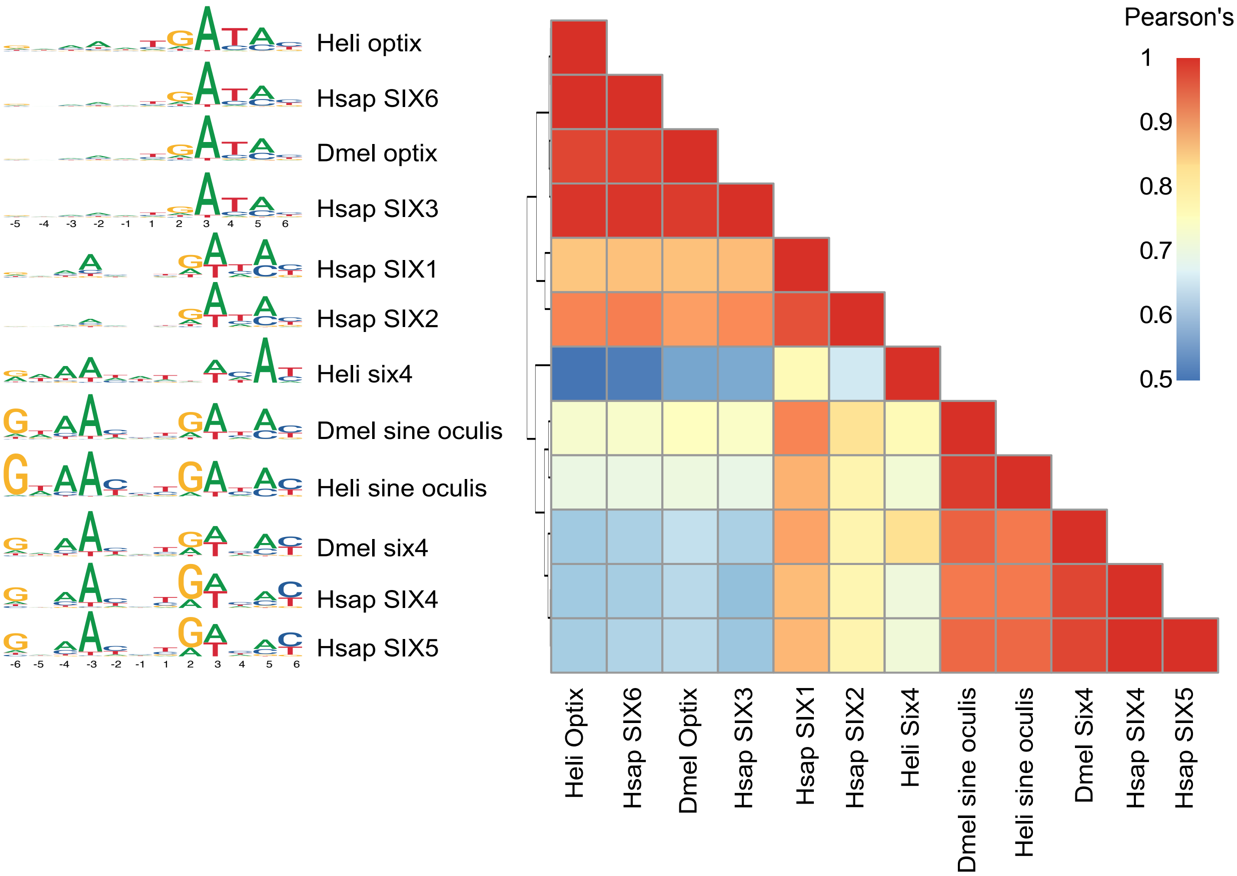


Figure 3.7. Comparison of SIX protein binding matrixes: RSAT binding correlation of PWMs from SELEX-seq data. Specificities are organized into four binding profiles. optix/SIX3/6 have their binding profile, while the SIX1/2 profile is found outside their orthologs and correlates closely with optix/SIX3/6 class. For the sine oculis proteins organizing within the same functional class. The six4/SIX4/5 proteins are found within their profile. Six4 from *Heliconius* is independent, possibly because of its low enrichment.

3.4. Discussion

A transcription factor's specificity is believed to be evolutionarily restricted to change since even minimal modifications could have detrimental effects on its role within gene regulatory networks. However, changes in specificity have been observed on some TFs, where members of the same family can interact with new genomic targets by having new specificity profiles, with some recognizing multiple sequence targets[31–33,105,106]. Genomic duplications that lead to new protein-coding genes allow for some specificity divergence to be plausible since one copy can retain the ancestral role while new copies can change or be deleted [103]. If change does happen, it can materialize between the specificity of a TF to the DNA.

It is possible that the SIX family can follow the same trend observed in other TFs families. In our EMSAs we were able to observe and determine that, indeed, there were different DNA-binding preferences between some SIX proteins. Even when we had successfully seen DNA-binding using the Wnt1 promoter with the majority of SIX proteins, 4 out of 12 (*D. melanogaster* sine oculis and optix) and (*Heliconius* sine oculis, and six4) could not bind to the Wnt1 sequence. At first, those proteins might have been expressed but nonfunctional. However, when using the soAE regulatory element, these 4 SIX proteins' DNA-binding capability was observed. The main difference between both elements is the presence of a 5' Flank found upstream of the core SIX binding motif of (TGATAC). This region has been reported to be critical for *Drosophila's* sine oculis to bind to DNA[134]. Sequence mutation within the 5'GA flank is known to effectively inhibit the capability of Drosophila's Sine oculis to bind to the regulatory element [134].

Surprisingly previous *in vitro* assays reported within the cis-bp database, a catalog of inferred sequence binding preferences, show only the SIX core binding motif without any mention of the flanking region[26,40]. Some of the reported motifs were obtained using SELEX. Our SELEX-seq experiment could observe the 5' Flanking region of the SIX proteins. It allowed us to notice both its enrichment and configurations that demonstrate that the SIX bind to an extended binding motif. This extended motif is only found within the cis-bp catalog in ChIP-seq experiments. The resolution of our SELEX-seq method allowed us to see the binding motif observable on *in vivo* experiments. Observing these enriched regions not reported in previous SELEX experiments can come from multiple sources. One of these could be an intrinsic advantage in our starting DNA library that allowed the extended motif to be found within the DNA library. Another possibility is based on our decision to express the complete protein sequence for each SIX protein. Most of the work done in determining the specificity of a TF is usually done using only the DNA-binding domain since it is challenging to express a TF in its entire length.

Most work done *in vitro* to the SIX has been done using only their DNA binding domain, in this case, its homeodomain. However, using only this region to determine specificity can allow for some information to be lost since the HD isn't entirely responsible for all the SIX binding specificities properties. Our capacity to observe the extended binding motif may result from our preference for using SIX full-length proteins. There are scientific works that support this idea. For example, work done between Six2 and Six6 has demonstrated that Six6 can bind to the Myo-MEF3 motif while Six2 couldn't[62]. Substituting the Six2 and Six6 C-terminal between factors revealed that Six2 could bind to the Myo-MEF3 motif while now Six6 couldn't. This demonstrated that the binding specificity between SIX proteins can be regulated by regions outside the HD and confirms that our decision to express full length SIX proteins was the correct method.

The specificity of our 12 SIX proteins is similar but with contrasting features. When comparing sine oculis/SIX1/2, it is notable that all four proteins share the same core binding motif of (YGATAC). The sine oculis from *Drosophila* and *Heliconius* share a similar extended motif of (GNAANNYGATAC) which shows that, at least in insects, the specificity of sine oculis hasn't diverged and has remained conserved. Interestingly sine oculis paralogs in humans (SIX1/2) share the core motif, but the 5' flank is not as enriched in both SIX1/2 but is still observable. This is interesting since the Wnt1 promoter used in our EMSA's and SELEX-seq lacks the 5' flank, and both SIX1/2 can bind in its absence (**Figure 3.3**). In contrast, *Drosophila* and *Heliconius* sine oculis need the 5'flanking region to bind successfully to the core motif (**Figure 3.4**). These differences could explain why some can bind with or without the 5' flank and highlight specificity differences between sine oculis/SIX1/2. The differences between orthologs and paralogs binding specificity profiles are observable in **Figure 3.7**, where the binding correlation between sine oculis and SIX1/2 demonstrates two functional classes. SIX1/2 may have diverged in a way that they can bind to DNA sequences lacking the 5' flank, unlike their insect counterparts.

The six4/SIX4/5 specificity shows low divergence when comparing insects and humans. The core motif between proteins and the flanking region has remained highlight conserved, with minimal alterations to the core motif, with an enrichment motif of `YGACAC` but comparable to the `TGATAC` canonical motif. The 5' Flank and the extended motif of six4/SIX4/5 is similar to the one observed in sine oculis/SIX1/2. This is interesting considering that, as mentioned previously in Chapter 2, Six4-like proteins are more evolutionary related to optix/SIX3/6 than to sine oculis/SIX1/2. Some functional conservation must have been retained from the last common ancestor between six4/SIX4/5 and optix/SIX3/6 with sine oculis/SIX1/2. Since these proteins originated in early Metazoans, some conservation may have been retained between six4 and sine oculis. Interestingly, no specificity change has occurred between six4 from insects to their human paralogs. A possible reason is that there is redundancy between both proteins, and neither has diverged enough to differentiate. Studies have demonstrated that SIX4 KO is not associated with any disease and no apparent effect is observable in its absence. Some reports have suggested that the SIX4 role has changed to interact with other proteins, with one report demonstrating kidney disruption when KO both SIX4 and SIX1 [82]. More research is needed to fully understand SIX4 role in development since it appears to have a cooperative role by interacting with other family members like SIX1. Regarding SIX5, it has been associated with BOR syndrome [84], a syndrome associated with malformations to the ears and kidneys. It is also possible that SIX4/5 share specificity but are spatial and temporally isolated, which could negate the effects of redundancy.

The six4/SIX4/5 subgroup is the least studied, and more research is needed to understand their roles thoroughly. The six4 binding motif from *Heliconius* is the least enriched of all four evaluated. However, the extended motif is still visible but with less enrichment. This could have happened for many reasons, but protein stability during the SELEX-seq process seems to be the possible cause.

The optix/SIX3/6 specificity is conserved between both orthologs and paralogs. The core motif is observable with similarities with sine oculis/SIX1/2. The 5' flank is observable, but its enrichment is the lowest between SIX protein subgroups. However, the length between the flank and the core motif is 1bp shorter than other SIX proteins, and this trait was only observable with all optix-related proteins. The effect on specificity from these proteins appears not to alter the core motif. The correlation between optix/SIX3/6 shows highly similar values, and all four proteins are located within their functional class. A shorter extended motif could result from the interaction between optix/SIX3/6 and the DNA. A critical consideration about optix/SIX3/6 is that, as mentioned previously, this subgroup can have unique characteristics not observable in other SIX subgroups, with its C-terminal tail contributing to their specificity. In our binding assays, we can observe that optix from *Drosophila* and *Heliconius* bind to different dsDNA probes (**Figure 3.4**). Even when both share the same binding specificity profile, they recognize DNA differently. Other contributing factors could explain these differences, but it is probable that the divergent C-terminal tail plays an essential role in differentiating optix/SIX3/6 between species.

3.5. Conclusion

Transcription factors, like all other components in biology, are under constant evolutionary pressure to be part of new phenotypic innovations. These proteins are expected to retain their specificity and continue to carry with their evolutionary roles, even when it has been shown that specificity is capable of diversifying. Based on the notion that specificity is set in stone limits our understanding of how TF families originate and how their cooption happens. Specificities can change even between a family. The SIX demonstrate divergent specificities that highlight the distinctiveness of each protein. Expanding the knowledge into how TF evolved is essential to understand the dynamics of gene regulation while also contributing to learning about animals' origins and how they came to be.

This work has demonstrated differences that make each protein subgroup unique. We showed the presence of an extended binding motif that, for some SIX proteins, is essential for DNA-binding, while others seem to have diverged to bind in its absence. SIX4/5 haven't changed since insects, and their binding profile has been conserved. At the same time, other SIX proteins have similar but shorter DNA-binding motifs, which could result from protein-specific modifications that extend their HD. We were able to observe that there are four functional binding profiles from our analysis. This work shows that specificity can change and that the SIX family is an example of specificity evolution. To further expand our knowledge about changes in specificity, it will be important to consider studying the specificity of more ancestral SIX protein lineage. Expanding our current knowledge of SIX specificities can provide insight into how the binding specificity of this family originated and how it has changed our current results.

Chapter 4: *Heliconius erato* DNA-binding specificity of optix.

4.1. Introduction

Animals have multiple defense mechanisms that improve their survival. Some animals use colors and patterns to deter predators that associate the colors and their patterns with a foul taste. For example, some species use Mullerian mimicry to copy color patterns to increase their survivability. The butterflies from the *Heliconius* genus have evolutionary converged to share the same color patterning in a geographical area. This convergence has been observed to occur between pairs of distance species from the same genus[135]. For example, it has been known that *Heliconius melpomene* mimics the color patterns of *Heliconius erato*[135]. The scales on the *Heliconius* wings have a mix of colors between blacks, yellow and red scales, with each pattern being copied between *H. erato* and *H. melpomene*. Extensive studies have been done to understand how these originate and the regulatory mechanism that promotes each color and its patterning[91,135–142].

The transcription factor optix was identified as the master regulator of the red color pigmentation in the *Heliconius* wings [90]. Transcription factors interact directly with the DNA to promote or repress gene expression. The dynamics in regulation are the mechanism by which multiple phenotypes develop. In the case of optix, it has been known to be involved during eye development in *Drosophila,* with some research demonstrating phenotypic roles within the wings [56,63,86,87]. The *Heliconius* optix appears to have been coopted into new regulatory pathways involved in red color pigmentation. This has been demonstrated in CRISPR optix KO cells, the red color pigmentation is substituted by nonred colors[90,143]. It has been proposed that for optix to promote the red-colored wing pigmentation, there must have occurred a cooption within the ommochrome pathway observed in flies' eyes and now used within the *Heliconius* wings. The ommochrome red pigment has been observed in both eyes and wings [144–146]. Amber trapped extinct Diptera have been observed to have a wing organ like dipterans eye[88,147].

As a transcription factor, optix must bind to DNA to regulate gene expression, with each species genome having all the regulatory information required for proper development. Determining the DNA sequences bound specifically by a TF can provide helpful information about the genes regulated by it. The specificity of a TF determines the sequences recognized by the TF and alternative specificities based on interactions with other proteins, like cofactors. In addition, interactions between TFs can also play regulatory roles. TFs can dimerize, forming both heterodimers and homodimers that can enhance functionality by improving the selectivity of the factor. Understanding how TF binds to DNA can provide helpful information explaining the phenotypes seen in multiple species.

Studying the mechanism of the diversity of red wing coloring and patterning observed in Heliconius butterflies provides an excellent model to study phenotypic variations and the regulation process underlying it[91]. The *Heliconius* genus has diversified during the past 12 million years and has morphologically evolved accordingly to the regions they are encounter[91]. These patterns have diversified and converged, with different species sharing the same geographical color patterning[89,148]. These color patterns are the results of gene regulation, with genomic elements being responsible for their variability[91,148]. Understanding how *Heliconius* optix interacts with DNA can provide insight into the dynamics of gene regulation that leads to the red wing-colored patterns that are observed.

The DNA-binding specificity of *Heliconius* optix is still not characterized. Here we determine the DNA-binding specificity of *Heliconius erato's* optix. By doing Systematic Evolution of Ligands by Exponential Enrichment (SELEX), we determined the binding preferences of optix.

Using bioinformatical tools allowed us to predict genome-wide targets of optix using available data. Using these predictions, we evaluate the binding of optix to cis-regulatory elements. Our work will provide essential information on the mechanism by that optix interacts with DNA that can provide an explanation of red color pigmentation dynamics in *Heliconius erato*.

4.2.Material and methods

4.2.1. *Heliconius erato* optix cloning

The *Heliconius erato* optix gene was synthesized using the https://www.genewiz.com/ Standard Gene Synthesis service (Genewiz, South Plainfield, NJ). The sequence for optix was obtained from the Universal Protein Resource (Uniprot https://www.uniprot.org/) and is identified by the ID: L7X1S6. The optix gene was inserted into two types of expression vectors. We use two expression systems, one for bacteria expression and the second for expressed proteins in a cell-free system. The vector utilized for bacterial expression was the pET32a (+) vector from Millipore Sigma (Cat. 69015) (MilliporeSigma, Burlington, MA, USA). This plasmid was chosen since it has a Thioredoxin solubility tag, a S•Tag, and a double 6xHis-tag for protein purification. The system we decided for cell-free protein expression was from CellFree Science (CellFree Science Ehime, Japan) (https://www.cfsciences.com/eg/). The manufacturer recommends the expression vector pEU, which is optimized for the system sold by the company. The plasmid selected was the pEU-E01-His-TEV-MCS-N1 expression vector. If needed, this plasmid has a TEV cleavage site and an N-terminal 6xHis-tag for protein purification.

To clone *optix* into these expression vectors, we used Gibson Assembly[131]. This process is more precise and streamlined than using restriction enzymes. Gibson cloning requires the linearized plasmid and the *optix* gene to have a 15bp flanking region on both sides. These regions overlap with the vectors. All primers used can be found in **Table 4.1**. The linearized vector and the optix insert were produced using a high-fidelity polymerase from New England BioLabs (New England Biolabs [NEB], Ipswich, MA, USA). We used the Phusion High-Fidelity (HF) DNA polymerase master mix. To determine the annealing temperature for all reactions, we used the NEB Tm Calculator (https://tmcalculator.neb.com/#!/main). The *optix* insert reaction was purified using the QIAGEN (Qiagen, Germantown, MD, USA) QIAquick PCR Purification Kit (Cat. No. / ID:28106). The pET32a(+) and the pEU-E01-His-TEV-MCS-N1 vector were linearized and purified using the QIAGEN (Qiagen, Germantown, MD, USA) QIAquix Gel Extraction Kit (Cat. No. / ID: 28707.) after running each reaction in a 1% Agarose gel.

To develop our plasmid constructs, we used the Gibson Assembly Cloning Kit from NEB (Cat. E5510). The reaction was done as the manufacturer suggested and using the NEB Ligation Calculator (https://nebiocalculator.neb.com/#!/ligation). The vector mass used was 30ng of the linearized plasmid, and the *optix* insert was added in a 5:1 ratio. The reaction was incubated at 50^{o}C for an hour. The products of this reaction were used to transform competent *E. coli* bacteria. We used the NEB 5 alpha Competent E. coli (Cat. C2987H). The process was done using the High-Efficiency Transformation Protocol suggested by the manufacturer and adding 5 µL of the ligation reaction. Luria-Bertani (LB) plates were used to validate the transformation procedure; these had ampicillin as the selective agent. Plates were left in an incubator at 37^{o}C overnight, and colony growth was evaluated the next day. Bacterial colonies were isolated from these plates and used as templates for Colony PCR. This procedure confirms that *optix* is within the vector. We used NEB Taq 2X Polymerase Master Mix (M0270L) and specific primers for each vector. In the case of pET32a(+), we used the standard T7 promoter and terminator primer sequences. For inserts within the pEU-E01-His-TEV-MCS-N1, we used custom-made primers based on the SP6 promoter found in the vector and a reverse primer named MCS-Rv (**Table 4.1**) to cover all the insert regions. All primer's annealing temperatures were determined using the NEB Tm Calculator, and reactions were done according to manufacturer protocol. PCR reaction mixes were run by electrophoresis in an Agarose 1.5% gel, and insert presence was done based on the expected weight for each amplicon. Bacterial colonies confirmed to have the *optix* were left overnight for no more than 16 hours. The plasmids were then purified using the Qiagen QIAprep Spin Miniprep Kit (Cat. No. / ID: 27106). These plasmids were then sent for Sanger Sequencing utilizing a selection of standard primers (T7 promoter and terminator) and custom-made primers for some genes of interest (**Table 4.1**).

Table 4.1 DNA oligos used for cloning and sequencing

Primers for Gibson Cloning

Primer Name	Primer Sequence
optix_Herato_pET32_Fw	GACGACGACGACAAGATGCGCGGTAGCTGG
optix_Herato_pET32_Rv	GTGGTGGTGGTGGTGCTCTTCGTCAACGTT
optix_Herato_pEU-E01-His-TEV-MCS-N1 _Fw	CTGTATTTTCAGGGCATGCGCGGTAGCTGG
optix_Herato_pEU-E01-His-TEV-MCS-N1 _Rv	GTAAATTCTATACAACTATTACTCTTCGTCAAC

Plasmid Sequencing Primers

Primer Name	Primer Sequence
T7 Promoter	TAATACGACTCACTATAGGG
T7 Terminator	GCTAGTTATTGCTCAGCGG
pET32_sequencing primer_Fw	ATGGACAGCCCAGATCTG
pEU-E01-His-TEV-MCS-N1_Fw_Seq_E01	CATTCAATCACTCTTTCCACTAACC
pEU-E01-His-TEV-MCS-N1_Rv pEU	CCTGATATAGGAAGGCCGG

4.2.2. *Heliconius erato* optix overexpression

4.2.2.1. Cell-free expression

The optix protein from *Heliconius erato* was expressed using a cell-free expression system bought from CellFree Sciences. The kit used was the WEPRO7240; it comes with all the necessary reagents and products for the cell-free reaction, including the pEU-E01-His-TEV-MCS-N1 vector used in our cloning. To express optix, we used the guidelines and protocols established by the manufacturer. We also used a positive control that comes with the kit, used to express the GFP protein. If the procedure is done successfully, the control reaction must demonstrate the green fluorescent light emitted by GFP.

The first step consists in promoting the transcription of the *optix* gene from the pEU-E01-His-TEV-optix-N1 plasmid. The transcription reactions consists of 300ng the His-TEV-optix-N1 plasmid, SP6 RNA Polymerase (Promega P1085), RNase Inhibitors (Promega N2611), NTP mix (NEB N0450L), transcription buffer LM (CellFree Sciences) and water, for a reaction volume of 20uL. After the transcription reaction is prepared, it was incubated at 37^{o}C for 6 hours. This step is done on a thermocycler to control the incubation temperature and time. After incubation, wheat germ extract and creatine kinase were added to the reaction tube. This mixture was then added to a mini dialysis cup with a 10K MWCO pore size (Thermo Scientific 88401). The dialysis cups were then placed in a deep-well plate filled with a mixture of reaction buffers provided by the manufacturer. The reaction buffer is a mixture of four buffers named 40x SUB-AMIX SGC (S1-S4). Each buffer has the reagents required to produce our proteins. The components of the buffers are proprietary. The reaction buffers were used to fill the deep-well plate wells. Each cup was left in its respective well, not letting it fully submerged in the buffer, only allowing the necessary depth for the membrane to be in direct contact with the buffer. The deep-well plate was then wrapped in plastic foil and incubated for 20 hours at 15°C

After incubation, the content in the dialysis cup is extracted and placed in a 1.5 mL collection tubes. This mixture is expected to have optix. The GFP positive control is evaluated to confirm that the reaction was done effectively. Western blot is done to validate that both GFP and optix were expressed. Since both have a His-tag, they can visualize using an Anti-HIS-HRP antibody [1 μL:10000 μL] (NOVUS NB100-63173) and ECL substrates (Bio-Rad 1705061). Proteins are confirmed based on their molecular weight.

4.2.2.2. optix bacterial expression and purification

The *Heliconius erato* optix was overexpressed using a bacterial system. The bacterial strain chosen for this work were the BL21-CodonPlus *E. coli* Competent Cells from Agilent (230245). This bacterium was transformed using the pET32a(+) plasmid carrying the *optix* gene. The transformation process was done based on the recommendations of the manufacturer. An overnight culture of 10 mL of the transformed bacteria was prepared in LB media. This overnight culture was used as inoculum for a 1 L TB media, split (500 mL) between two conical flasks of 2L. These are left at 37°C with constant shaking until they have an OD_{600} measurement of 0.5 – 0.7. The production of optix was done by adding Isopropyl ß-D-1-thiogalactopyranoside (IPTG) at 0.5 mM. The induction reaction was left overnight for 15-19 hours. Cells were harvested by centrifugation (~15000 Gs') and stored at -80 °C overnight.

Cell lysis was done by sonication to disrupt the cell membrane the following day. The bacterial pellet was resuspended using a sonication buffer (NaCl [500 mM], Tris -HCl [20 mM], Tween-20 [0.2 %], and Imidazole [30mM] to a final pH of 8). Protease inhibitors are also added to the sonicator buffer (ThermoFisher A32955). The sonication process was done on an amplitude of 40% and done in an On/Off cycle of 30 seconds for a total of five rounds. Soluble protein was separated from the insoluble fraction by centrifugation at max speed (~150000 G's) for 30 minutes. The supernatant is then used for the purification process of the protein.

The presence of a His-tag in optix allows for purification using affinity chromatography. The purification of optix was done by adding Ni-NTA agarose (QIAGEN Cat.No ID: 30210) to a purification column. The column was washed with column buffer (NaCl [500 mM], Tris-HCl [20 mM] pH 8, Imidazole [30 mM], and Tween-20 [0.2%]) to equilibrate the agarose with 20 mL of buffer. The supernatant fraction of optix was then added to the purification column, and the column was mixed and incubated for 1 hour at 4°C. After incubation, the purification column was opened to allow the supernatant to flow through the column. After the flow-through, the column was washed with the column buffer four times. To elute the protein, an elution buffer was used (NaCl [500mM], Tris-HCl [20 mM], and Imidazole [500 mM]). A total of 6 elutions were done using 2 mL for each one and collected in 2mL collection tubes. Samples were then dialyzed using protein concentrators (Millipore UFC501096) and using an exchange buffer (Tris-HCl [10 mM], NaCl [50 mM], with Glycerol [10 %]). Afterwards the elutions were run on an SDS-PAGE to validate the purification process. Further confirmation of protein production is done with Western blot using anti-his antibodies [1 μL: 10000 μL] (NOVUS NB100-63173).

4.2.3. Electrophoretic mobility shift assay

An Electrophoretic Mobility Shift Assay (EMSA) was done to confirm that optix can bind to DNA. This is done by identifying a DNA sequence that can be recognized by optix or relatable SIX. The **Wnt1** enhancer sequence recognized by Six3 was used for this assay[132]. A 60bp-long DNA probe labeled with an infrared fluorescent dye was design. A fluorescent dye must be incorporated to prepare the ssDNA probe for EMSA. This is done using an already marked ssDNA primer complementary to constant regions of the ssDNA probe. This is done by producing dsDNA in a PCR reaction. The DNA oligos used to prepare our probes with the corresponding fluorescent dye can be found in **(Table 4.2.)**. The DNA oligos were purchased from Integrated DNA Technologies as standard desalted, except when noted otherwise. PCR purification columns are then used to purify the resulting dsDNA probe.

The EMSA is done using a mix consisting of our protein of interest and a fluorescent label DNA probe. A specific binding buffer for optix was determined and named the optix binding buffer, initially in a 10X concentration (250 mM HEPES, 750 mM NaCl, 10 mM EDTA, 100 mM $MgCl_2$). The optix binding buffer is then diluted to a 5X working concentration, and we add Glycerol to a final concentration of 50 %. We also make an additional dilution of the optix binding buffer to a 1X working concentration using the 5X concentration, and this 1X buffer is used for control reactions or protein dilutions. The samples are run in a final volume of 15 μL. The reaction mix is made of (5.1 μL of SIX binding buffer [5X], 4 μL of pdI-DC [1000 ng/μL], 0.5 μL of Bovine Serum Albumin [1 mg/mL], 1 μL of Tween-20 [1%], 0.2 μL of Dithiothreitol [1 M], 0.75 μL of the label DNA probe [1000 nM] and 5 μL of our expressed optix protein from both expression systems. In one control reaction, we don't add protein. A different reaction is done in which no optix is added, and its volume is substituted with SIX binding buffer 1X. This sample will function as the negative control. Each reaction is left at room temperature for an hour of incubation. The reaction is then run on a 6% Acrylamide Native Gel for a 22 cm-high crystal chamber with 0.5X TBE buffer. The gel is left to pre-ran with the TBE buffer for 15 minutes, and samples are added to each well. The samples are left to run for 2 hours and then visualized in a biomolecular imager using the 700 nM wavelength to see the label DNA probes, allowing for the visualization of the DNA on the gel.

Table 4.2. DNA oligos used for EMSA.

DNA oligo name	Oligo Sequences
Wnt1 Promoter	CTTTACTCTCTCCCCAAGGGACATCTAATGATAAGCACAGGACACTTCTGCCCAGGCGAG
Wnt1 Promoter Scrambled	CTTTACTCTCTCCCCAAGGGAATTCAGTCCCGAAAGTAAAGACACTTCTGCCCAGGCGAG
IR700_rPCR_v2 (HPLC purify)	/5IRD700/CTCGCCTGGGCAGAAGTGTC

4.2.4. Systematic evolution of ligands and exponential enrichment (SELEX)

The DNA binding preferences of optix were determined by employing the Systematic Evolution of Ligand and Exponential Enrichment (SELEX) assay. This is an *in vitro* technique used to determine the DNA targets of any DNA-binding protein. To determine the optix binding preference, we used a SELEX variant called SELEX-seq[25]. To start, we use a 60bp long DNA-Library, with a 20bp randomized region flanked by constant regions used to enrich selected sequences and add Illumina compatible barcodes to each sample (See **Table 4.3.** for all DNA oligos used during the SELEX-seq). The DNA oligos were purchased from Integrated DNA Technologies (IDT, Coralville, Iowa, USA) as standard desalted except where noted otherwise. The 20bp randomized region gives 10^{12} unique sequences, which in theory, can provide coverage for each possible 20mer. The ssDNA library is made into a dsDNA library by PCR, and this is done using the rPCR primer (from a 100 μM stock), which makes de dsDNA biotinylated. This PCR reaction is done in two independent PCR reactions. PCR purification columns are used to purify the dsDNA libraries, and concentration is determined by UV/Vis Nanodrop (expecting a working range from 0.5-2 μM).

The binding reaction is done at the same concentrations as the EMSA procedure but in a final volume of 20 μL. In total, three binding reactions are prepared, the dsDNA library (200 nM) + optix, the dsDNA probe + optix, and one only with the dsDNA probe. Since the library cannot be visualized, we used the dsDNA probe as a guide when running and visualizing the gel. Samples are run as described for EMSA, with each dsDNA library reaction run beside their homologous dsDNA probe reaction. After the samples are run, they are visualized with the Azure Sapphire™ Biomolecular Imager (Azure Biosystems Inc., Dublin, California, USA) with laser excitation wavelength 658 nm and filter Red 710BP40. This step allows us to determine if DNA-binding is observable in those lanes with dsDNA probe + optix. The produced image is then printed on a 1:1 scale, allowing the gel and the image to overlap fully. The gel with its back crystal cover is placed over the scaled image. This enables us to locate the samples on the gel using the controls seen in the picture. Using the dsDNA probe + optix lane as a guide, the corresponding lane of dsDNA library + optix. All bound regions and unbound (free DNA) regions are cut from this lane. The cut gels are placed in EB buffer and incubated at room temperature with constant shaking overnight, allowing the DNA oligos to be released into the buffer. The bound sequences are purified using Dynabeads Streptavidin magnetic beads (Thermo Fisher 11205D). Using a magnetic well plate to isolate biotinylated oligos and magnetically selected oligos. DNA oligos corresponding to the bound fraction are washed and enriched (15 cycles) using fPCR and rPCR primers (both from a 10 μM stock) in a PCR reaction. The products are then purified, and their concentration is measured by UV/Vis Nanodrop, and this concentration is used to determine the required volume for another round of selection. This method shows what is considered the first round of SELEX-seq. This process is done two more times for a total of three rounds. Always using the bound fraction purified from the previous step to carry on to subsequent rounds. If any bound sequence resulted from a possible dimeric binding, we treated each bound shift independently during the SELEX-seq. The products of each round, including bound and unbound oligos, were then uniquely barcoded to be sequenced by Illumina sequencing. To barcode, each sample, 5 µL of every DNA sample, including DNA from all three rounds, negative controls (blanks), and the starting library, are run in a PCR reaction independently for each sample. This is done using the rPCR + barcode and the fPCR + adapter (both from a 5µM stock). The sequencing was done using the services of Novogene. Samples were prepared and sent as recommended by the company with the corresponding sequencing primer. After sequencing, de novo motif analysis is done to determine the enriched sequences in our datasets.

Table 4.3. DNA oligos used during SELEX-seq

SELEX-seq DNA oligo	Oligo Sequences
20 library (20N)	CTGATCCTACCATCCGTGCT(N)$_{20}$CACAGCTTCGTACCGAGCGG
Fw_primer (fPCR)	CTGATCCTACCATCCGTGCT
Rv_primer (rPCR) *Adds biotin	CCGCTCGGTACGAAGCTG
Fw_primer + Illumina adapter (fPCR + Ad)	AATGATACGGCGACCACCGAGATCTGCTCTTCCGATCTCTGATCCTACCATCCGTGCT
Rv_primer + Barcode (6p) +Illumina adapter (rPCR+barcode)	CAAGCAGAAGACGGCATACGAGATXXXXXXTCTTCCGATCCCGCTCGGTACGAAGCTGTG
Sequencing primer (HPLC purify)	GCTCTTCCGATCTCTGATCCTACCATCCGTGCT

4.2.5. De novo motif discovery

Enriched sequences are detected using the bioinformatical pipeline described by Nitta et al.[26]. Detection of sequences selected by optix was done with the Autoseed algorithm developed by Nitta and colleagues. This algorithm works by identifying all subsequences that are enriched more than any other relatable sequence. To do this, the program scans the datasets from our SELEX-seq and detects enriched motifs based on a baseline, in this case, the starting library. We used two input files for our analysis: the library file that functions as the background and our optix DNA bound sequences from SELEX-seq rounds. We ran our analysis natively on our computers. The code that we used to determine the enrichment of our sequences is the following:

- **Autoseed running example for optix Round 3 Bound.**

```
./totalautoseed -20N R47_555.txt R47_168.txt 1 8 10 0.35 - 50 40
HeraOptix_R3_168_R47_555.txt;cp Kmer_summary8to10.svg HeraOptix_R3_168_R47_555.tx
t
```

This command was run for each of our proteins and was used to determine the sequences that were bound and enriched by optix. Following analysis, we use Autoseed to determine the binding matrices for each enriched seed we have identified. This is done to represent our analysis in a DNA Logo format to provide information regarding the binding of optix. To determine the binding matrices for our proteins, we used the following code:

- **Autoseed command to determine binding matrices-based sequence seed**

```
./spacek40 --f -dinuc -nocall -m=1 -20N -q R47_555.txt R47_168.txt NYGATACN 200 |
grep "One Hit" > HeliOptix_R47_555_R47_168_NYGATACN.pfm


./spacek40 --logo HeliOptix_NYGATACN.pfm > HeliOptix_R47_555_R47_168_NYGATACN.svg
```

Subsequently, the PWM was determined by Autoseed and was used as input in an R pipeline to better represent the information stored within the matrices.

4.2.6. Homodimer analysis

optix demonstrated dimeric binding in our binding assays and from the results obtained from the De novo motif analysis. The Moder software[149] is used for the discovery of both monomeric and dimeric motifs[149]. To analyze our data, we obtained a randomized sample (100,000 reads) of each sequencing file corresponding to both optix monomer and dimeric sequencing results. To obtain a random sample of 100000 reads we used the following command directly on the computer terminal:

```
sort -R R47_168.txt | head -n 100000 >R47_168_100mil_R3.txt
```

To run the Moder software, we used the running commands described by the developer. The program was run remotely on a computer running UBUNTU, an open-source operating system based on Linux. We ran into some compatibility issues locally and were able to run the program successfully using UBUNTU. An example of the command used in this step is the following:

```
./moder2 --model ppm --names OptixT --outputdir OptixT_100mil_flanks --cob
all data/R47_168_100mil_R3.txt GATAC --number-of-threads 4  --flanks >
results_ R47_168_100mil_R3.txt
```

4.2.7. Genome-wide predictions

Genomic predictions were made using cis-regulatory elements identified from ChIP-seq data from wing tissue on day 3 using antibodies directed towards optix[143]. Scanning was done using the MOODS algorithm[150–152] https://github.com/jhkorhonen/MOODS, which scans the ChIP-seq file for binding sites based on a matrix. The matrix provided in our analysis was obtained from Autoseed and Moder. Each software provides a probability position matrix which is used as one of the inputs used by MOODS. We used multiple matrices in our predictions. From the Moder software, we used four matrices corresponding to optix monomeric binding and three homodimers matrices with different spacing between binding sites, going from 1-3bp between sites. In addition, we also scanned using a matrix obtained from Autoseed, which corresponded to the monomeric binding of optix. An additional binding matrix was obtained using the MEME suite from analyzing enriched motifs within the ChIP-seq data. MOODS was run using the following command for each position frequency matrix (pfm):

```
moods-dna.py -m optix.pfm -s optix_CRE.fa -p 0.0001 > optix_CRE.csv
```

From the data obtained from MOODS, we identified hits to all the cis-regulatory elements (CRE) described previously[143].

4.2.8. Genome-wide predictions sequence validation

To evaluate that optix could bind to predicted CRE regions, we use a genomic browser to visualize the CRE's locations in the *Heliconius erato lativitta* V1 genome scaffold. To visualize the genome, we use the Integrative Genomics Viewer (IGV) software[153] (V.2.13.1) (https://software.broadinstitute.org/software/igv/), last accessed June 16, 2022. Using our MOODS predictions, we located the CRE regions identified by the software in the IGV browser.

To validate if optix could bind to the predicted CRE's hits, we designed EMSA probes using 40bp of genomic sequence. Within those 40bp will be located the predicted optix binding site within the CRE. We also designed control probes (scrambled sequences) and negatives. The scrambled controls are randomized CRE sequences maintaining the same proportion of DNA bases and length. The negatives were designed by going 100bp 5' of the optix binding site and selecting the following 40bp from the Heliconius genome region.

Using EMSA, as previously described, we put to the test if optix can bind to our designed genomic probes. The catalog of genomic probes designed for this step can be found in Table **4.4**.

Table 4.4. DNA oligos of optix genomic targets probes of predicted binding to cis-regulatory elements.

Primer Name	Primer Sequence
optix Dimer 0bp spacer SELEX sequence	CTTTACTCTCTCCCCAAGGGGCATGATACGTATCACTCCTGACACTTCTGCCCAGGCGAG
optix Dimer 1bp spacer SELEX sequence	CTTTACTCTCTCCCCAAGGGACGTGATACAGTATCACTGCGACACTTCTGCCCAGGCGAG
optix Dimer 2bp spacer SELEX sequence	CTTTACTCTCTCCCCAAGGGCCTTGATACATGTATCATATGACACTTCTGCCCAGGCGAG
optix Dimer 3bp spacer SELEX sequence	CTTTACTCTCTCCCCAAGGGGTGATACTTGGTATCAATTGGACACTTCTGCCCAGGCGAG
optix Monomer CRE Prediction	AACAACATAGTAGCGGTGTAATCAACCTTTTATTAATTGCCGTGCCAATGCCGCCGTAAG
optix Dimer 1bp spacer CRE Prediction	CATTTTTTAGAAGAAATATGGTATCAGTTACAGGATTGGCCGTGCCAATGCCGCCGTAAG
optix Dimer 2bp spacer CRE Prediction Hit #1	TACATGAGCTAGGTTGCAGTTAATAACAGTATCAATTAGCCGTGCCAATGCCGCCGTAAG
optix Dimer 2bp spacer CRE Prediction Hit #2	TAATAGGTCATTATAGCCGTATAATAGAATCATTTTTATACGTGCCAATGCCGCCGTAAG
optix Autoseed OptixP CRE Prediction	GGGGGGGCGCGCGCGGTAACCGATAGCCAATCGGGCGCGGCGTGCCAATGCCGCCGTAAG
OptixP Negative N122	GGGGGCCGCGTCCCGCCGCCCTCGTTAAACGCGCCCGCCGCGTGCCAATGCCGCCGTAAG
optix_Scrambled_OptixP	GGAAGGGGCGGGAGGCCACGTGCAGCCTGCCGCCGAGGTGCGTGCCAATGCCGCCGTAAG
optix Autoseed Obs214 CRE Prediction	TAGTAGAACTAGGGGTATCATTTACAAAGGTGCCTTTTGGCGTGCCAATGCCGCCGTAAG
OptixHw Obs132 CRE Prediction	TCATTGCGTGTTAAGTGTCAAAATGAAAATAATTCAATCTCGTGCCAATGCCGCCGTAAG
OptixFw LR1 CRE Prediction	AAGTAGCGAATCGTGACATTTGAAAAGGTAATAAGTAATTCGTGCCAATGCCGCCGTAAG

OptixFw LR1 Negative CRE N100 Prediction	TGCGCCAATTAGTTTGGAAAATCGGACGGTGCCTGTTGAACGTGCCAATGCCGCCGTAAG
OptixFw LR2 CRE Prediction	AAGTAGCGAATCGTGACATTTGAAAAGGTAATAAGTAATTCGTGCCAATGCCGCCGTAAG

4.3. Results

4.3.2. optix overexpression, binding and SELEX-seq Analysis

The expression of optix was done successfully using both expression systems (**Figure 4.1 & Figure 4.2**). Not reported in this work, we previously were successful in cloning and expressing optix in the pTXB1 vector. The expressed nt15-optix was confirmed using a specific anti-optix antibody specific for *Heliconius* optix (**Supplementary Image 6**). The specific anti-optix antibody bound to our insert which confirmed that gene being used is indeed *Heliconius* optix.

Using the Wnt1 promoter, it was noticed that optix demonstrated a double binding shift during EMSA seen in both expression systems (**Figure 4.3**) (**Figure 4.4A**). SELEX-seq was done using cell-free expressed optix. The double-binding shifts were analyzed independently through this work. The autoseed output showed that the heavier band revealed enriched motifs associated with dimeric binding (**Figure 4.4B**). The lower weight band revealed only monomeric bindings with the canonical binding motif of the SIX proteins (TGATAC).

The dimeric binding motifs obtained from Autoseed showed different configurations and spacing between sites while sharing the SIX canonical binding motif. Most homodimers had a Head-to-head configuration with spacings varying from 0-3bp between binding sites. Analysis from the Moder software confirmed the presence of dimeric sites in the SELEX data (**Figure 4.5**). Moder analysis of the higher weight band showed a preference for a dimeric binding site in a Head-to-Head configuration with a spacing of 2bp between binding sites. Having 1bp and 3bp between sites showed the second and third preference in dimeric binding. Data from the dimeric binding file demonstrated a higher presence (0.521) for an optix-optix configuration and a lower association (0.179) for the monomeric binding of optix. Analysis of the monomeric binding file (those coming from the lower weight band) contrasted, with the monomeric association yielding a higher value (0.594) while optix-optix (0.147). Dimeric binding of optix was observed (**Figure 4.6**) and (**Figure 4.7**) when using dsDNA probes from SELEX-seq data.

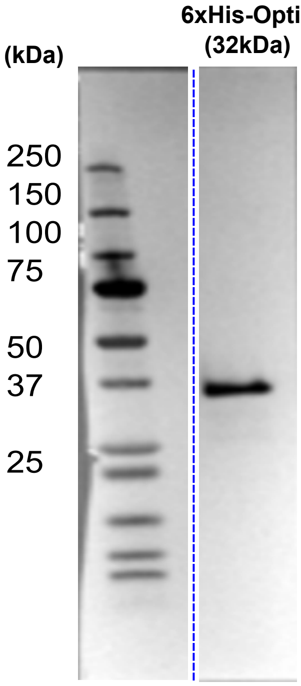


Figure 4.1. Anti-his Western blot of *Heliconius erato*'s optix was expressed in a Cell-Free System: Western blot confirmed the production of optix from *Heliconius erato* when using the wheat germ extract expression system.

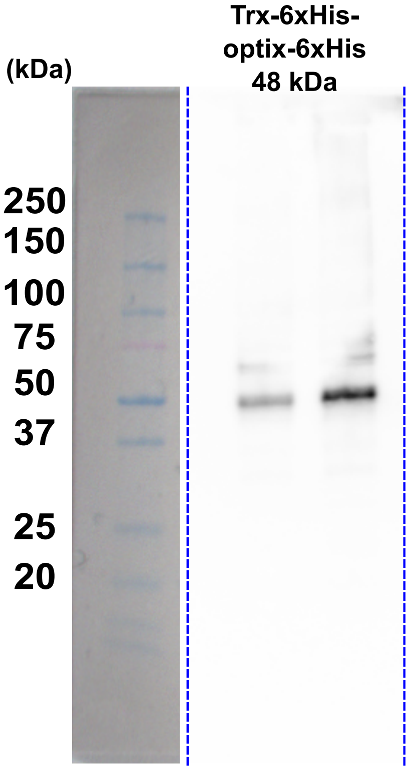


Figure 4.2. Anti-his Western blot of *Heliconius erato*'s optix was expressed and purified from a bacterial expression system: Western blot confirmed the production of optix from *Heliconius erato*, when using BL21 expression system.

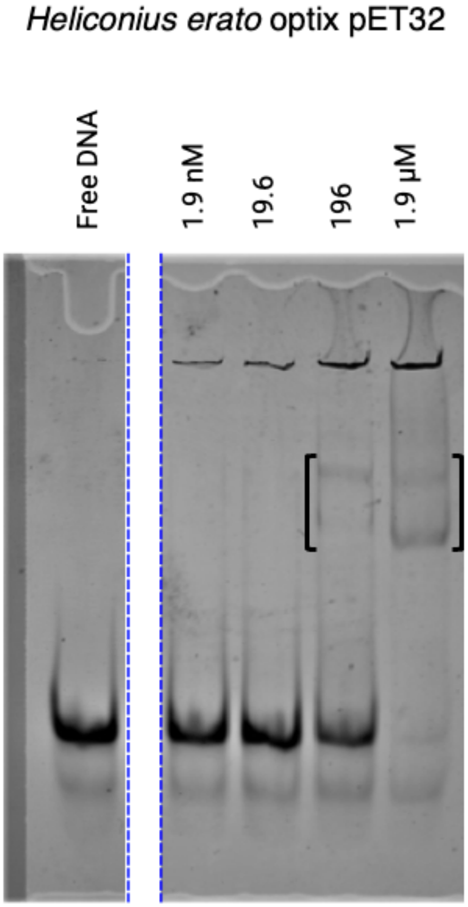


Figure 4.3. optix demonstrates two different weight bands: EMSA done with optix overexpressed from the pET32 vector showed a mix of bands when analyzed with the Wnt1 promoter DNA probe.

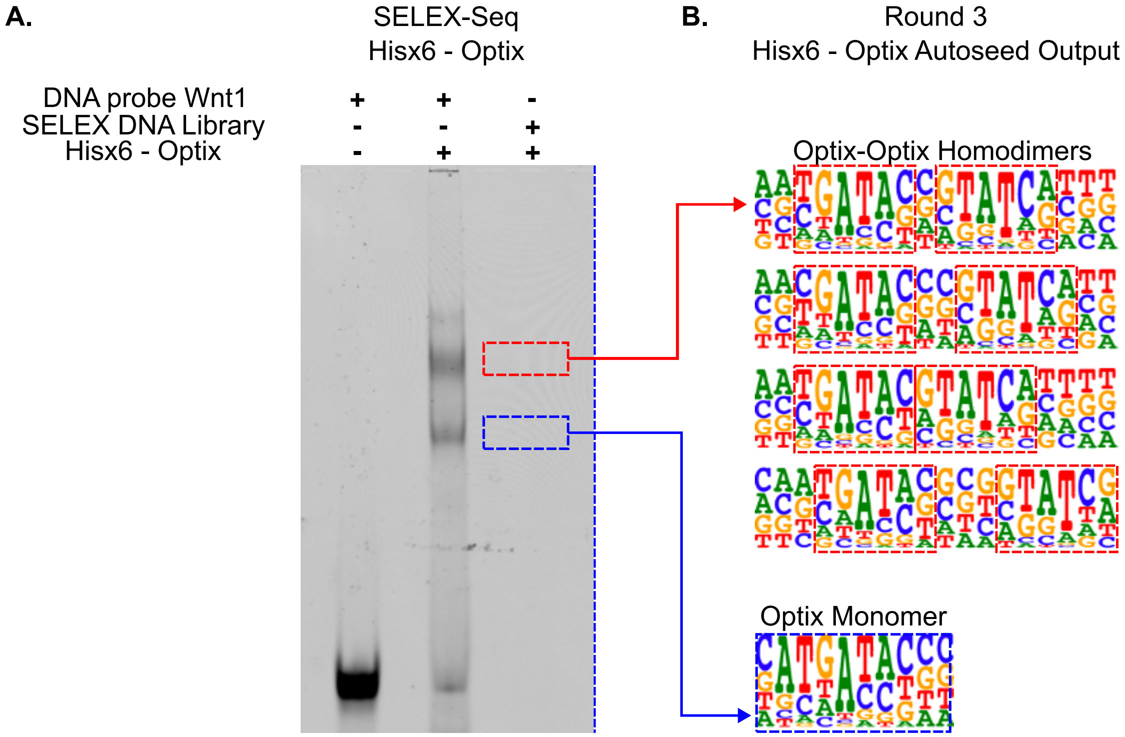
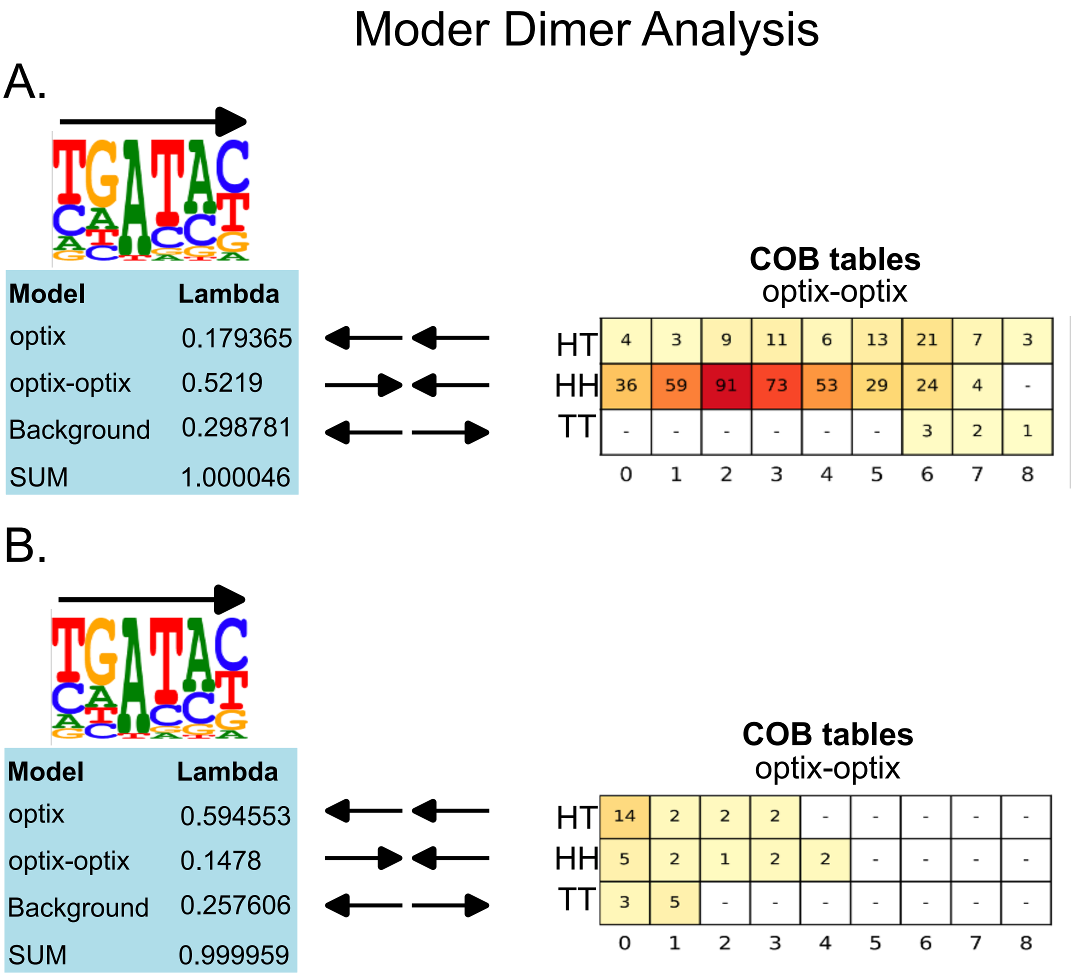


Figure 4.4. optix generates two binding shifts: (A) Double bands were observed with 6xHis-optix expressed on Cell-free. (B) Autoseed analysis shows different enriched motifs based on the band being analyzed. A higher weight band corresponds to dimeric binding, with lower weight showing monomeric binding.



A.

Model	Lambda
optix	0.179365
optix-optix	0.5219
Background	0.298781
SUM	1.000046

COB tables
optix-optix

	0	1	2	3	4	5	6	7	8
HT	4	3	9	11	6	13	21	7	3
HH	36	59	91	73	53	29	24	4	-
TT	-	-	-	-	-	-	3	2	1

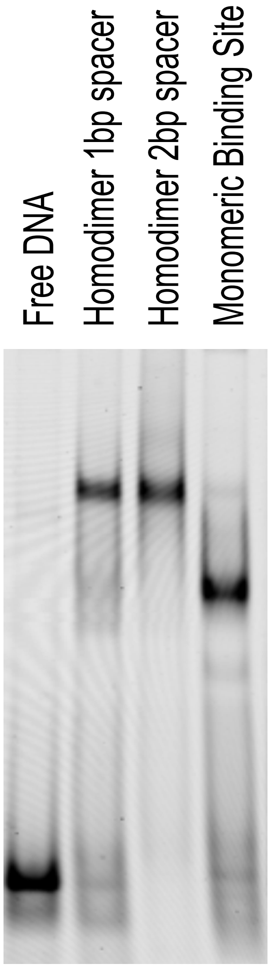
B.

Model	Lambda
optix	0.594553
optix-optix	0.1478
Background	0.257606
SUM	0.999959

COB tables
optix-optix

	0	1	2	3	4	5	6	7	8
HT	14	2	2	2	-	-	-	-	-
HH	5	2	1	2	2	-	-	-	-
TT	3	5	-	-	-	-	-	-	-

Figure 4.5. Moder software analysis showed dimeric association: (A) Dimer analysis showed a higher association for optix-optix configuration with preferred spacing between 1-3bp spacing between sites. (B) Monomeric analysis shows a lack of dimeric binding, highlighting that the higher weighted band corresponds to enriched dimeric sites.



Homodimer 1bp spacer
ACG**TGATAC**A**GTATCA**CTGC

Homodimer 2bp spacer
CCT**TGATAC**AT**GTATCA**TAT

Monomeric Binding Site
CAATACGCA**TGATAC**CTCTG

Figure 4.6. 6xHis-optix can bind as a homodimer. DNA probes with dimer and monomeric binding sites were used based on information from Moder. 6xHis-optix demonstrated a preference for dimeric binding and only binds as a monomer in the absence of a dimeric binding site.

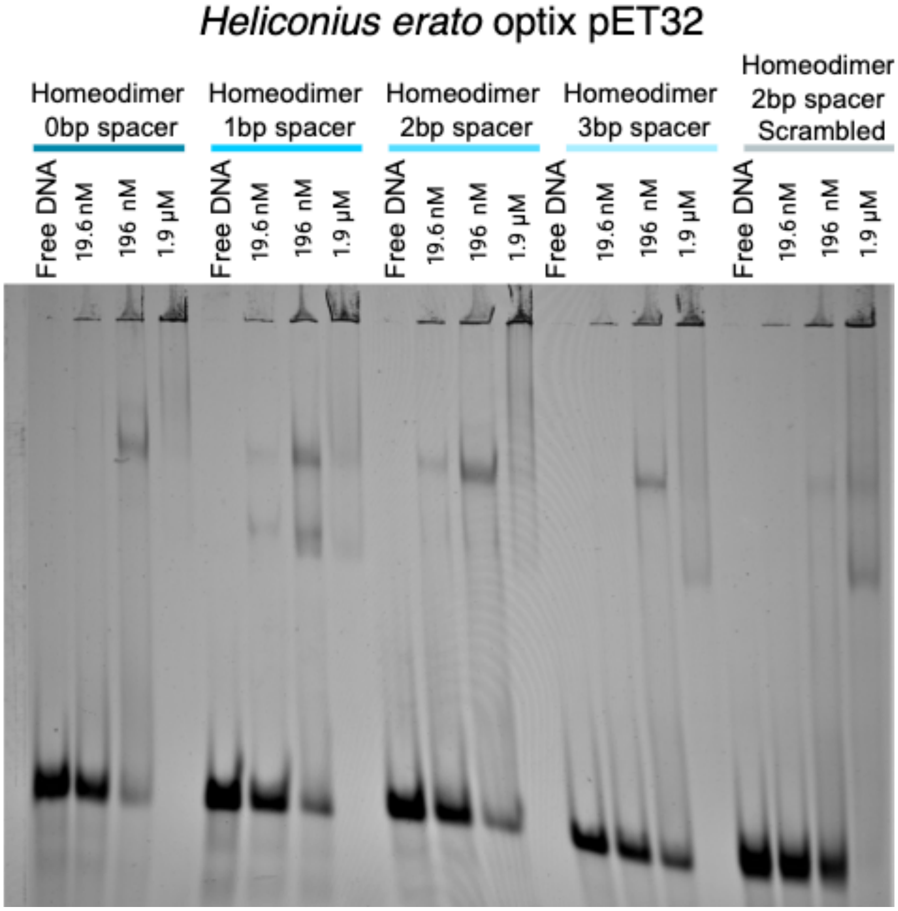


Figure 4.7. Trx-6xHis-optix-6xHis binds in a dimeric form with multiple spacing configurations. The bacterial expressed optix demonstrated similar results to optix from cell-free. Dimeric binding is observed in all Dimeric DNA probes, but some have a mix of binding shifts. As predicted by Moder spacing of 2bp is the preferred spacing in our assay. Both monomeric and dimeric shifts are observable with 1bp between binding sites. Both 0bp and 3bp showed dimeric binding.

4.3.3. Genomic predictions and binding validations

The SELEX-seq data from optix was analyzed by MOODS which allowed us to predict optix binding to (7) cis-regulatory element regions, described in a previous article done with ChIP-seq. Moder homodimer binding matrices return (3) CREs hits (U1, U2 and Obs214). Only dimeric bindings with 1bp and 2bp spacing returned hits within a CREs. For the homodimer binding with 1bp (D1), MOODS predicts binding to the Obs214 CRE. While the dimeric binding with 2bp (D2) resulted in two hits within the 3'UTR of optix (U2). We also used the monomeric matrix from Moder and obtained one prediction within a CRE, especially to another 3' UTR region of optix (U1). Binding was evaluated with EMSA and showed that optix was capable to bind to these cis-regulatory elements of optix (**Figure 4.8.**)

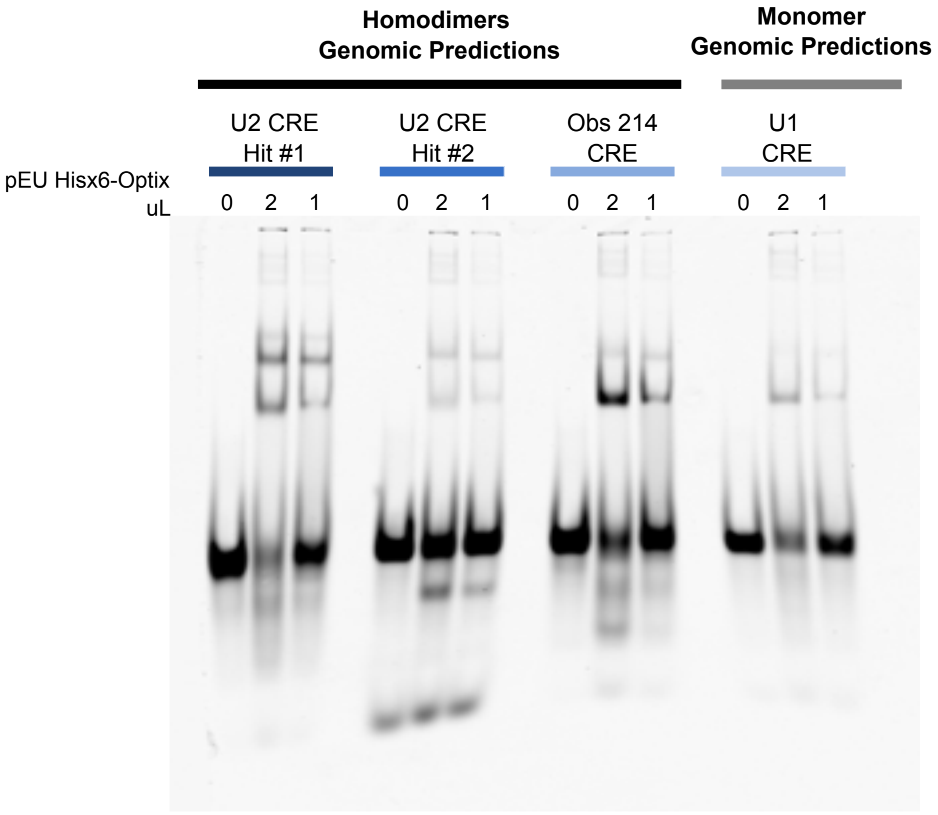


Figure 4.8. 6xHis-optix binds to predicted CRE: Prediction of genomic CRE demonstrates that optix does indeed bind within the evaluated CRE regions. Moder prediction of dimers does demonstrate a mixture of dimeric and monomeric binding in dimeric hits, w those associated with monomeric binding show only one binding shift.

The evaluated CREs demonstrated a mixture of dimeric and monomeric binding. The U2 Hit #1 shows that optix can bind as both a dimer and a monomer to this sequence, the same is observable with U2 Hit #2 but with lesser binding. The Obs214 CRE binding is predicted from a dimeric matrix, however, much of the binding is observable as a monomer. The 3' UTR (U1) monomeric hit does show that optix bind as a monomer to this CRE.

The remaining 4 CREs had corresponding MOODS matches when using matrices from Autoseed. The optix promoter and the Obs214 CREs shows that optix is capable of binding to the predicted sites. Negative control for the optix promoter (negative and scrambled sequences) showed that optix binds to the predicted sequence in a specific manner when using optix from cell-free (**Figure 4.9**). The same CREs were also evaluated with pet32a (+) optix and it was found that optix can bind also with its promoter and with Obs214, however unspecific binding is observable in control lanes (**Figure 4.10**). Interestingly, Obs214 shows a gradual increase in optix binding as concentration with monomeric binding observable at lower concentration while a 1 μM concentration shows a mix of both monomer and dimers. Unspecific binding was also observed with the LR1, LR2, and Obs132 CREs (**Figure 4.11.**). Binding was observable in all the previously mentioned CREs, however unspecific binding was observable in the LR1 Negative DNA probe.

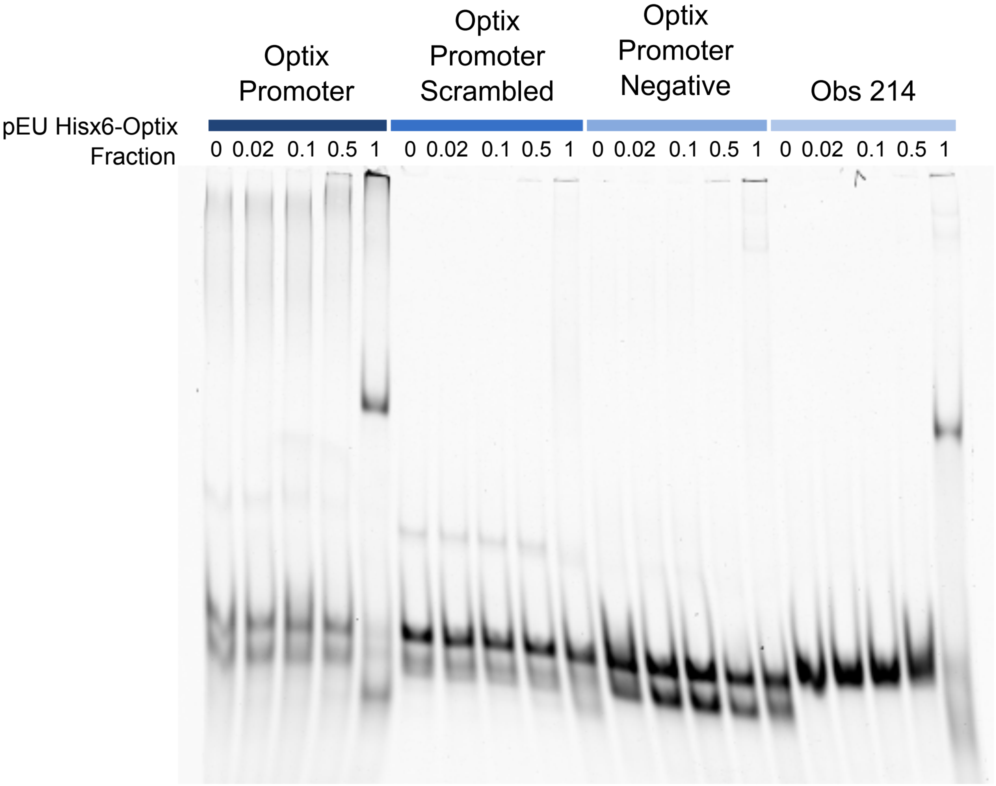


Figure 4.9. 6xHis-optix binds to genomic optix promoter and Obs214 CREs. Specific binding is observable in both the optix promoter and Obs214 CREs. No binding is observable in control lanes. No binding is observable at lower proteins concentration and could be the result of low protein concentration in our original stock.

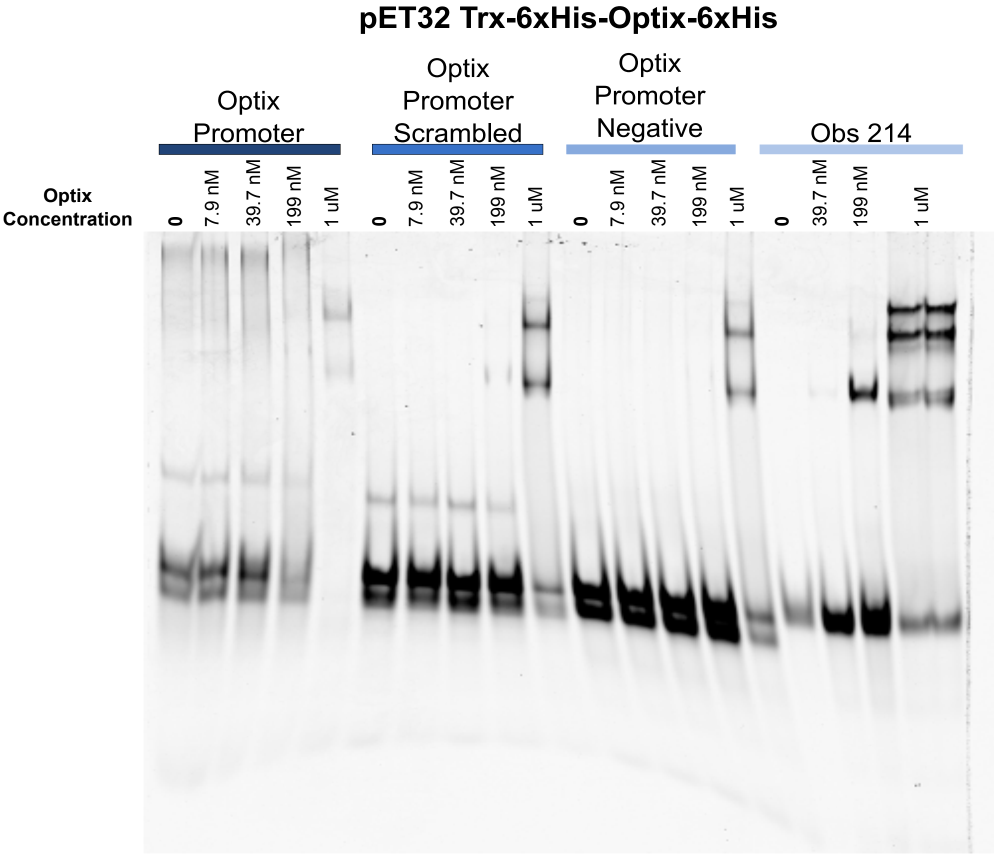


Figure 4.10. Trx-6xHis-optix-6xHis binds unspecifically to control lanes. Binding is observable in the optix promoter but at 1 µM of optix. Less free DNA is observable at 199nM of optix, but no specific shift is observable. Binding is observable in in the Obs 214 CRE from in all lanes where optix is added. Unspecific binding is observable at control lanes (negative & scrambled) this could be the result of too much optix being added.

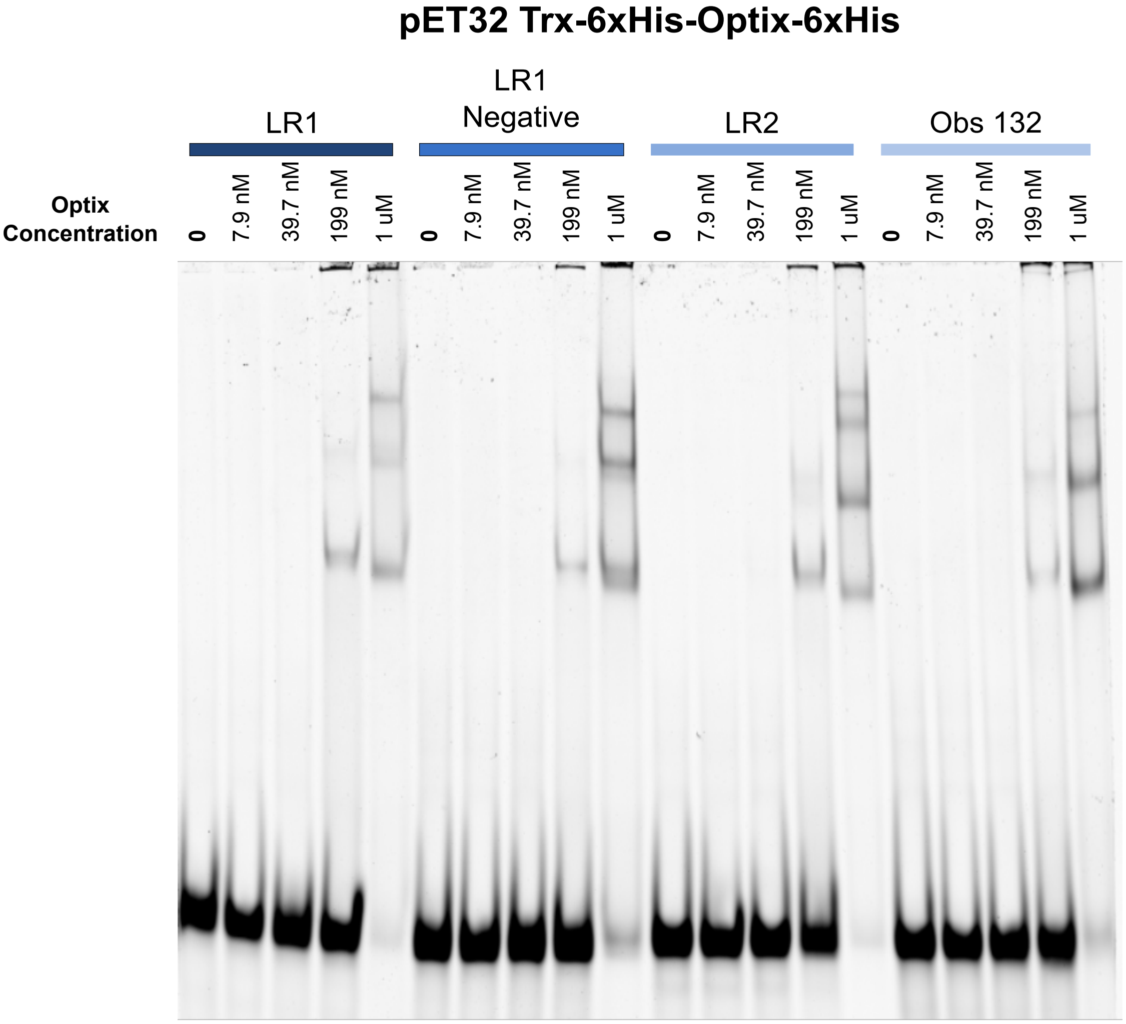


Figure 4.11. Trx-6xHis-optix-6xHis binds to remaining CREs. Binding is observable on all CRE analyzed in this gel. Unspecific binding is observable in the LR1 negative probe. However, binding is similar in all lanes where 1 μM of optix was used. Binding is noticeable at a 199 nM concentration, an uncharacterized optix binding site may be found in the LR1 negative sequence.

4.4. Discussion

In this work, we successfully expressed optix from *Heliconius erato* in multiple expression vectors. The dual expression systems allowed for the development of multiple successful methodologies to express optix both in a cell-free expression system and in bacteria. This is the first reported production of optix from *Heliconius erato* and will allow an opening for more research initiatives. Binding results from EMSA done to optix from both expression systems were the same. When using the Wnt1 promoter, both optix's successfully bind to the sequence, a shared trait with optix paralog SIX3. What made our binding assays interesting was the presence of a double binding shift when using the Wnt1 promoter with both optix samples. These suggested the possibility that optix could be binding as a homodimer. Since the double binding was consistent in multiple EMSA replicates, on SELEX-seq, both binding shifts were analyzed independently. Analyzing both bands independently allows for motifs to enrich based on their binding configurations.

The Autoseed program was used to evaluate each band by comparing the sequencing file from each sample with the starting library. Both samples shared the same core binding motif of the SIX protein family `TGATAC`. When comparing both binding shifts, the main difference was based on the presence of a homodimeric binding site associated with the heavier-weight band shift (**Figure 4.4A**). The homodimeric sequences were found enriched in a head-to-head configuration with multiple spacing between the optix binding sites (**Figure 4.4B**). Based on this observation, we wanted to determine if our results were also replicable within a software specialized in analyzing dimers. In this case, we decided to use the Moder software. Again, both shifts were analyzed independently to compare the differences between shifts. We analyzed a subset of the sequencing data using a random sample of 100,000 sequences from each file. These subsets were necessary since using all the sequencing data was not possible with Moder. Our results showed that optix was binding as a homodimer, with a spacing preference of 2bp between optix binding sites. In addition, it found a preference from 0 to 3bp between sites (**Figure 4.5A**). It was also shown that the preferred dimeric configuration was indeed head-to-head. The Moder analysis of the lower weight binding shift was associated with monomeric binding. These results gave weight to the Autoseed analysis done on both samples and gave confidence to both results.

To confirm these observations, dsDNA probes were designed using sequences found in our SELEX data that follow the dimeric spacing described in Moder. This was done to evaluate the binding preference of the homodimer configurations reported by Moder. To test dimeric binding, an EMSA was done with a small subset of probes, only considering spacing from 1 to 2bp and a monomeric dsDNA probe. Using these binding configurations, it was observed that optix indeed bound to these sequences accordingly to the presence of a dimeric or monomeric binding site (**Figure 4.6**). The dsDNA probes with homodimeric binding sites showed that functionally optix did indeed bind to dimeric bindings sites. However, the dimeric binding observed was not the same for the tested probes. When comparing 2bp and 1bp spacing, the preference for a 2bp spacing was noticeable. Confirming the results from Modern, where 2bp had a greater representation in our data. The 1bp spacer showed most of the binding is found in the dimeric form; however, some binding is observable at the lower-weight monomeric binding. A single binding shift was visible when using the monomeric binding DNA probe.

Using optix produced from bacterial overexpression, we analyzed the dimeric probes having a 0-3bp spacing between optix binding sites. This binding assay showed that homodimers with 0, 2, and a 3bp spacer between sites demonstrate only dimeric binding when using 194 nM of optix protein (**Figure 4.7**). Higher concentrations of optix bind to all the DNA probes with no free DNA is observable, this is expected when using TF at higher concentrations. For example, TF when in the nucleus are bound to DNA. Higher protein concentrations can lead to unspecific binding in EMSA assays. This is observable in the control scramble sequence used for the gel, which shows binding at 1 μM of optix, it is also notable that there's no observable free DNA on 1μM optix lanes. When using 194 nM of optix we were able to observe similar results between optix from both expression systems. Results are as expected based on our Moder results, again noticing that 1bp between binding sites leads to a mix of monomeric and homodimeric binding. These observations show that optix has spacing preferences when binding as a homodimer.

The SIX transcription factor family is known to participate in protein-protein interaction, for example, optix has been known to interact with the corepressor Groucho[56], while other SIX proteins are known to interact with other cofactors like Eya or Eyeless[55,56,66,78] and even between some family members SIX1:SIX4[61,83,100]. No homodimeric binding has been reported for optix before, and it will seem to be a trait found only in *Heliconius* optix. Based on these observations, it will appear that optix in the *Heliconius* genus has evolutionarily allowed optix to make homodimers.

Dimeric association by a TF allows an additional layer of genomic regulation in which some genes can be expressed or repressed based on the nature of the binding. For example, it has been reported that Gsx2/Ind homeodomain, a TF family of which the SIX are members, can bind as a homodimer. What is interesting about this type of binding is that the genomic role of the TF can depend on if it binds as a monomer or homodimer. It was reported that if Gsx binds as a monomer, it functions as a repressor, while binding as a homodimer carries an activator role[154]. Cooperative binding is also possible when one protein promotes another's binding to its ligand. In the case of FOXL2 and FOXA1, the binding of one factor promotes changes within the DNA structure, in which the minor groove becomes narrower and promotes the binding of the secondary protein[155]. Both examples are ways cooperative binding can promote the binding of another protein, in this case, a TF, and allow for the expression or repression of a target gene. The current role of optix homodimer in the Heliconius erato is currently not known, but it is possible that similar mechanisms, like the ones reported, can occurring with optix. This trait could play potential roles in how color pattern phenotype develops based on the interactions between optix and the DNA.

The regulatory roles of optix are based on the cis-regulatory elements (CRE)s to which it binds. Using data from our SELEX-seq analysis, we successfully predicted optix binding to previously described cis-regulatory elements in developing wings from previously reported CREs[143]. The MOODS predictions identified binding to 7 CREs. The predictions were made using dimeric and monomeric position frequency matrixes, and our EMSA validations showed that optix could bind to our predicted binding regions.

When using Moder position frequency matrices (**Figure 4.8**), two hits were predicted within the 3'UTR2 of optix, one hit to the Obs 214 CRE, and one hit for the 3'UTR1 of optix. The 3'UTR2 first hit prediction showed that optix could bind as a homodimer, but we could also observe monomeric binding. The second prediction shows the same binding configuration, but this set of predictions showed a weaker binding overall. Interestingly the binding to the Obs 214 CRE showed a monomeric binding preference. All previous predictions were done using Moder homodimeric matrices. Lastly, our prediction in 3'UTR1 of optix was made using a monomeric matrix, and we obtained the predicted monomeric binding.

To explain why some predictions bind in a mixture of homodimer and monomer depends on the DNA's characteristic and the binding's nature. Flanking regions near the binding sites can play a role in the mechanism that the protein binds to DNA. These predicted regions may have suboptimal flanking that doesn't fully promote homodimer binding. In addition, it is possible that it could depend on other factors not taken into consideration, DNA looping or other markers that could promote either dimeric or monomeric binding. The Obs214 is interesting since we expected a preference for the homodimeric binding; however, we obtained a monomeric preference; we suspect that it could also be based on the flanking regions near the core binding motif. It is also possible that these regions are bound as a dimer in higher optix concentration.

The remaining CREs were predicted using nondimeric position frequency matrices from Autoseed data. One key prediction was that optix could bind within its promoter (**Figure 4.9**) and with Obs 214. For optix to bind to its promoter could mean an autoregulatory feature of optix, something already described for sine oculis, another SIX protein[56]. Previous observations were done using optix from expressed on cell-free, and during this process, no more optix could be produced using the system. Redoing the same CREs EMSA gel (**Figure 4.10**) using optix expressed in bacteria shows similar results but with unspecific binding being observed. Again, these unspecific binding are only observable in high optix concentrations (1 μM) and could explain the observed unspecific binding on both control probes. The Obs214 results are interesting, binding is observable at 39.7 nM optix concentration with an evident monomeric binding at 199nM with observable homodimeric binding, which could provide some insight into how optix concentration can contribute to optix binding configurations.

The remaining CREs (LR1, LR2, and Obs 132) were only validated using bacterially expressed optix (**Figure 4.11**). At lower protein concentrations, LR1, LR2, and Obs 132 showed optix binding with only monomeric binding, with a faint dimeric binding observable in LR1 and LR2. High optix concentration shows multiple binding shifts, a trait noticeable in the remaining CREs. We did notice binding in the LR1 negative probe, further analysis of the genomic region used, and we notice the presence of a (`TAAT`), the canonical binding site of the Homeodomains. This binding site is not associated with the SIX however, Nitta et al. (2015) [26] SELEX work showed that *Drosophila's* optix could bind to the (`TAAT`) sequence. It is plausible that in the absence of the canonical SIX binding motifs, optix can bind to the ancestral homeodomain binding site with lower specificity. What role can this possible alternative binding is an interesting future endeavor. For the remaining CRE LR2 and Obs132, we observe the same binding patterns previously described for LR1.

4.5. Conclusion

The *Heliconius* butterflies' diversity is not limited to their diverse color palette or the way they use their colors in multiple patterns and arrangements but also to what makes their regulation mechanism unique. optix has been extensively studied in *Drosophila*, but the way it is used in a novel form in *Heliconius* shows that they are not alike. We were able to observe that optix from *Heliconius* can bind DNA as a homodimer, a trait that has not been reported in *Drosophila*. This unique characteristic of optix could provide an example of how these color patterns are achieved since it could demonstrate a layer of gene regulation only found in *Heliconius*. My observations were validated with numerous software analyses which showed that optix can bind a homodimer with binding assays validating our predictions. *In vivo* data from the developing wing allowed us to confirm that our *in vitro* data could predict optix binding to genomic targets. The validation of our data opens the possibility for new studies into what genes are regulated by optix and to be able to fully understand the red color diversity of the *Heliconius* butterflies.

Chapter 5: Significance and Future Directions

5.1. Significance

In this thesis, we have explored multiple aspects of the SIX transcription factor family. First, it was established that this family can be traced to the simplest of animals, the sea sponges, where at least an uncharacterized SIX protein gave origin to sine oculis, six4, and optix. These three proteins have been in animals' genomes since 810MYA.

This work also explored if changes in specificity could explain the diverse phenotypic variations associated with SIX proteins. When comparing SIX orthologs in *Drosophila melanogaster* and *Heliconius erato* with their respective paralogs in *Homo sapiens*, we can observe the presence of at least five functional classes based on their specificity. These results highlight that the differences between SIX members make each protein an unique target for research. This uniqueness was observable from the presence of a 5'-GA extension of the binding motif, being the extension shorter in optix.

The *Heliconius erato* optix demonstrated that it could bind to DNA as a homodimer, a trait not previously reported for optix in any species. The capability of optix to bind as a homodimer can provide another layer of regulation that could be how the red color patterns on the wings are regulated. Using the data from SELEX-seq, it was also predicted that optix could bind to previously reported cis-regulatory elements and binding assays confirm our predictions.

The current work paves the way for new scientific endeavors that can provide new insights into how TF evolve by changing their specificity. We also determined the rich evolutionary history of this family, and as sequencing is done to more species, a more precise view will be developed with our observations. This work has also provided insight into the unique trait of optix from *Heliconius* to bind as a homodimer while validating previously mentioned cis-regulatory elements.

How optix regulates red color pigmentation is still open for further studies. Still, this work provides a gateway for future stories by being the first to express the protein in two expression systems.

5.2. Future directions

This work raises the potential for numerous projects and endeavors to give answers to the countless new questions that have arisen. The SIX are a complex family of TF, and there is still much information needed to be unraveled to understand how they function and what makes them so important in development. The current work has provided some answers to previously established questions but has also given origin to new questions that will be enhanced by future work on the topic.

In Chapter 2, we were able to determine the evolutionary history of the SIX family but even with the quality of our information, much work is needed. The low branch of our tree needs more information to understand the dynamic observed in sponges. How many proto-SIX proteins were really on sponges? To answer this, more sequencing will be needed in Porifera to have a more precise idea of the dynamics and establishment of the SIX lineages. We were also unable to detect any SIX proteins in unicellular organisms, are the SIX only found in multicellular organisms? In addition, some proteins appear to have been lost in the evolution on some species. Did redundancy lead to the deletion, or the current genome assemblies have been unable to identify the missing protein? Furthermore, *C. elegans* two SIX proteins are not identified with any of the other SIX proteins; it will be interesting to study what sets them apart from the canonical SIX proteins. This could be done using the method I have employ in this work (SELEX-seq), this could provide useful information if there are traits exclusive for this nematode.

The SIX's binding motif has long been established to be a `TGATAC` binding motif; however, in our work (Chapter 3), we were able to observe that the motif is more extended. A 5'GA flanking region is observed in all SIX proteins, with some degrees of difference. We could even observe that some SIX proteins could not bind in the absence of this flank, a trait only previously reported in sine oculis. How the flanking region contributes to DNA-binding is an interesting question that could be answered by determining the protein arrangement when interacting with DNA. The atypical SIX HD must make the protein adopt a novel structure to bind to DNA properly, is the extended motif the result of this? Why are some SIX proteins capable of binding without the flanking sequence? The structure of the protein, which at this moment hasn't been determined for a SIX protein interacting with DNA, will provide answers to these lingering questions. Although some predictions could be done using the new structure prediction models like Alphafold, doing X-Ray crystallography to at least one SIX TF will provide information that is currently lacking, especially if this structure is done with the protein interacting with DNA. Likewise, it will be interesting to determine the DNA-binding specificity of "ancestral" SIX proteins, like those from sponges, jellyfish, and comb jellies. This is achievable since I have already found the sequences and cloning, and protein expression are already optimized. Doing the "ancestral" SIX proteins allows to establish a specificity timestamp for this TF, that can be used to compare with other members. Is the binding specificity established since early SIX proteins, or is it the result of millions of years of evolution?

Finally, the *Heliconius* butterflies are an excellent model for studying the evolution of color patterns and the underlying regulatory process that achieves it. The master regulator optix is known to be responsible for regulating the red coloring process, probably by using the ommochrome pathway seen in fly eyes. It would be interesting if, using our binding matrices, it could be proven that, indeed, optix does bind to components of the ommochrome pathway. Does optix achieve this by suppressing other pigments to be expressed? The first could be explained by directly identifying these genes on the *Heliconius* genome and using our binding matrices to do predictions. Reporter assays can be done however in *Heliconius* it will be more useful to determine the homologous genes in Drosophila and compare results between both datasets. Could we, in the future, be able to use optix as a genomic paintbrush to manipulate colors in the future? Another interesting question will be what is the function of the homodimer during genomic regulation? Could monomeric or dimeric binding function as repressors or activators, or is there an underlying cooperative interaction between optix and other factors? Being able to express optix *in vitro* can provide new pathways previously not attainable.

Supplementary Images

Supplementary Image 1: SIX proteins multiple sequence alignment

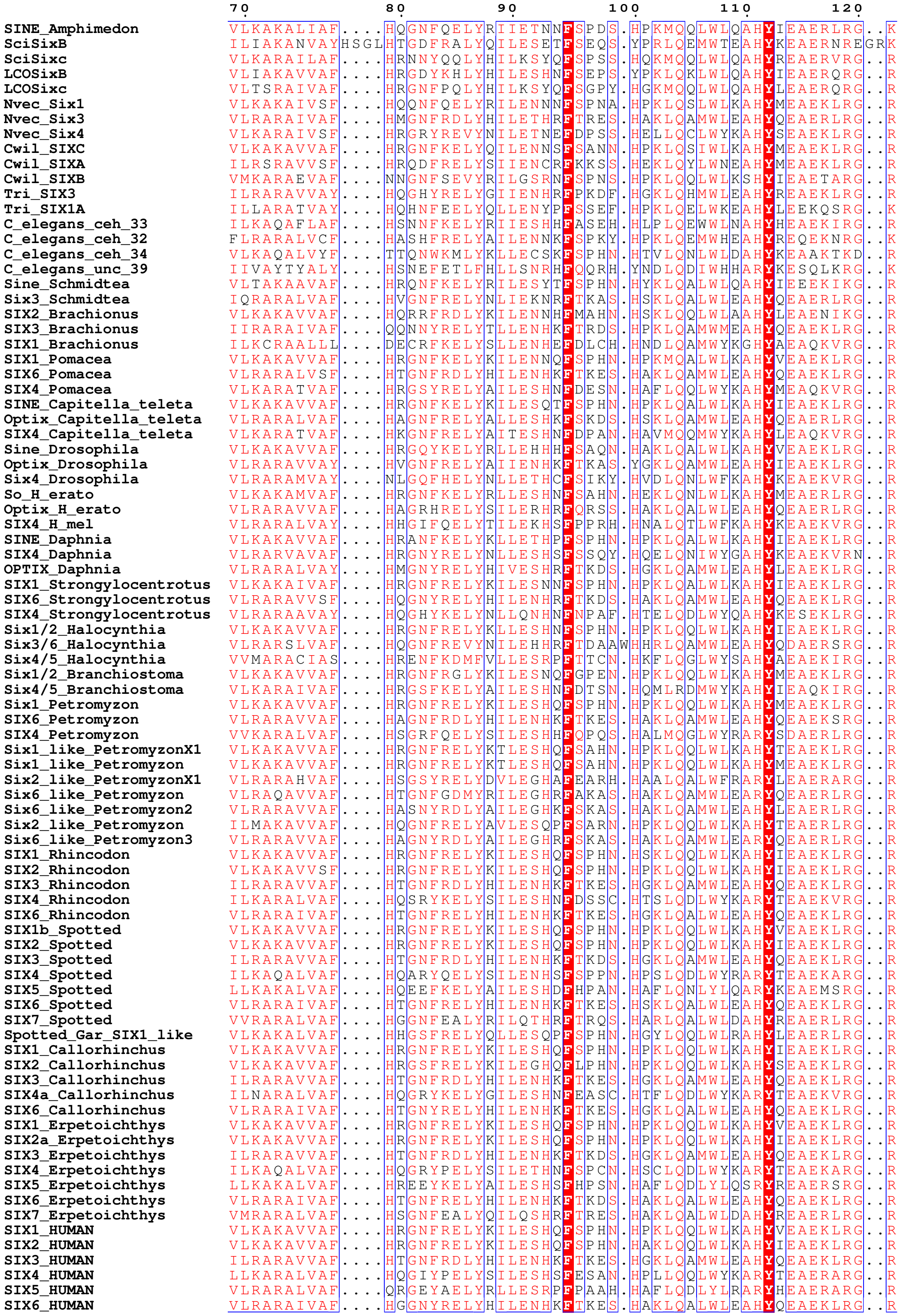
```
SINE_Amphimedon             ............................................................
SciSixB                     ....................................MSRDRGRNSHLPLPSASVIASPPR
SciSixc                     ..................................MKTANFTSPSAVRDQRLSSSSASNVE
LCOSixB                     ............................................................
LCOSixc                     ............................................PVTTVGGGGGTALALL
Nvec_Six1                   ............................................................
Nvec_Six3                   ............................................................
Nvec_Six4                   ............................................................
Cwil_SIXC                   .................................................MFEATSTVKLP
Cwil_SIXA                   ............................................................
Cwil_SIXB                   ............................................................
Tri_SIX3                    ...................................................MYQLRHAPY
Tri_SIX1A                   ............................................................
C_elegans_ceh_33            ............................................................
C_elegans_ceh_32            ............................................................
C_elegans_ceh_34            ............................................................
C_elegans_unc_39            ........................................................MTDH
Sine_Schmidtea              ............................................................
Six3_Schmidtea              ............................................................
SIX2_Brachionus             ............................................................
SIX3_Brachionus             ............................................................
SIX1_Brachionus             ............................................................
SIX1_Pomacea                ............................................................
SIX6_Pomacea                ............................................................
SIX4_Pomacea                .....................................................MDNNGRP
SINE_Capitella_teleta       ............................................................
Optix_Capitella_teleta      ............................................................
SIX4_Capitella_teleta       ..........................................................MD
Sine_Drosophila             .............................................MLQHPATDFYDLAAA
Optix_Drosophila            ............................................................
Six4_Drosophila             .......................................MFDKNLDGNNLSVSIGGDLDS
So_H_erato                  ............................................................
Optix_H_erato               ............................................................
SIX4_H_mel                  ...........................................................M
SINE_Daphnia                ..............................................MIIGPSAAGGMSDL
SIX4_Daphnia                MASGSLSSGASLLHPFGAGSSVNETTSTGSVSPPPSGHHTLEPATGNHPLLSHRLVTKSE
OPTIX_Daphnia               ............................................................
SIX1_Strongylocentrotus     ............................................................
SIX6_Strongylocentrotus     ............................................................
SIX4_Strongylocentrotus     ............................................................
Six1/2_Halocynthia          ............................................................
Six3/6_Halocynthia          ............................................................
Six4/5_Halocynthia          ..................................MSSELSHEARVTPTKSKNGTELESCG
Six1/2_Branchiostoma        ............................................................
Six4/5_Branchiostoma        ............................................................
Six1_Petromyzon             ............................................................
SIX6_Petromyzon             ............................................................
SIX4_Petromyzon             ............................................................
Six1_like_PetromyzonX1      ............................................................
Six1_like_Petromyzon        ............................................................
Six2_like_PetromyzonX1      ............................................................
Six6_like_Petromyzon        ............................................................
Six6_like_Petromyzon2       ............................................................
Six2_like_Petromyzon        ............................................................
Six6_like_Petromyzon3       ............................................................
SIX1_Rhincodon              ............................................................
SIX2_Rhincodon              ............................................................
SIX3_Rhincodon              ............................................................
SIX4_Rhincodon              ............................................................
SIX6_Rhincodon              ............................................................
SIX1b_Spotted               ............................................................
SIX2_Spotted                ............................................................
SIX3_Spotted                ............................................................
SIX4_Spotted                ............................................................
SIX5_Spotted                ............................................................
SIX6_Spotted                ............................................................
SIX7_Spotted                ............................................................
Spotted_Gar_SIX1_like       ............................................................
SIX1_Callorhinchus          ............................................................
SIX2_Callorhinchus          ............................................................
SIX3_Callorhinchus          ............................................................
SIX4a_Callorhinchus         ............................................................
SIX6_Callorhinchus          ............................................................
SIX1_Erpetoichthys          ............................................................
SIX2a_Erpetoichthys         ............................................................
SIX3_Erpetoichthys          ............................................................
SIX4_Erpetoichthys          ............................................................
SIX5_Erpetoichthys          ............................................................
SIX6_Erpetoichthys          ............................................................
SIX7_Erpetoichthys          ............................................................
SIX1_HUMAN                  ............................................................
SIX2_HUMAN                  ............................................................
SIX3_HUMAN                  ............................................................
SIX4_HUMAN                  .......................................MSSSSPTGQIASAADIKQENG
SIX5_HUMAN                  ............................................................
SIX6_HUMAN                  ............................................................
```

```
SINE_Amphimedon               ............................................................
SciSixB                       PPPLPALSTQATDPDESRTYQPAFPLPPTSQQQQQEEKELCTPLAVALGDSGQQLDLYRC
SciSixc                       YFGASPALHAALEAERIAAYEAANPMPLPMSSNASSQSAGTYLPPVESTEPNYLAHQAAQ
LCOSixB                       ......................QYPIQHVAQNQLVQTAASPRISAPVTVVASTAVHGNLS
LCOSixc                       AAQEAERIASFEVANTPPLPLPLASRTAHAVQTFQPPSLTSLTRSVAAETDSHYLALRAA
Nvec_Six1                     ............................................................
Nvec_Six3                     ............................................................
Nvec_Six4                     ............................................................
Cwil_SIXC                     STTVTMETTQPMELSPRHHHVMTGTSSFYRPPSPTTTDHRQGARSPGPSYRHHGIPSSPL
Cwil_SIXA                     ................................................MDFGVKIEQSSI
Cwil_SIXB                     .......RKDESPPSPPSPPIPSDPQRAHHLRQGGGAGGGGAPTPPPPPLRPVIVGEDEE
Tri_SIX3                      VKPAPTINGAIPIIAGMPASPHHQALATANLFRAPTMVPAVAPLPPHNVALAALASGMAQ
Tri_SIX1A                     ....................MAYLPYMYCYPEHRHSNSNLILNYNYLSGFQPSKQPYPKQ
C_elegans_ceh_33              ............................................................
C_elegans_ceh_32              ........................MFTPEQFTKVMSQLGNFSQLGQMFQPGNVAMLQALQ
C_elegans_ceh_34              ............................................................
C_elegans_unc_39              PPIDTSSYFDCYQQHQLPLQYTFTSSSNSNTSNSSTSPSHISDQFSSSGGPPYELSSHIL
Sine_Schmidtea                ......MRSISTSQATDIQQNICDQFSNSGDYKSLDGPVVSSPSESDSNVNYASFHQFGM
Six3_Schmidtea                ............................................................
SIX2_Brachionus               ...................................MYEPVSNKTQNLLIMTSSSQNSTGP
SIX3_Brachionus               ............................................................
SIX1_Brachionus               .....MDYHHAQVSTEGRGAPLCYEAPYPDPFSHFAGGYQPPQFYYGNYALPHDTESNED
SIX1_Pomacea                  ............................................................
SIX6_Pomacea                  ..MDLSRTSKLGPVALSTLVGSQPQHPPPPPPPPVLAQPHAHLPAITAAAAAAAMYPGLP
SIX4_Pomacea                  PSARGCMAQASDDDTLTLGDDVLGASLLHGDNSHLDTTVNSNNSSNGGALNATLSNSSSS
SINE_Capitella_teleta         .......................................................MLHEQ
Optix_Capitella_teleta        ....................................MEKVSPSRPKPIFPPFSPGVFHGA
SIX4_Capitella_teleta         LGTSNFVDCRRAASDCRLMQDENMEPLNDLVKSEDSYDATADEIENISSFKSPPQSSSSS
Sine_Drosophila               NAAAVLTARHTPPYSPTGLSGSVALHNNNNNNSSTSNNNNSTLDIMAHNGGGAGGGLHLN
Optix_Drosophila              ...................................................MAVGPTEGK
Six4_Drosophila               TSSGGTSSDHSAVHQDNLSSPMAYGSLFLPNAGYRGNLSCKTVLQLDKFAPYEGVEKDHL
So_H_erato                    ............................................................
Optix_H_erato                 ...............................MRGSWDESTTAALHARGPAAERAAEPAAC
SIX4_H_mel                    ESCSDRLSPSSDMTSESETSLPSFPYEQNNFYTQSDINEKQYFSCKQKSPRTDERKDYLQ
SINE_Daphnia                  VAHHHHMSGFAAHHHHHHMGLGGMMAAGMMPGMSSTSTGMTVQPTSQSPPSSNNNNNNNN
SIX4_Daphnia                  RHHQQHRLYQPYHLPLHPHHHIQHHHHLQQQPQHMENQYDLNQHYPVASTSAVGGTDSRC
OPTIX_Daphnia                 ..............MALAASSSATSVTTAGPRSSAATPPDQPTFPLMPTPLFAGAGNGGG
SIX1_Strongylocentrotus       ............................................................
SIX6_Strongylocentrotus       ..............MVWTETETSPSDRGAMRAAQALAQYYSLHDPRSSEGLARLAHMYTS
SIX4_Strongylocentrotus       ..MEAAPTPLETVGSELDTAMSVGESDINLNSAGDTSSTTDSNAGISLPVPTSLLSNVEN
Six1/2_Halocynthia            .....................MATALAVSSYTTVLPQVVQHHSAAAAAAAAAAAAQLSAV
Six3/6_Halocynthia            ............................................MFPKQYPGLGQDLSAS
Six4/5_Halocynthia            GFLPSVAAVISPEDETNSRAVPPTSSLCNRLENNTSPQDLASLANMLTIDKQVSYNISNN
Six1/2_Branchiostoma          ............................................................
Six4/5_Branchiostoma          .............................................MANAVENVNPNVNQG
Six1_Petromyzon               ............................................................
SIX6_Petromyzon               ............................................................
SIX4_Petromyzon               ....................................MSCSTQVAGRAPQVPVAPALLQSP
Six1_like_PetromyzonX1        ............................................................
Six1_like_Petromyzon          ............................................................
Six2_like_PetromyzonX1        ...........................................................M
Six6_like_Petromyzon          ............................................................
Six6_like_Petromyzon2         ............................................................
Six2_like_Petromyzon          ............................................................
Six6_like_Petromyzon3         ............................................................
SIX1_Rhincodon                ............................................................
SIX2_Rhincodon                ............................................................
SIX3_Rhincodon                ..........................................MVFRSQFELYSALPLLPN
SIX4_Rhincodon                ......................MSSASNEITNTIEIKEENDSLQPKINNTLGSPATLEPE
SIX6_Rhincodon                ............................................................
SIX1b_Spotted                 ........................................................MSFW
SIX2_Spotted                  ............................................................
SIX3_Spotted                  .....................................HSMVFRSPLELYPSHFFLPNFAD
SIX4_Spotted                  .....................................................LALDAAG
SIX5_Spotted                  .....................................MASFSLEAGVQAESPSEVSVQDS
SIX6_Spotted                  ...........................................................G
SIX7_Spotted                  .......................................QIYHPKLKTHPTCCFCANLRG
Spotted_Gar_SIX1_like         ............................................................
SIX1_Callorhinchus            ............................................................
SIX2_Callorhinchus            ............................................................
SIX3_Callorhinchus            .........................................MVFRSQLELYSAALPFLPN
SIX4a_Callorhinchus           ....................................................MVVEPEVS
SIX6_Callorhinchus            ............................................................
SIX1_Erpetoichthys            ............................................................
SIX2a_Erpetoichthys           ............................................................
SIX3_Erpetoichthys            .......................................MVFRSPLELYPSHFFLPNFAD
SIX4_Erpetoichthys            ...............................MSSSSNEVRSAGEIKKENVSSGEEPELLA
SIX5_Erpetoichthys            .................................MASLSLEAGTPKESPGETSSGELSSPS
SIX6_Erpetoichthys            ............................................................
SIX7_Erpetoichthys            ............................................................
SIX1_HUMAN                    ............................................................
SIX2_HUMAN                    ............................................................
SIX3_HUMAN                    .MVFRSPLDLYSSHFLLPNFADSHHRSILLASSGGGNGAGGGGGAGGGSGGGNGAGGGGA
SIX4_HUMAN                    MESASEGQEAHREVAGGAAVGLSPPAPAPFPLEPGDAATAAARVSGEEGAVAAAAAGAAA
SIX5_HUMAN                    ....MATLPAEPSAGPAAGGEAVAAAAATEEEEEEARQLLQTLQAAEGEAAAAAGAGAGA
SIX6_HUMAN                    ............................................................
```

```
                          1       10        20
SINE_Amphimedon           .MQCLSFARNIMNGYQLTAAMAIPYN..................................
SciSixB                   ECESVGRQDRSARSSLDGADQAEPEQADRNTSRPVQEAGPYNTSGGSHREPYQVSEQNMP
SciSixc                   AASLQAAQSVQQQQNHSLNNPLLPT...................................
LCOSixB                   ATPAAGDPVENRTDTSSMPMVTAPD...................................
LCOSixc                   REAANQAAESAHHQQHTLNNPLLPS...................................
Nvec_Six1                 .....................MLP....................................
Nvec_Six3                 ..................MFAPLP....................................
Nvec_Six4                 ......MESLDSVSPIKGDGSTLA....................................
Cwil_SIXC                 FARATAPHYSEPSHHHPHHYEKMPEHRHSVQGLNP.........................
Cwil_SIXA                 NDLKSECSVITTFQDETGIPTTFP....................................
Cwil_SIXB                 DEEEDVKREHHLAALESGILPPFPDYPD................................
Tri_SIX3                  PTNANGGNQEAPILPPGSSSPSMP....................................
Tri_SIX1A                 QPQSQYQQYHPSLSSLANGYSSLP....................................
C_elegans_ceh_33          ........MQLNSSSFHPHHFTCD....................................
C_elegans_ceh_32          ANGASSTPSLFPAMPSVIPSLAAPSSPT................................
C_elegans_ceh_34          ...........MQQSYNPQNSLTA....................................
C_elegans_unc_39          TPSSVIPTPSPSVASASISSPTIP....................................
Sine_Schmidtea            VSYEQSCLAAPALDYQTNQDSVISPDSG................................
Six3_Schmidtea            ............QMNCKQNSKLLA....................................
SIX2_Brachionus           NSSSSSASSASELVNLANGDAKTP....................................
SIX3_Brachionus           ....MNLFSALPALFPGLPAPTIP....................................
SIX1_Brachionus           ADSKTSADSVHQTDEEGPAGSVHV....................................
SIX1_Pomacea              ..MLHDQATTSVGSPTGGGPTMLP....................................
SIX6_Pomacea              GLMFGLCGFESPAAAAALPMLPLP....................................
SIX4_Pomacea              TPSCSGSGKISIIGSSGPNSNGVAASNSSSSNNNNNSAASDLLSGK..............
SINE_Capitella_teleta     MPTTTGMAPHSGAGAHAGGMPPSPRL..................................
Optix_Capitella_teleta    FGVLQSPLPLHPRASHGWPLMMLP....................................
SIX4_Capitella_teleta     KCGDLAVVSVPTPSPLNDKNEGSPGQAEGSEK............................
Sine_Drosophila           SSSNGGGGGGVVSGGGSGGRENLP....................................
Optix_Drosophila          QPPSESFSPTHHQIIAPSPILAVP....................................
Six4_Drosophila           LERRFQDITNDYDKSPPPTASTTPTHYPSLNSIIFENGSSGNLGDLNGNTKTDLCAGLQR
So_H_erato                ...MLGGPEWGQREATPPRDAPLP....................................
Optix_H_erato             ADPPAPLSLAAIELAAPTPLLPLP....................................
SIX4_H_mel                ESNKNSFENYLSPRSKKYLNFELKLPPINDH..FFSQIDNDERRTAYYNDNMKNVQNENN
SINE_Daphnia              NNNSSSIGGTSSSSLSSTGNGGLP....................................
SIX4_Daphnia              SSVSSSGGADQHHHVALALKSEQPQHNSLDQTSSLELLDSANLSYGEDPQQRHSSSGGDR
OPTIX_Daphnia             GGGGGGPGGGGGPGGLIPPPPALP....................................
SIX1_Strongylocentrotus   .....................MLP....................................
SIX6_Strongylocentrotus   MQPFVPFPAAAPGGTLFPLPTALP....................................
SIX4_Strongylocentrotus   LTCSKSKEVECMNSQLSVQHGIISGGAGTGGPG...........................
Six1/2_Halocynthia        TSAGQNVHLSSGHMASHSGMSLLPAAPPTSL.............................
Six3/6_Halocynthia        MDRLKMMLSLLQRPAALPTLFPFP....................................
Six4/5_Halocynthia        NSRRFGEDSPDLYRDKTPEIDKISTI..................................
Six1/2_Branchiostoma      .....................MLP....................................
Six4/5_Branchiostoma      VDMGTGTPGVQTSQTINGGDTAS.....................................
Six1_Petromyzon           ...................MSMLP....................................
SIX6_Petromyzon           ...................MFHLP....................................
SIX4_Petromyzon           VGMAQGLVTLAGVAPIPVPASSVPAFIAEIFG............................
Six1_like_PetromyzonX1    ...................MSLLP....................................
Six1_like_Petromyzon      ...................MSMLP....................................
Six2_like_PetromyzonX1    ASVPGPQGAAGSGAAGVPEGGLMT....................................
Six6_like_Petromyzon      ..MKQEAAFSVHGPSRCSSMLHLP....................................
Six6_like_Petromyzon2     ...................MFALP....................................
Six2_like_Petromyzon      .............MPSHAAAAAEM....................................
Six6_like_Petromyzon3     ...................MFHLP....................................
SIX1_Rhincodon            ...................MSVLP....................................
SIX2_Rhincodon            ...................MSMLP....................................
SIX3_Rhincodon            SAERAFLLLASSRPQEELSMFQLP....................................
SIX4_Rhincodon            VSALSTSPEISDQVPVELLTNPAA....................................
SIX6_Rhincodon            ...................MFQLP....................................
SIX1b_Spotted             LRGRDRFGARGDTTQRPDIMSMLP....................................
SIX2_Spotted              ...................MSMLP....................................
SIX3_Spotted              RPLVLASSAPSARSPEELSMFQLP....................................
SIX4_Spotted              LSMEQATSAAEQIHGELLASAAAS....................................
SIX5_Spotted              GSLKEETAALDEVSEELLQTFQSS....................................
SIX6_Spotted              RERVRVWCERETGEQTEASMFQLP....................................
SIX7_Spotted              SRLVLAPSGLRGAPPPGRTMFPLP....................................
Spotted_Gar_SIX1_like     .......................M....................................
SIX1_Callorhinchus        ...................MSVLP....................................
SIX2_Callorhinchus        ...................MSMLP....................................
SIX3_Callorhinchus        SADRAFLLLASSRPQEEPSMFQLP....................................
SIX4a_Callorhinchus       AAVLPSSPELPGQLPVELLAHPAS....................................
SIX6_Callorhinchus        ...................MFQLP....................................
SIX1_Erpetoichthys        ...................MSMLP....................................
SIX2a_Erpetoichthys       ...................MSMLP....................................
SIX3_Erpetoichthys        RSALLASSAAGSRAPEELSMFQLP....................................
SIX4_Erpetoichthys        LNSALLPMEHADQVHTELLLSASS....................................
SIX5_Erpetoichthys        PKAPDELGALDDASEQLLRTLRGS....................................
SIX6_Erpetoichthys        ...................MFQLP....................................
SIX7_Erpetoichthys        ...................MFSLPGL..................................
SIX1_HUMAN                ...................MSMLP....................................
SIX2_HUMAN                ...................MSMLP....................................
SIX3_HUMAN                GGAGGGGGGGSRAPPEELSMFQLP....................................
SIX4_HUMAN                DQVQLHSELLGRHHHAAAAAAQTP....................................
SIX5_HUMAN                AAAGAEGPGSPGVPGSPPEAASEPPT..................................
SIX6_HUMAN                ...................MFQLP....................................
```

```
                                                                            .
SINE_Amphimedon          ....................................AY.....G......LSQEQVACVC
SciSixB                  MQMGSD..............................DF.....S......FTLEQVACLA
SciSixc                  ....................................TY.....T......FSLDQVACVC
LCOSixB                  ....................................QF.....G......FTLEQIACLA
LCOSixc                  ....................................TY.....T......FTLDQVACVC
Nvec_Six1                ....................................SF.....S......FTPEQVACVC
Nvec_Six3                ....................................AL.....S......FSAHQIAQVC
Nvec_Six4                ..................................................FSADQVACVC
Cwil_SIXC                ....................................SY.....G......FTQEQVACVC
Cwil_SIXA                ....................................PF.....S......FSVEQVASVC
Cwil_SIXB                ....................................SYPGDHLK......FTAGQVACVC
Tri_SIX3                 ....................................GL.....H......FSVHQVASVC
Tri_SIX1A                ....................................IG.....N......FSTDQFASVC
C_elegans_ceh_33         ....................................TT.....R......YSEEQVACIC
C_elegans_ceh_32         ....................................TS.....N......LTADQIVKTC
C_elegans_ceh_34         ....................................TT.....S......YSEQEIVCIC
C_elegans_unc_39         ....................................AFGCTMSE......YSMEQMEAIS
Sine_Schmidtea           ....................................AM.....G......FTQEQVACVC
Six3_Schmidtea           ....................................PFMCHNQL......FSVEQITKVC
SIX2_Brachionus          ....................................SF.....G......FTQEQVACVC
SIX3_Brachionus          ....................................SL.....N......FSAQQVAQVC
SIX1_Brachionus          ....................................VYLRV..GEQQVVQFSLAQLECII
SIX1_Pomacea             ....................................SF.....G......FTQEQVACVC
SIX6_Pomacea             ....................................TL.....H......FTPSQVAKVC
SIX4_Pomacea             ....................................NL.....T......FSPEQVACVC
SINE_Capitella_teleta    ....................................QF.....G......FTQEQVACVC
Optix_Capitella_teleta   ....................................PV.....PPPAPPVFSSSQVTQVC
SIX4_Capitella_teleta    ....................................AL.....T......FSPEQVACVC
Sine_Drosophila          ....................................SF.....G......FTQEQVACVC
Optix_Drosophila         ....................................TL.....A......FSAAQVEIVC
Six4_Drosophila          SGGGLGGNAGSGGHLISNLTAAHNMSAVSSFPIDAKML.....Q......FSTDQIQCMC
So_H_erato               ....................................SF.....G......FTQEQVACVC
Optix_H_erato            ....................................TL.....S......FSAAQVATVC
SIX4_H_mel               PMVEVEINNFENVVNERERQYFADSDNNMRR.....CL.....N......FNAEQVQCVC
SINE_Daphnia             ....................................SF.....G......FTQEQVACVC
SIX4_Daphnia             LDLHHNGPSSASSTSSSSLGHNNNNNNNNNNNK...PMTGPSGSSSMAMAFSKDQVACVC
OPTIX_Daphnia            ....................................TL.....N......FSVSQVATVC
SIX1_Strongylocentrotus  ....................................SF.....G......FTQEQVACVC
SIX6_Strongylocentrotus  ....................................TL.....C......FSPTQIASVC
SIX4_Strongylocentrotus  ....................................IL.....S......FSAQQVVCVC
Six1/2_Halocynthia       ....................................SF.....G......FTQEQVACVC
Six3/6_Halocynthia       ....................................AP.....S......LNASQIATVC
Six4/5_Halocynthia       ....................................RLNSDAPS......YSLDNVSCIC
Six1/2_Branchiostoma     ....................................SF.....G......FTQEQVACVC
Six4/5_Branchiostoma     ....................................TL.....T......FSPEQVACAC
Six1_Petromyzon          ....................................SF.....G......FTQEQVACVC
SIX6_Petromyzon          ....................................IL.....S......FSPQQVASVC
SIX4_Petromyzon          ....................................ASLPTASGAGSLLAFSSEQVAGAC
Six1_like_PetromyzonX1   ....................................SF.....G......FTQEQVACVC
Six1_like_Petromyzon     ....................................TF.....G......FTQEQVACVC
Six2_like_PetromyzonX1   ...........................................G......FSDEQVACVC
Six6_like_Petromyzon     ....................................TL.....T......FSAQQVAGVC
Six6_like_Petromyzon2    ....................................TL.....S......FSPQQVASVC
Six2_like_Petromyzon     ....................................SF.....G......FTQEQVACVC
Six6_like_Petromyzon3    ....................................VL.....S......FSPQQVASVC
SIX1_Rhincodon           ....................................SF.....G......FTQEQVACVC
SIX2_Rhincodon           ....................................TF.....G......FTQEQVACVC
SIX3_Rhincodon           ....................................TI.....N......FTPEQVASVC
SIX4_Rhincodon           ....................................AL.....T......FSPEQVACVC
SIX6_Rhincodon           ....................................IL.....N......FSPQQVAGVC
SIX1b_Spotted            ....................................SF.....G......FTQEQVACVC
SIX2_Spotted             ....................................TF.....G......FTQEQVACVC
SIX3_Spotted             ....................................TL.....N......FSPEQVASVC
SIX4_Spotted             ....................................SL.....A......FSPEQVACVC
SIX5_Spotted             ....................................VL.....S......FSADQVACLC
SIX6_Spotted             ....................................IL.....N......FSPQQVAGVC
SIX7_Spotted             ...........................................M......FTPEQVARVC
Spotted_Gar_SIX1_like    ....................................AF.....G......FSQDQVACVC
SIX1_Callorhinchus       ....................................SF.....G......FTQEQVACVC
SIX2_Callorhinchus       ....................................AF.....G......FTQEQVACVC
SIX3_Callorhinchus       ....................................TI.....N......FTPEQVASVC
SIX4a_Callorhinchus      ....................................AL.....T......FSPEQVACVC
SIX6_Callorhinchus       ....................................IL.....N......FSPQQVAGVC
SIX1_Erpetoichthys       ....................................SF.....G......FTQEQVACVC
SIX2a_Erpetoichthys      ....................................TF.....G......FTQEQVACVC
SIX3_Erpetoichthys       ....................................TL.....N......FSPEQVASVC
SIX4_Erpetoichthys       ....................................AL.....A......FSPEQVACVC
SIX5_Erpetoichthys       ....................................GL.....N......FSAEQVSCVC
SIX6_Erpetoichthys       ....................................IL.....N......FSPQQVAGVC
SIX7_Erpetoichthys       ....................................PM............FTPEQVARVC
SIX1_HUMAN               ....................................SF.....G......FTQEQVACVC
SIX2_HUMAN               ....................................TF.....G......FTQEQVACVC
SIX3_HUMAN               ....................................TL.....N......FSPEQVASVC
SIX4_HUMAN               ....................................LA............FSPDHVACVC
SIX5_HUMAN               ....................................GL.....R......FSPEQVACVC
SIX6_HUMAN               ....................................IL.....N......FSPQQVAGVC
```

```
                         40            50                               60
                          .             .                                .
SINE_Amphimedon          DVL....QQSGNIERLARFLWSL..................P.....ACE.Q..IQKNES
SciSixB                  EFI....LHSEDVHRLERFLYAL..................P.....KCA.Q..VQRHEK
SciSixc                  EVL....QQNGQIERLERFLWSL..................P.....LHD.E..VQRHES
LCOSixB                  EFL....QQTGDVSRLESFLSSL..................P.....KCA.R..LQGNES
LCOSixc                  EVL....QQSHQIDRLERFLWSL..................P.....PCD.E..VQRNES
Nvec_Six1                EVL....QQSGDIERLGRFLWSL..................P.....ECE.T..IQKNES
Nvec_Six3                ETL....EESGDVERLARFLWSL..................PVAPG.TLE.A..LGKHES
Nvec_Six4                DAL....RQAGDIERLSRFLWSL..................P.....PDD.L..LNGSES
Cwil_SIXC                EVL....SQGGNMERLARFLWSL..................P.....SCD.H..LHKNES
Cwil_SIXA                DSL....EASGDIDRLARFLWSL..................P.....LSQ.MEEFNKNEK
Cwil_SIXB                EAL....LQSGNIKRLAAFLWSL..................P.....CHDSN..LMNNES
Tri_SIX3                 EAL....ESSGDIERLSRFLWSL..................PSTLDGYTN....LLNHDA
Tri_SIX1A                NIL....LQSNHIDRLATFLWSL..................P.....PND.E..LKVNQN
C_elegans_ceh_33         EAL......SNDARKLSQFVWTV..................L.....ERD.E..MRNNQY
C_elegans_ceh_32         EQL....ETDGDVDGLFRFMCTI..................P.....PQK.TQEVAGNEA
C_elegans_ceh_34         ESLFNEGLQTGRTEQLANFIYNL..................P.....QC.....YQVMES
C_elegans_unc_39         TSL....FQARDGDRLVAFFKQL.............ESLYGPN....AVD....HLRSEA
Sine_Schmidtea           EVL....ENGGNIDRLALFIWSL..................P.....PCQ.Q..LQTNES
Six3_Schmidtea           ETL....EEAGDIDRLSRFLWSL..................PSFSS.LWD.S..LSRQES
SIX2_Brachionus          EVL....QQSNSIDRLARFLWSL..................P.....PCE.H..LHKNES
SIX3_Brachionus          EQL....EESGDTDRLARFLWSL...............SVIP.....GSN.L..LTQHES
SIX1_Brachionus          EAL....LQMNNLKKVRSMLAMLAIDVHEGSVALDAANLNDPH....TLK.Y..LCTHDS
SIX1_Pomacea             EVL....QQGGNIDRLARFLWSL..................P.....ACE.H..LHKNES
SIX6_Pomacea             ETL....EESGDIERLGRFLWSI..................PVNPS.ACE.A..LNKHES
SIX4_Pomacea             EAL....QQKGDIERLARFLWSL..................P.....PSE.L..LRGSEA
SINE_Capitella_teleta    EVL....QNSGNIERLARFLWSL..................P.....ACE.H..LHKNES
Optix_Capitella_teleta   ETL....EESGDVERLARFLWSL..................PPPGPGLSSSD..PARCEA
SIX4_Capitella_teleta    EAL....QQSGNMERLARFLWSL..................P.....PSE.L..LRGSEA
Sine_Drosophila          EVL....QQAGNIERLGRFLWSL..................P.....QCD.K..LQLNES
Optix_Drosophila         KTL....EDSGDIERLARFLWSL..............PVALP.....NMH.E..ILNCEA
Six4_Drosophila          EAL....QQKGDIEKLTTFLCSL..................P.....PSE.F..FKTNES
So_H_erato               EVL....QQAGNIERLGRFLWSL..................P.....ACE.R..LHAHES
Optix_H_erato            ETL....EESGDVERLARFLWSL..................PVAHPNVAE....LERCEA
SIX4_H_mel               EAL....QQKGDMEKLAAFLWSL..................P.....TTE.L..LRGNET
SINE_Daphnia             EVL....QQSGNIERLARFLWSL..................P.....ACD.Q..LHKNES
SIX4_Daphnia             EAL....QQAGDMERLSRFLWSL..................P.....ASE.LSGSASSES
OPTIX_Daphnia            ETL....EESGDIERLGRFLWSL..............PVAHP.....NIG.E..LNKSEA
SIX1_Strongylocentrotus  EVL....QQSGNIERLGRFLWSL..................P.....ACE.H..LHKNES
SIX6_Strongylocentrotus  ETL....EESGDIERLARFLWSL..................PVAPG.TCE.A..LSKNES
SIX4_Strongylocentrotus  EAL....RQEGNIDRLARFLWTL..................P.....ADE.T..LQNDET
Six1/2_Halocynthia       EVL....QQGGNIERLARFLWSL..................P.....ACE.H..LHKNES
Six3/6_Halocynthia       DAL....AESGDMERLARFLWSL..................PAIPS.VME.A..LQTNES
Six4/5_Halocynthia       KAL....MQSKDPDRLERYLETL..................PT....E.A.L..NSGKEY
Six1/2_Branchiostoma     EVL....QQSGQIERLGRFLWSL..................P.....ACE.H..LHKNES
Six4/5_Branchiostoma     EAL....QQGGDIEGLARFLWSL..................P.....PNE.L..LRGSES
Six1_Petromyzon          EVL....QQGGNIERLGRFLWSL..................P.....ACE.H..LHKNES
SIX6_Petromyzon          ETL....EESGDIERLGRFLWSL..................PVAPS.AWE.A..LNKHES
SIX4_Petromyzon          ETL....LRGGDMERLGRFVHSL..................P.....TAD.L..LRSSEV
Six1_like_PetromyzonX1   EVL....QQGGNIERLGRFLWSL..................P.....ACE.H..LHKNES
Six1_like_Petromyzon     EVL....QQGGNIERLGRFLWSL..................P.....ACE.H..LHKNES
Six2_like_PetromyzonX1   EAL....QQGGDVARLARFLWAL..................PG....PAA....ASRDEA
Six6_like_Petromyzon     ETL....EESGDVERLARFLWSL..................PAAALSAGHAGDPLSRNEA
Six6_like_Petromyzon2    ETL....EESGDIERLGRFLWSL..................PVAPG.AWE.A..LNRQEA
Six2_like_Petromyzon     EAL....LQGGNMERLGRFLWSL..................P.....ACE.Q..IQRSES
Six6_like_Petromyzon3    ETL....EESGDVERLARFLWSL..................PVAPG.AWE.T..LNKNEA
SIX1_Rhincodon           EVL....QQGGNLERLGRFLWSL..................P.....ACD.H..LHKNES
SIX2_Rhincodon           EVL....QQGGNIERLGRFLWSL..................P.....ACE.H..LHKNES
SIX3_Rhincodon           ETL....EETGDIERLGRFLWSL..................PVAPG.ACE.A..INKHES
SIX4_Rhincodon           EAL....QQGGNLDRLAQFLWSL..................P.....PSD.L..LRGNES
SIX6_Rhincodon           ETL....EESGDIERLGRFLWSL..................PVAPA.ACE.A..LNKNES
SIX1b_Spotted            EVL....QQGGNLERLGRFLWSL..................P.....ACD.H..LHKNES
SIX2_Spotted             EVL....QQGGNIERLGRFLWSL..................P.....ACE.H..LHKNES
SIX3_Spotted             ETL....EETGDIERLGRFLWSL..................PVAPG.ACE.A..INKHES
SIX4_Spotted             EAL....QQGGNVDRLARFLWSL..................P.....QSD.L..LRGNES
SIX5_Spotted             EAL....LQAGNVERLGRFLSTI..................PP....SAE.L..LRGNET
SIX6_Spotted             ETL....EESGDIERLGRFLWSL..................PVAPA.ACE.V..LNKNES
SIX7_Spotted             ENL....EETGDIERLARFLWSL..................PAAVPGSAGEA..LSRHES
Spotted_Gar_SIX1_like    EVL....LQRGSVERLGRFLWSL..................P.....PCD.L..LHRNES
SIX1_Callorhinchus       EVL....QQGGNLERLGRFLWSL..................P.....ACD.H..LHKNES
SIX2_Callorhinchus       EVL....QQGGNIDRLGRFLWSL..................P.....ACD.H..LHKNES
SIX3_Callorhinchus       ETL....EETGDIERLGRFLWSL..................PVAPG.ACE.A..INKHES
SIX4a_Callorhinchus      EAL....QQGGNLDRLAQFLWSL..................P.....ASE.L..LRGNES
SIX6_Callorhinchus       ETL....EESGDIERLGRFLWSL..................PVAPA.ACE.A..LNKNES
SIX1_Erpetoichthys       EVL....QQGGNLERLGRFLWSL..................P.....ACD.H..LHKNES
SIX2a_Erpetoichthys      EVL....QQGGNIERLGRFLWSL..................P.....ACE.H..LHKNES
SIX3_Erpetoichthys       ETL....EETGDIERLGRFLWSL..................PVAPG.ACE.A..INKHES
SIX4_Erpetoichthys       EAL....QQGGNLDRLARFLWSL..................P.....QSD.L..LRGNES
SIX5_Erpetoichthys       EAL....LQAGNVEQLGRFLSTI..................PA....APE.L..QRGNET
SIX6_Erpetoichthys       ETL....EESGDIERLGRFLWSL..................PVAPA.ACE.V..LNKNES
SIX7_Erpetoichthys       ENL....EDTGDMERLGRFLWSL..................PSAMPGSATDA..LNRHES
SIX1_HUMAN               EVL....QQGGNLERLGRFLWSL..................P.....ACD.H..LHKNES
SIX2_HUMAN               EVL....QQGGNIERLGRFLWSL..................P.....ACE.H..LHKNES
SIX3_HUMAN               ETL....EETGDIERLGRFLWSL..................PVAPG.ACE.A..INKHES
SIX4_HUMAN               EAL....QQGGNLDRLARFLWSL..................P.....QSD.L..LRGNES
SIX5_HUMAN               EAL....LQAGHAGRLSRFLGAL..................P.....PAE.R..LRGSDP
SIX6_HUMAN               ETL....EESGDVERLGRFLWSL..................PVAPA.ACE.A..LNKNES
```



```
                          70             80        90        100       110       120
SINE_Amphimedon           VLKAKALIAF....HQGNFQELYRIIETNNFSPDS.HPKMQQLWLQAHYIEAERLRG..K
SciSixB                   ILIAKANVAYHSGLHTGDFRALYQILESETFSEQS.YPRLQEMWTEAHYKEAERNREGRK
SciSixc                   VLKARAILAF....HRNNYQQLYHILKSYQFSPSS.HQKMQQLWLQAHYREAERVRG..R
LCOSixB                   VLIAKAVVAF....HRGDYKHLYHILESHNFSEPS.YPKLQKLWLQAHYIEAERQRG..R
LCOSixc                   VLTSRAIVAF....HRGNFPQLYHILKSYQFSGPY.HGKMQQLWLQAHYLEAERQRG..R
Nvec_Six1                 VLKAKAIVSF....HQQNFQELYRILENNNFSPNA.HPKLQSLWLKAHYMEAEKLRG..R
Nvec_Six3                 VLRARAIVAF....HMGNFRDLYHILETHRFTRES.HAKLQAMWLEAHYQEAERLRG..R
Nvec_Six4                 VLKARAIVSF....HRGRYREVYNILETNEFDPSS.HELLQCLWYKAHYSEAEKLRG..R
Cwil_SIXC                 VLKAKAVVAF....HRGNFKELYQILENNSFSANN.HPKLQSIWLKAHYMEAEKLRG..R
Cwil_SIXA                 ILRSRAVVSF....HRQDFRELYSIIENCRFKKSS.HEKLQYLWNEAHYMEAEKLRG..R
Cwil_SIXB                 VMKARAEVAF....NNGNFSEVYRILGSRNFSPNS.HPKLQQLWLKSHYIEAETARG..R
Tri_SIX3                  ILRARAVVAY....HQGHYRELYGIIENHRFPKDF.HGKLQHMWLEAHYREAEKLRG..R
Tri_SIX1A                 ILLARATVAY....HQHNFEELYQLLENYPFSSEF.HPKLQELWKEAHYLEEKQSRG..K
C_elegans_ceh_33          ILKAQAFLAF....HSNNFKELYRIIESHHFASEH.HLPLQEWWLNAHYHEAEKIRG..R
C_elegans_ceh_32          FLRARALVCF....HASHFRELYAILENNKFSPKY.HPKLQEMWHEAHYREQEKNRG..K
C_elegans_ceh_34          VLKAQALVYF....TTQNWKMLYKLLECSKFSPHN.HTVLQNLWLDAHYKEAAKTKD..R
C_elegans_unc_39          IIVAYTYALY....HSNEFETLFHLLSNRHFQQRH.YNDLQDIWHHARYKESQLKRG..K
Sine_Schmidtea            VLTAKAAVAF....HRQNFKELYRILESYTFSPHN.HYKLQALWLQAHYIEEEKIKG..R
Six3_Schmidtea            IQRARALVAF....HVGNFRELYNLIEKNRFTKAS.HSKLQALWLEAHYQEAERLRG..R
SIX2_Brachionus           VLKAKAVVAF....HQRRFRDLYKILENNHFMAHN.HSKLQQLWLAAHYLEAENIKG..R
SIX3_Brachionus           IIRARAIVAF....QQNNYRELYTLLENHKFTRDS.HPKLQAMWMEAHYQEAEKLRG..R
SIX1_Brachionus           ILKCRAALLL....DECRFKELYSLLENHEFDLCH.HNDLQAMWYKGHYAEAQKVRG..R
SIX1_Pomacea              VLKAKAVVAF....HRGNFKELYKILENNQFSPHN.HPKMQALWLKAHYVEAEKLRG..R
SIX6_Pomacea              VLRARALVSF....HTGNFRDLYHILENHKFTKES.HAKLQAMWLEAHYQEAEKLRG..R
SIX4_Pomacea              VLKARATVAF....HRGSYRELYAILESHNFDESN.HAFLQQLWYKAHYMEAQKVRG..R
SINE_Capitella_teleta     VLKAKAVVAF....HRGNFKELYKILESQTFSPHN.HPKLQALWLKAHYIEAEKLRG..R
Optix_Capitella_teleta    VLRARALVAF....HAGNFRELYALLESHKFSKDS.HSKLQAMWLEAHYQEAERLRG..R
SIX4_Capitella_teleta     VLKARATVAF....HKGNFRELYAITESHNFDPAN.HAVMQQMWYKAHYLEAQKVRG..R
Sine_Drosophila           VLKAKAVVAF....HRGQYKELYRLLEHHHFSAQN.HAKLQALWLKAHYVEAEKLRG..R
Optix_Drosophila          VLRARAVVAY....HVGNFRELYAIIENHKFTKAS.YGKLQAMWLEAHYIEAEKLRG..R
Six4_Drosophila           VLRARAMVAY....NLGQFHELYNLLETHCFSIKY.HVDLQNLWFKAHYKEAEKVRG..R
So_H_erato                VLKAKAMVAF....HRGNFKELYRLLESHNFSAHN.HEKLQNLWLKAHYMEAERLRG..R
Optix_H_erato             VLRARAVVAF....HAGRHRELYSILERHRFQRSS.HAKLQALWLEAHYQEAERLRG..R
SIX4_H_mel                VLRARALVAY....HHGIFQELYTILEKHSFPPRH.HNALQTLWFKAHYKEAEKVRG..R
SINE_Daphnia              VLKAKAVVAF....HRANFKELYKLLETHPFSPHN.HPKLQALWLKAHYIEAEKLRG..R
SIX4_Daphnia              VLRARVAVAF....HRGNYRELYNLLESHSFSSQY.HQELQNIWYGAHYKEAEKVRN..R
OPTIX_Daphnia             VLRARALVAY....HMGNYRELYHIVESHRFTKDS.HGKLQAMWLEAHYLEAEKLRG..R
SIX1_Strongylocentrotus   VLKAKAIVAF....HRGNFRELYKILESNNFSPHN.HPKLQALWLKAHYIEAEKLRG..R
SIX6_Strongylocentrotus   VLRARAVVSF....HQGNYRELYHILENHRFTKDS.HAKLQAMWLEAHYQEAEKLRG..R
SIX4_Strongylocentrotus   VLRARAAVAY....HQGHYKELYNLLQNHNFNPAF.HTELQDLWYQAHYKESEKLRG..R
Six1/2_Halocynthia        VLKAKAVVAF....HRGNFRELYKLLESHNFSPHN.HPKLQQLWLKAHYIEAEKLRG..R
Six3/6_Halocynthia        VLRARSLVAF....HQGNFREVYNILEHHRFTDAAWHHRLQAMWLEAHYQDAERSRG..R
Six4/5_Halocynthia        VVMARACIAS....HRENFKDMFVLLESRPFTTCN.HKFLQGLWYSAHYAEAEKIRG..R
Six1/2_Branchiostoma      VLKAKAVVAF....HRGNFRGLYKILESNQFGPEN.HPKLQQLWLKAHYMEAEKLRG..R
Six4/5_Branchiostoma      VLKARAIVAF....HRGSFKELYAILESHNFDTSN.HQMLRDMWYKAHYIEAQKIRG..R
Six1_Petromyzon           VLKAKAVVAF....HRGNFRELYKILESHQFSPHN.HPKLQQLWLKAHYMEAEKLRG..R
SIX6_Petromyzon           VLRARAVVAF....HAGNFRDLYHILENHKFTKES.HAKLQAMWLEAHYQEAEKSRG..R
SIX4_Petromyzon           VVKARALVAF....HSGRFQELYSILESHHFQPQS.HALMQGLWYRARYSDAERLRG..R
Six1_like_PetromyzonX1    VLKAKAVVAF....HRGNFRELYKTLESHQFSAHN.HPKLQQLWLKAHYTEAEKLRG..R
Six1_like_Petromyzon      VLKAKAVVAF....HRGNFRELYKTLESHQFSAHN.HPKLQQLWLKAHYMEAEKLRG..R
Six2_like_PetromyzonX1    VLRARAHVAF....HSGSYRELYDVLEGHAFEARH.HAALQALWFRARYLEAERARG..R
Six6_like_Petromyzon      VLRAQAVVAF....HTGNFGDMYRILEGHRFAKAS.HAKLQAMWLEARYQEAERLRG..R
Six6_like_Petromyzon2     VLRARAVVAF....HASNYRDLYAILEGHKFSKAS.HAKLQAMWLEAHYLEAERLRG..R
Six2_like_Petromyzon      ILMAKAVVAF....HQGNFRELYAVLESQPFSARN.HPKLQQLWLKAHYTEAERLRG..R
Six6_like_Petromyzon3     VLRARAVVAF....HAGNYRDLYAILEGHRFSKAS.HAKLQAMWLEARYQEAERLRG..R
SIX1_Rhincodon            VLKAKAVVAF....HRGNFRELYKILESHQFSPHN.HSKLQQLWLKAHYIEAEKLRG..R
SIX2_Rhincodon            VLKAKAVVSF....HRGNFRELYKILESHQFSPHN.HPKLQQLWLKAHYIEAEKLRG..R
SIX3_Rhincodon            ILRARAVVAF....HTGNFRDLYHILENHKFTKES.HGKLQAMWLEAHYQEAEKLRG..R
SIX4_Rhincodon            ILKARALVAF....HQSRYKELYSILESHNFDSSC.HTSLQDLWYKARYTEAEKVRG..R
SIX6_Rhincodon            VLRARAIVAF....HTGNFRELYHILENHKFTKES.HGKLQALWLEAHYQEAEKLRG..R
SIX1b_Spotted             VLKAKAVVAF....HRGNFRELYKILESHQFSPHN.HPKLQQLWLKAHYVEAEKLRG..R
SIX2_Spotted              VLKAKAVVAF....HRGNFRELYKILESHQFSPHN.HPKLQQLWLKAHYIEAEKLRG..R
SIX3_Spotted              ILRARAVVAF....HTGNFRDLYHILENHKFTKDS.HGKLQAMWLEAHYQEAEKLRG..R
SIX4_Spotted              ILKAQALVAF....HQARYQELYSILENHSFSPPN.HPSLQDLWYRARYTEAEKARG..R
SIX5_Spotted              LLKAKALVAF....HQEEFKELYAILESHDFHPAN.HAFLQNLYLQARYKEAEMSRG..R
SIX6_Spotted              VLRARAIVAF....HTGNFRELYHILENHKFTKES.HSKLQALWLEAHYQEAEKLRG..R
SIX7_Spotted              VVRARALVAF....HGGNFEALYRILQTHRFTRQS.HARLQALWLDAHYREAERLRG..R
Spotted_Gar_SIX1_like     VLKAKALVAF....HHGSFRELYQLLESQPFSPHN.HGYLQQLWLRAHYLEAERLRG..R
SIX1_Callorhinchus        VLKAKAVVAF....HRGNFRELYKILESHQFSPHN.HPKLQQLWLKAHYIEAEKLRG..R
SIX2_Callorhinchus        VLKAKAVVAF....HRGSFRELYKILEGHQFLPHN.HPKLQQLWLKAHYSEAEKLRG..R
SIX3_Callorhinchus        ILRARAVVAF....HTGNFRDLYHILENHKFTKES.HGKLQAMWLEAHYQEAEKLRG..R
SIX4a_Callorhinchus       ILNARALVAF....HQGRYKELYGILESHNFEASC.HTFLQDLWYKARYTEAEKVRG..R
SIX6_Callorhinchus        VLRARAIVAF....HTGNYRELYHILENHKFTKES.HGKLQALWLEAHYQEAEKLRG..R
SIX1_Erpetoichthys        VLKAKAVVAF....HRGNFRELYKILESHQFSPHN.HPKLQQLWLKAHYVEAEKLRG..R
SIX2a_Erpetoichthys       VLKAKAVVAF....HRGNFRELYKILESHQFSPHN.HPKLQQLWLKAHYIEAEKLRG..R
SIX3_Erpetoichthys        ILRARAVVAF....HTGNFRDLYHILENHKFTKDS.HGKLQAMWLEAHYQEAEKLRG..R
SIX4_Erpetoichthys        ILKAQALVAF....HQGRYPELYSILETHNFSPCN.HSCLQDLWYKARYTEAEKARG..R
SIX5_Erpetoichthys        LLKAKALVAF....HREEYKELYAILESHSFHPSN.HAFLQDLYLQSRYREAERSRG..R
SIX6_Erpetoichthys        VLRARAIVAF....HTGNFRELYHILENHKFTKDS.HAKLQALWLEAHYQEAEKLRG..R
SIX7_Erpetoichthys        VMRARALVAF....HSGNFEALYQILQSHRFTRES.HAKLQALWLDAHYREAERLRG..R
SIX1_HUMAN                VLKAKAVVAF....HRGNFRELYKILESHQFSPHN.HPKLQQLWLKAHYVEAEKLRG..R
SIX2_HUMAN                VLKAKAVVAF....HRGNFRELYKILESHQFSPHN.HAKLQQLWLKAHYIEAEKLRG..R
SIX3_HUMAN                ILRARAVVAF....HTGNFRDLYHILENHKFTKES.HGKLQAMWLEAHYQEAEKLRG..R
SIX4_HUMAN                LLKARALVAF....HQGIYPELYSILESHSFESAN.HPLLQQLWYKARYTEAERARG..R
SIX5_HUMAN                VLRARALVAF....QRGEYAELYRLLESRPFPAAH.HAFLQDLYLRARYHEAERARG..R
SIX6_HUMAN                VLRARAIVAF....HGGNYRELYHILENHKFTKES.HAKLQALWLEAHYQEAEKLRG..R
```

```
                                  130       140       150       160        170       180
                                    .         .         .         .          .         .
SINE_Amphimedon             PLGAVGKYRIRRKFPLPRTIWDGEETSYCFKEKSRVVLRQWYTK.NAYPSPREKRQLAEQ
SciSixB                     PLGAVGKYRVRKKNPFPRTIWDGEETNYCFKEKSRVRLREWYSK.SPYPSPQQKKDLARD
SciSixc                     ALGAVGKYRIRRKYPLPRTIWDGEETSYCFKDKSRNRLKDYYKS.NRYPSPAQKRKLAAE
LCOSixB                     ALGAVGKYRIRRKFPLPHSIWDGEETNYCFKEKSRVLLREWYKK.SPYPSPPQKKELAQK
LCOSixc                     ALGAVGKYRIRRKYPLPRTIWDGEETSYCFKDRSRNRLKQYYAK.NAYPSPTQKQELASE
Nvec_Six1                   PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRNILREWYSH.NPYPSPREKRELAEG
Nvec_Six3                   PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRSLLREWYLQ.DPYPNPTKKRELAQA
Nvec_Six4                   SLGAVDKYRIRRKFPLPRTIWDGEETVYCFKEKARAALKDCYEQ.NKYPTPQEKRLIAKQ
Cwil_SIXC                   SLGAVGKYRVRKKFPLPRTIWDGDETSYCFKEKSRTVLRDWYAH.NPYPSPREKRELAEA
Cwil_SIXA                   PLGAVDKYRVRKKFPLPRTIWDGKIQNHCFKEKSRNILKEWYSK.NPYPSPHTKRELADA
Cwil_SIXB                   PLGAVDKYRIRRKYPLPNTIWDGEETSYCFKEKSRNRLREWYAQ.NKYPSPHEKRQLAES
Tri_SIX3                    SLGPVDKYRIRKKYPLPVTIWDGEQKTHCFKEKTRNLLREWYLR.DPYPNPGKKRELANA
Tri_SIX1A                   ELDAVTKYRVRKKYPLPLTISDGEKITYSFKESSRKMLVEYYQR.NPYPTSEEKAIIAEA
C_elegans_ceh_33            QLGAVGKYRIRRKYPLPRTIWDGEETSYCFRDKSRVLLRDWYCR.NSYPSPREKRELAEK
C_elegans_ceh_32            SLCAVDKYRVRKKYPMPRTIWDGEQKTHCFKERTRSLLREWYLK.DPYPNPPKKKELANA
C_elegans_ceh_34            ELGAVCKYRIRKKNPFPNTIWDGEETNYCFKSKSRNVLRDAYKK.CQYPSVEDKRRLAQQ
C_elegans_unc_39            ELNPVEKYRLRRKFPAPKTIWDGEEIVYSFKDSSRKFLKQFFRNVSEYPTQEQKREISRA
Sine_Schmidtea              SLGAVAKYRIRRKYPLPRTIWDGEETSYCFKEKSRAVLRQWYLH.NPYPSPREKKDLAEM
Six3_Schmidtea              SLGPVDKYRVRKKFPMPRTIWDGEQKTHCFKERTRNLLRECYLD.DPYPNPSKKRQLASA
SIX2_Brachionus             PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRTVLREWYIH.NPYPSPREKRELAEA
SIX3_Brachionus             PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRGLLREWYLQ.DPYPNPAKKRELAQA
SIX1_Brachionus             SLGAVDKYRIRRKFPLPKSIWDGEETIYCFKEKSRQALKDCYRQ.NRYPTPDEKRALAKR
SIX1_Pomacea                PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRTILREWYAH.NPYPSPREKRELAEA
SIX6_Pomacea                PLGPVDKYRVRKKYPLPRTIWDGEQKTHCFKERTRNLLREWYLQ.DPYPNPTKKRELAQA
SIX4_Pomacea                ALGAVDKYRLRRKYPLPKTIWDGEETIYCFKEKSRQALKECYKQ.NRYPTPDEKRGLAKK
SINE_Capitella_teleta       PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRTVLREWYAH.NPYPSPREKRELAEA
Optix_Capitella_teleta      PLGPVDKYRVRKKFPFPCSIWDGEQKSHCFKEKTRNLLREWYLQ.DPYPNPTKKRELAKA
SIX4_Capitella_teleta       PLGAVDKYRLRRKYPLPKTIWDGEETIYCFKEKSRQALKECYKQ.NRYPTPDEKRALAKK
Sine_Drosophila             PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRSVLRDWYSH.NPYPSPREKRDLAEA
Optix_Drosophila            SLGPVDKYRVRKKFPLPPTIWDGEQKTHCFKERTRSLLREWYLQ.DPYPNPTKKRELAKA
Six4_Drosophila             PLGAVDKYRLRKKYPLPKTIWDGEETVYCFKEKSRNALKDCYLT.NRYPTPDEKKTLAKK
So_H_erato                  PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRSVLRDWYLH.NPYPSPREKRELAET
Optix_H_erato               PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRSLLREWYLQ.DPYPNPTKKRELAAA
SIX4_H_mel                  PLGAVDKYRLRKKYPLPKTIWDGEETVYCFKEKSRNALKDCYYR.NRYPTPDEKRALAQK
SINE_Daphnia                PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRSVLRDWYGH.NPYPSPREKRELAEA
SIX4_Daphnia                ALGAVDKYRLRRKYPLPKTIWDGEETIYCFKEKSRAALKDCYRQ.NRYPTPDEKRTLAKK
OPTIX_Daphnia               PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRSLLREWYLQ.DPYPNPTKKRELAQA
SIX1_Strongylocentrotus     PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRSILREWYSH.NPYPSPREKRELAEA
SIX6_Strongylocentrotus     PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRSLLREWYLQ.DPYPNPTKKRELAQA
SIX4_Strongylocentrotus     PLGAVDKYRIRRKHPLPRTIWDGEEMAYCFKEKSRNMLKECYKQ.NRYPTPDEKRNLAKV
Six1/2_Halocynthia          PLGAVGKYRVRRKFPLPRCIWDGEETSYCFKEKSRAALREWYAH.NPYPSPREKRELAEA
Six3/6_Halocynthia          ALGPVDKYRIRKKFPLPRSIWNGEQKSHCFKERTRNSLRESYLR.DPYPNPSKKRELARL
Six4/5_Halocynthia          PLGAVDKYRIRRKHPLPRTIWDGEEMVYCFKERSRKALKDCYMS.NRYPTPDEKRQLAKI
Six1/2_Branchiostoma        PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRGVLREWYAH.NPYPSPREKRELAEA
Six4/5_Branchiostoma        PLGAVDKYRLRRKFPLPRTIWDGEETVYCSKEKARQALKEMYNN.NRYPTPDEKRNLAKK
Six1_Petromyzon             PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRGVLREWYAH.NPYPSPREKRELAEA
SIX6_Petromyzon             PLGPVDKYRVRKKFPLPKTIWDGEQKTHCFKERTRNLLREWYLQ.DPYPNPSKKRELAQA
SIX4_Petromyzon             ALCAVDKYRLRRKFPLPRTIWDGEETVYCFKEKSRNFLKDCYRR.TRYPAPDEKRRLAKL
Six1_like_PetromyzonX1      PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRGVLREWYAH.NPYPSPREKRELAEA
Six1_like_Petromyzon        PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRSILREWYAH.NPYPSPREKRELAEA
Six2_like_PetromyzonX1      ALGAVDKYRVRKRHPPPRTIWDGERTSYCFKERARGALMESYGG.ARYPSPEQKLQLALA
Six6_like_Petromyzon        PLGPVDKYRVRKKFPLPKTIWDGEQKSHCFKERTRSLLREWYLQ.DPYPNPAKKRELAQA
Six6_like_Petromyzon2       PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRSLLREWYLQ.DPYPNPAKKRELAQA
Six2_like_Petromyzon        PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKERSRGVLRDWYAH.NPYPSPREKRELAQA
Six6_like_Petromyzon3       PLGPVDKYRVRKKFPLPPTIWDGEPKTHCFKERTRRVLREWYLQ.DPYPNPAKKRELAQA
SIX1_Rhincodon              PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRGVLREWYAH.NPYPSPREKRELAEA
SIX2_Rhincodon              PLGAVGKYRVRRKFPLPRSIWDGEETSYCFKEKSRSVLREWYAH.NPYPSPREKRELAEA
SIX3_Rhincodon              PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRSLLREWYLQ.DPYPNPSKKRELAQA
SIX4_Rhincodon              PLGAVDKYRLRRKFPLPRTIWDGEERVYCFKEKSRNALKELYKQ.NRYPSPAEKRNLAKI
SIX6_Rhincodon              PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRHLLREWYLQ.DPYPNPSKKRELAQA
SIX1b_Spotted               PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRGVLREWYTH.NPYPSPREKRELAEA
SIX2_Spotted                PLGAVGKYRVRRKFPLPRSIWDGEETSYCFKEKSRSVLREWYTH.NPYPSPREKRELAEA
SIX3_Spotted                PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRSLLREWYLQ.DPYPNPSKKRELAQA
SIX4_Spotted                PLGAVDKYRLRRKYPLPRTIWDGEETVYCFKERSRNALKDLYKQ.NRYPSPAEKRNLAKI
SIX5_Spotted                SLGAVDKYRLRKKFPLPKTIWDGEETVYCFKEKSRNALKECYRS.NRYPTLDEKRHLAKI
SIX6_Spotted                PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRHLLREWYLQ.DPYPNPSKKRELAQA
SIX7_Spotted                PLGPVEKYRIRKKFPLPRTIWDGEQKTHCFKERTRSLLREWYLQ.DPYPNPSRKRHLAQA
Spotted_Gar_SIX1_like       PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRGVLRQWYLH.KPYPSPREKRELAEA
SIX1_Callorhinchus          PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRGVLREWYAH.NPYPSPREKRELAEA
SIX2_Callorhinchus          PLGAVGKYRVRRKFPLPRTIWDGDETSYCFKEKSRSVLREWYNH.NPYPSPCEKRELAEA
SIX3_Callorhinchus          PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRSLLREWYLQ.DPYPNPSKKRELAQA
SIX4a_Callorhinchus         PLGAVDKYRLRRKFPLPRTIWDGEERVYCFKEKSRNALKELYKQ.NRYPSPADKRNLAKL
SIX6_Callorhinchus          PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRHLLREWYLQ.DPYPNPSKKKELAQA
SIX1_Erpetoichthys          PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRGVLREWYTH.NPYPSPREKRELAEA
SIX2a_Erpetoichthys         PLGAVGKYRVRRKFPLPRSIWDGEETSYCFKEKSRSVLREWYTH.NPYPSPREKRELAEA
SIX3_Erpetoichthys          PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRSLLREWYLQ.DPYPNPSKKRELAQA
SIX4_Erpetoichthys          PLGAVDKYRLRRKYPLPRTIWDGEETVYCFKERSRNALKELYKQ.NRYPSPAEKRNLAKI
SIX5_Erpetoichthys          RLGAVDKYRLRKKYPLPKTIWDGEETVYCFKEKSRNALKECYKS.NRYPTPDEKRNLARL
SIX6_Erpetoichthys          PLGPVD.................NKKTHCFKERTRHLLREWYLQ.DPYPNPSKKRELAQA
SIX7_Erpetoichthys          PLGPVEKYRIRKKFPLPRTIWDGEQKTHCFKERTRSLLREWYLQ.DPYPNPSRKRHLAQA
SIX1_HUMAN                  PLGAVGKYRVRRKFPLPRTIWDGEETSYCFKEKSRGVLREWYAH.NPYPSPREKRELAEA
SIX2_HUMAN                  PLGAVGKYRVRRKFPLPRSIWDGEETSYCFKEKSRSVLREWYAH.NPYPSPREKRELAEA
SIX3_HUMAN                  PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRSLLREWYLQ.DPYPNPSKKRELAQA
SIX4_HUMAN                  PLGAVDKYRLRRKFPLPRTIWDGEETVYCFKEKSRNALKELYKQ.NRYPSPAEKRHLAKI
SIX5_HUMAN                  ALGAVDKYRLRKKFPLPKTIWDGEETVYCFKERSRAALKACYRG.NRYPTPDEKRRLATL
SIX6_HUMAN                  PLGPVDKYRVRKKFPLPRTIWDGEQKTHCFKERTRHLLREWYLQ.DPYPNPSKKRELAQA
```

```
                                190       200                     210       220
                                  .         .                       .         .
SINE_Amphimedon          TGLTTTQVSNWFKNRRQRDRA......AE.....T.KS..TDPKFKQDLSLNSTPSSSTS
SciSixB                  TGLTITQVSNWFKNRRXRDRA......AEGRQDGGSGDDGREDLDDDEQTGMHSLDQATS
SciSixc                  TGLTVTQVSNWFKNRRXRDRA......SDRSAGKS.SS..RKTSTSSLATS.....DSGR
LCOSixB                  TGLTITQVSNWFKNRRXRDRA......ADFRQDGVNDDDDRLEDEDGEEMQHGLSHEPSS
LCOSixc                  TELTTTQVSNWFKNRRQRDRA......AE...KVS.KGVRRQSSSSLATSDSGRDSVNSG
Nvec_Six1                TGLTTTQVSNWFKNRRQRDRA......AE.....A.KI..RETTVDGRNPK.........
Nvec_Six3                TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RMS.................
Nvec_Six4                TNLTLKQVSNWFKNRRQRDRI......PS.....N.....KSSSGTLRRNS.........
Cwil_SIXC                TGLTTTQVSNWFKNRRQRDRA......AE.....S.KD..RDDDGDPRERM.........
Cwil_SIXA                AGLTPTQVSNWFKNRRQRDRA......AI.....S.KT..RHETK...............
Cwil_SIXB                TGLSLTQVSNWFKNRRQRDRA......AE.......TKTKRGGKEVEEGSC.........
Tri_SIX3                 TGLTPTQVGNWFKNRRQRDRA......AA.....A.KH..KMASQNRKQSS.........
Tri_SIX1A                ASLTKVQVSNWFKNKRQRDRA......KS.....C.TT..SSESQDEDLDI.........
C_elegans_ceh_33         THLTVTQVSNWFKNRRQRDRAGV....PE.....P.KDCLKDISEEEDLKLI........
C_elegans_ceh_32         TGLTQMQVGNWFKNRRQRDRA......AA.....A.KNKQNIIGVELKKTS.........
C_elegans_ceh_34         TELSIIQVSNWFKNKRQRERA......AG...........QLDRSSARSND.........
C_elegans_unc_39         TGLKIVQISNWFKNRRQRDKS...............NN......................
Sine_Schmidtea           TSLTTTQVSNWFKNRRQRDRA......AE.....N.KD..KHDDLSASEADSESLKDIKE
Six3_Schmidtea           TGLTPTQVGNWFKNRRQRDRA......AA.....A.KNGRQMSQSEDEKFE.........
SIX2_Brachionus          TGLTTTQVSNWFKNRRQRDRA......AE.....A.KD..RDPMLDNSQNDSFDSLADSS
SIX3_Brachionus          TGLTATQVGNWFKNRRQRDRA......AA.....A.KN..RNSNNQYD............
SIX1_Brachionus          TGLTLTQVSNWFKNRRQRDRS......TS...........RATCNTITPAL.........
SIX1_Pomacea             TGLTTTQVSNWFKNRRQRDRA......AE.....A.....KEREHGGSGGG.........
SIX6_Pomacea             TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RLQQQGHLSKL.........
SIX4_Pomacea             TGLTLTQVSNWFKNRRQRDRT.......................................
SINE_Capitella_teleta    TGLTTTQVSNWFKNRRQRDRA......AE.....A.KD..REPGQSSLDVG.....GQSQ
Optix_Capitella_teleta   TSLTPTQVGNWFKNRRQRDRA......AA.....Q.KN..R...................
SIX4_Capitella_teleta    TGLTLTQVSNWFKNRRQRDRT......PH.......GG..HQAQCRSMYGPCDDPMTMSA
Sine_Drosophila          TGLTTTQVSNWFKNRRQRDRA......AE.....H.KDGSTDKQHLDSSSD.........
Optix_Drosophila         TGLNPTQVGNWFKNRRQRDRA......AA.....A.KN..RIQHSQNSSGM.....GCRS
Six4_Drosophila          TGLTLTQVSNWFKNRRQRDRT......PQ.....Q.RP......................
So_H_erato               TGLTTVQVSNWFKNRRQRDRQ......AE.....H.KD......................
Optix_H_erato            TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RSAVLGRGFAS.........
SIX4_H_mel               TGLTLTQVSNWFKNRRQRDRT......PQ.....Q.QN..RSEMLVSAQYV.........
SINE_Daphnia             TGLTTTQVSNWFKNRRQRDRA......AE.....H.KD..RVSGGSSAKDGGDGANDASA
SIX4_Daphnia             TGLTLTQVSNWFKNRRQRDRT......PP.......QQQRSDDCMGSDNGLPTTPVNGSS
OPTIX_Daphnia            TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RRMGGHGGHHNSCNYSSGSS
SIX1_Strongylocentrotus  TGLTTTQVSNWFKNRRQRDRA......AE.....A.KE..RENAEQESKTK.....MATP
SIX6_Strongylocentrotus  TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RMQQQQQQALQ.........
SIX4_Strongylocentrotus  TGLTMTQISNWFKNRRQRDKL......PMSLKGGG.DGEYSNGKDSPNGDFSLDGETEGS
Six1/2_Halocynthia       TGLTVTQVSNWFKNRRQRDRA......AE.....A.KE..RDGSEAGMSGGISGDGMVTP
Six3/6_Halocynthia       TGLSPTQVGNWFKNRRQRDRA......AA.....A.KN..RLMNEQQNGGQISTPSSSNN
Six4/5_Halocynthia       TSLSVTQVSNWFKNRRQRDRS......PHTSPISN.RPPQHDTLMTGKTDN.........
Six1/2_Branchiostoma     TGLTTTQVSNWFKNRRQRDRA......AE.....A.KEREQQEQTKLGPDQ.........
Six4/5_Branchiostoma     TGLTLTQVSNWFKNRRQRDRT......PS.....IHKGDDRHSGLDSGDEDLKGHLNGDL
Six1_Petromyzon          TGLTTTQVSNWFKNRRQRDRA......AE.....A.KE..RENSENTNSNSSGTAPTQLL
SIX6_Petromyzon          TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RLQV................
SIX4_Petromyzon          TGLSVVQVSNWFKNRRQRERG......PQDGTHLK.SENDTDDGEQSAKEDHVDFQEQHM
Six1_like_PetromyzonX1   TGLTTTQVSNWFKNRRQRDRA......AE.....A.KD..RENSENAHHHHSNAGGGTGG
Six1_like_Petromyzon     TGLTTTQVSNWFKNRRQRDRA......AE.....A.KE..RYGENNENANSNTGGGGGGG
Six2_like_PetromyzonX1   TGLTATQVANWFKNRRQRDRA......PG.....P.AGDRRQPRRSEPTGGSSAGDDDGD
Six6_like_Petromyzon     TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RMLLLLQP............
Six6_like_Petromyzon2    TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RLQQQTVTHSV.........
Six2_like_Petromyzon     TGLTTTQVSNWFKNRRQRDRA......AE.....S.RD..RDGSDNADGDGTARPPGNTT
Six6_like_Petromyzon3    TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RLHQPSLSPGA.........
SIX1_Rhincodon           TGLTTTQVSNWFKNRRQRDRA......AE.....A.KE..RENTENTNTSGNKQNQL...
SIX2_Rhincodon           TGLTTTQVSNWFKNRRQRDRA......AE.....A.KE..RYEENSENSNS.....NVTA
SIX3_Rhincodon           TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RLQHQAVGQSG.........
SIX4_Rhincodon           TGLSLTQVSNWFKNRRQRDRN......PS.......ENQSKSESDGNHSTE.....DESS
SIX6_Rhincodon           TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RLQQQILSQAS.........
SIX1b_Spotted            TGLTTTQVSNWFKNRRQRDRA......AE.....A.KE..RENSENNNSGN.........
SIX2_Spotted             TGLTTTQVSNWFKNRRQRDRA......AE.....A.KE..RENNENSNSNS.........
SIX3_Spotted             TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RLQHQAIGQNG.........
SIX4_Spotted             TGLSLTQVSNWFKNRRQRDRN......PSEAQSKS.ESDGNHSTEDESSKGQDDLSPRPL
SIX5_Spotted             TGLSLTQVSNWFKNRRQRDRT......PSGTNSKS.ESDGNHSTEDEASRGLEDAGMAPA
SIX6_Spotted             TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RLQQQVLSQGS.........
SIX7_Spotted             TGLTPTQVGNWFKNRRQRDRA......AS.....A.KN..R...................
Spotted_Gar_SIX1_like    TGLTTTQVSNWFKNRRQRDRA......SS.....S.RDRERESSGQPCGSQ.........
SIX1_Callorhinchus       TGLTTTQVSNWFKNRRQRDRA......AE.....A.KE..RENNENTNSSTNKQNQL...
SIX2_Callorhinchus       TGLTTTQVSNWFKNRRQRDRA......SE.....T.KE..RYGEGNERVLS.........
SIX3_Callorhinchus       TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RS..................
SIX4a_Callorhinchus      TGLSLTQVSNWFKNRRQRDRNP.....CE.....T.HS..KSESDGNHSTE.....DESS
SIX6_Callorhinchus       TGLTPTQVGNWFKNRRQRDRA......AA.......AKNSRLQQQVLSQAS.........
SIX1_Erpetoichthys       TGLTTTQVSNWFKNRRQRDRA......AE.....A.KE..RENSENNNSGGSKQNQL...
SIX2a_Erpetoichthys      TGLTTTQVSNWFKNRRQRDRA......AE.....A.KE..RYEENNENSNS.........
SIX3_Erpetoichthys       TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RLQHQAIGQNG.........
SIX4_Erpetoichthys       TGLSLTQVSNWFKNRRQRDRN......PSEAQSKS.ESDGNHSTEDESSKGQDDPSPRPL
SIX5_Erpetoichthys       TGLSLTQVSNWFKNRRQRDRTPSGTHSKS.....E.SDGNRSESDDSTRGLEDAAHLTPV
SIX6_Erpetoichthys       TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RLQQQVLSQGS.........
SIX7_Erpetoichthys       TGLTPTQVGNWFKNRRQRDRA......AS.....A.KN..RSVLQQNASLV.........
SIX1_HUMAN               TGLTTTQVSNWFKNRRQRDRA......AE.....A.KE..RENTENNNSSS.........
SIX2_HUMAN               TGLTTTQVSNWFKNRRQRDRA......AE.....A.KE..RENNENSNSNSHN.......
SIX3_HUMAN               TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RLQHQAIGPSG.........
SIX4_HUMAN               TGLSLTQVSNWFKNRRQRDRN......PSETQSKS.ESDGNPSTEDESSKGHEDLSPHPL
SIX5_HUMAN               TGLSLTQVSNWFKNRRQRDRT......GAGGGAPC.KS..ESDGNPTTEDESSRSPEDLE
SIX6_HUMAN               TGLTPTQVGNWFKNRRQRDRA......AA.....A.KN..RLQQQVLSQGS.........
```

```
                          230       240       250       260
                            .         .         .         .
SINE_Amphimedon          SASLSLELHSPKEGLSVTNELPDDPINPSDLSIKEEIS....S.................
SciSixB                  SMDEMYTADSSGGSHLTSQPQSSSHHARQQQQQ...........................
SciSixc                  DSVTSGDFPTALDSAFQSDECPEEEEEDGDDDEAMTAPHH....................
LCOSixB                  GQLNDVYRDRDHMGPPPPSDHERDWHRPDKLTL...........................
LCOSixc                  DFAEAHGLDSAFQNESGGDECDDEDNDVDAEEGDDTQQ......................
Nvec_Six1                ............................................................
Nvec_Six3                ............................................................
Nvec_Six4                ............................................................
Cwil_SIXC                ............................................................
Cwil_SIXA                ............................................................
Cwil_SIXB                ............................................................
Tri_SIX3                 ............................................................
Tri_SIX1A                ............................................................
C_elegans_ceh_33         ............................................................
C_elegans_ceh_32         ............................................................
C_elegans_ceh_34         ............................................................
C_elegans_unc_39         ............................................................
Sine_Schmidtea           TNHENAIKKIQSY...............................................
Six3_Schmidtea           ............................................................
SIX2_Brachionus          SDVKAD......................................................
SIX3_Brachionus          ............................................................
SIX1_Brachionus          ............................................................
SIX1_Pomacea             ............................................................
SIX6_Pomacea             ............................................................
SIX4_Pomacea             ............................................................
SINE_Capitella_teleta    SGLEPKQ.....................................................
Optix_Capitella_teleta   ............................................................
SIX4_Capitella_teleta    HMHVQQG.....................................................
Sine_Drosophila          ............................................................
Optix_Drosophila         RRADGAA.....................................................
Six4_Drosophila          ............................................................
So_H_erato               ............................................................
Optix_H_erato            ............................................................
SIX4_H_mel               ............................................................
SINE_Daphnia             LDESGSE.....................................................
SIX4_Daphnia             SSSGIGLH....................................................
OPTIX_Daphnia            AGGGPTSNKRHRPDHSPNGNDSDDDVDLLMDDANGSGP....L.................
SIX1_Strongylocentrotus  STSSEED.....................................................
SIX6_Strongylocentrotus  ............................................................
SIX4_Strongylocentrotus  RIGPDGGLHGVNRDMRPSNGLDPMLGHHQSGSAITVISTS....................
Six1/2_Halocynthia       PHPGGMEGPASVGHVGSNHGDMME....................................
Six3/6_Halocynthia       ICAEQRRFDERFNVTQLRHS........................................
Six4/5_Halocynthia       ............................................................
Six1/2_Branchiostoma     ............................................................
Six4/5_Branchiostoma     SHDLSQLT....................................................
Six1_Petromyzon          ............................................................
SIX6_Petromyzon          ............................................................
SIX4_Petromyzon          LNLSPPTEADGNTMSNNQEVESLLEEVFSSDSSFSNQMTPLL..................
Six1_like_PetromyzonX1   AGGLVL......................................................
Six1_like_Petromyzon     GSQQNQLTS...................................................
Six2_like_PetromyzonX1   DEEGEDGHEDDGDAGRRRR.........................................
Six6_like_Petromyzon     ............................................................
Six6_like_Petromyzon2    ............................................................
Six2_like_Petromyzon     TTTAAAAATTTTTTTTASAAASTVAAGGVV..............................
Six6_like_Petromyzon3    ............................................................
SIX1_Rhincodon           ............................................................
SIX2_Rhincodon           GGHNPMN.....................................................
SIX3_Rhincodon           ............................................................
SIX4_Rhincodon           KGQDDMS.....................................................
SIX6_Rhincodon           ............................................................
SIX1b_Spotted            .NKQNQL.....................................................
SIX2_Spotted             ..HNPLT.....................................................
SIX3_Spotted             ............................................................
SIX4_Spotted             SNSSDGMANHSTHPLPGSTDSIIIQQIGEAKMSSSSASSAVFNGNLVSTNAPSVFLNG..
SIX5_Spotted             PQEEPGAAPVLLSPGPPCSSGSPILLNGSFLTTSAQTPLLLNGGSLLPGPAGGVIINGLA
SIX6_Spotted             ............................................................
SIX7_Spotted             ............................................................
Spotted_Gar_SIX1_like    ............................................................
SIX1_Callorhinchus       ............................................................
SIX2_Callorhinchus       ....NPG.....................................................
SIX3_Callorhinchus       ............................................................
SIX4a_Callorhinchus      KGQDDLS.....................................................
SIX6_Callorhinchus       ............................................................
SIX1_Erpetoichthys       ............................................................
SIX2a_Erpetoichthys      NSHNPLT.....................................................
SIX3_Erpetoichthys       ............................................................
SIX4_Erpetoichthys       SSSSDGLASHSSIQLSSDSEGLIIQQLGDAKLPSTSPGAMFNGGLVTASTPTVFLNGSSY
SIX5_Erpetoichthys       SQEDHNTINNNILLSTTSTPCSTSSSMLLNGSLINANT....Q.................
SIX6_Erpetoichthys       ............................................................
SIX7_Erpetoichthys       ............................................................
SIX1_HUMAN               .NKQNQL.....................................................
SIX2_HUMAN               ............................................................
SIX3_HUMAN               ............................................................
SIX4_HUMAN               SSSSDGITNLSLSSHMEPVYMQQIGNAKISLSSSGVLLNG....................
SIX5_HUMAN               RGAAPVSAEAAAQGSIFLAGTGPPAPCPASSSILVNGS....F.................
SIX6_HUMAN               ............................................................
```

```
SINE_Amphimedon          ............................................................
SciSixB                  ............................................................
SciSixc                  ............................................................
LCOSixB                  ............................................................
LCOSixc                  ............................................................
Nvec_Six1                ............................................................
Nvec_Six3                ............................................................
Nvec_Six4                ............................................................
Cwil_SIXC                ............................................................
Cwil_SIXA                ............................................................
Cwil_SIXB                ............................................................
Tri_SIX3                 ............................................................
Tri_SIX1A                ............................................................
C_elegans_ceh_33         ............................................................
C_elegans_ceh_32         ............................................................
C_elegans_ceh_34         ............................................................
C_elegans_unc_39         ............................................................
Sine_Schmidtea           ............................................................
Six3_Schmidtea           ............................................................
SIX2_Brachionus          ............................................................
SIX3_Brachionus          ............................................................
SIX1_Brachionus          ............................................................
SIX1_Pomacea             ............................................................
SIX6_Pomacea             ............................................................
SIX4_Pomacea             ............................................................
SINE_Capitella_teleta    ............................................................
Optix_Capitella_teleta   ............................................................
SIX4_Capitella_teleta    ............................................................
Sine_Drosophila          ............................................................
Optix_Drosophila         ............................................................
Six4_Drosophila          ............................................................
So_H_erato               ............................................................
Optix_H_erato            ............................................................
SIX4_H_mel               ............................................................
SINE_Daphnia             ............................................................
SIX4_Daphnia             ............................................................
OPTIX_Daphnia            ............................................................
SIX1_Strongylocentrotus  ............................................................
SIX6_Strongylocentrotus  ............................................................
SIX4_Strongylocentrotus  ............................................................
Six1/2_Halocynthia       ............................................................
Six3/6_Halocynthia       ............................................................
Six4/5_Halocynthia       ............................................................
Six1/2_Branchiostoma     ............................................................
Six4/5_Branchiostoma     ............................................................
Six1_Petromyzon          ............................................................
SIX6_Petromyzon          ............................................................
SIX4_Petromyzon          ............................................................
Six1_like_PetromyzonX1   ............................................................
Six1_like_Petromyzon     ............................................................
Six2_like_PetromyzonX1   ............................................................
Six6_like_Petromyzon     ............................................................
Six6_like_Petromyzon2    ............................................................
Six2_like_Petromyzon     ............................................................
Six6_like_Petromyzon3    ............................................................
SIX1_Rhincodon           ............................................................
SIX2_Rhincodon           ............................................................
SIX3_Rhincodon           ............................................................
SIX4_Rhincodon           ............................................................
SIX6_Rhincodon           ............................................................
SIX1b_Spotted            ............................................................
SIX2_Spotted             ............................................................
SIX3_Spotted             ............................................................
SIX4_Spotted             ............................................................
SIX5_Spotted             LGDGQTITLSPVAGGTPLLLNGAPVVSKQPSGSADVGLKEQAAAGVPTIVLNAAGAALSL
SIX6_Spotted             ............................................................
SIX7_Spotted             ............................................................
Spotted_Gar_SIX1_like    ............................................................
SIX1_Callorhinchus       ............................................................
SIX2_Callorhinchus       ............................................................
SIX3_Callorhinchus       ............................................................
SIX4a_Callorhinchus      ............................................................
SIX6_Callorhinchus       ............................................................
SIX1_Erpetoichthys       ............................................................
SIX2a_Erpetoichthys      ............................................................
SIX3_Erpetoichthys       ............................................................
SIX4_Erpetoichthys       IQTPGNVLFNG.................................................
SIX5_Erpetoichthys       ............................................................
SIX6_Erpetoichthys       ............................................................
SIX7_Erpetoichthys       ............................................................
SIX1_HUMAN               ............................................................
SIX2_HUMAN               ............................................................
SIX3_HUMAN               ............................................................
SIX4_HUMAN               ............................................................
SIX5_HUMAN               ............................................................
SIX6_HUMAN               ............................................................
```

```
                            270                             280
                              .                               .
SINE_Amphimedon           TPGPSGDF......................PSPSASPVSESSFPK...............
SciSixB                   .PSHGGHDADWNRPSKL..............ALSVHGPSHSPPHAN...............
SciSixc                   .RTSGSRS......................SPVAETPDGEEPSLP...............
LCOSixB                   .SGHQHSS......................EITGTLTSNRLPPPS...............
LCOSixc                   .ARSAASP......................VAVISPSDDGERPLV...............
Nvec_Six1                 .VLTLDKS......................SLMVQDVKKCMMPDF...............
Nvec_Six3                 ..............................QQQGTDLSHCKPPLS...............
Nvec_Six4                 .TIDASSP......................PLMNASFEHDAHHVM...............
Cwil_SIXC                 .DSKEEDM......................MDSINHPDAKVKMED...............
Cwil_SIXA                 ..............................PLAEVWPDEVPQHHY...............
Cwil_SIXB                 .SNGSDDE......................GKAPSPLKDCKPDNS...............
Tri_SIX3                  .QESDRDD......................VESMSGFDSLSDCDP...............
Tri_SIX1A                 ............................................................
C_elegans_ceh_33          .RKTASKL......................SNSFHNPSDLSSYSA...............
C_elegans_ceh_32          .SDMSDSD......................DDFEDSMTDSPSPID...............
C_elegans_ceh_34          ..............................SDDGSSGCESKPPMN...............
C_elegans_unc_39          ..............................SAKCSPPSSSSSTNG...............
Sine_Schmidtea            .PIGVGNP......................SLMFPGFNESVSPFS...............
Six3_Schmidtea            .DEIAEDD......................DEIAYESLEIVPKTH...............
SIX2_Brachionus           .FGSSGEM......................APRNMVSSTSPSNDT...............
SIX3_Brachionus           ..............................SQTASSSNHNYSPPS...............
SIX1_Brachionus           ..............................STSSSSSSNSSSNVN...............
SIX1_Pomacea              .LGSQQDP......................MSPTTDIKQEKSPDS...............
SIX6_Pomacea              .SGGEGSR......................ADTLSPSSTATDRRI...............
SIX4_Pomacea              ..............................PISQLTMCTTLTAET...............
SINE_Capitella_teleta     .EPDSPDL......................NGQTKLDLDDSPPPP...............
Optix_Capitella_teleta    ............................................................
SIX4_Capitella_teleta     .HSGGGMQ......................AMQAMNFGMNLGMNH...............
Sine_Drosophila           .SEMEGSM......................LPSQSAQHQQQQQQQ...............
Optix_Drosophila          .SPTPSDS......................SDSDISLGTHSPVPS...............
Six4_Drosophila           ..............................DIMSVLPVGQLDGNG...............
So_H_erato                .SGGTGDK......................QLDSSTDDDSDAPHA...............
Optix_H_erato             .....SST......................YDEDSADSEINVDEE...............
SIX4_H_mel                ............................................................
SINE_Daphnia              .SLDGDGP......................DVVMATSSASARHSA...............
SIX4_Daphnia              .HHHGGND......................LLGELKPSNFAMSGN...............
OPTIX_Daphnia             SPGAESLI......................DSDNSSISLTGPPSP...............
SIX1_Strongylocentrotus   .LPMNGKE......................NLGETTSVDMGVNHM...............
SIX6_Strongylocentrotus   ..............................NSSVSNSAHSPSLTE...............
SIX4_Strongylocentrotus   .SGGGNTT......................TDLLHPQAVKVEPPE...............
Six1/2_Halocynthia        .QKPGLES......................SMHPHSHHQQLPPHV...............
Six3/6_Halocynthia        .SPSNNSD......................SSDASDRKERKSSDE...............
Six4/5_Halocynthia        .RLSLGDI......................GTDLANSCRTRRCTP...............
Six1/2_Branchiostoma      .VGQPGED......................QVDRNPGTELQHEDS...............
Six4/5_Branchiostoma      .ARVDKMD......................SSIATTPVHAQQQQA...............
Six1_Petromyzon           .SPLNGAR......................SLLSSSDDEKSPDHT...............
SIX6_Petromyzon           ..............................PIALCSSERAAHPDS...............
SIX4_Petromyzon           .SPCMSDH......................DPIDPLDDSPFPDMESGAGPVGDANQSERD
Six1_like_PetromyzonX1    .SPLHGGR......................GLLSSDEERSPAPSP...............
Six1_like_Petromyzon      .PSLNGSK......................HLLASSDEDKSPVGT...............
Six2_like_PetromyzonX1    .VGPRHGH......................AGALAPEGSPRPPSA...............
Six6_like_Petromyzon      .PLLSESP......................SPPLLSFAARDSPAG...............
Six6_like_Petromyzon2     ....ARFL......................PERAESPGERSPPSS...............
Six2_like_Petromyzon      .ARSAGGG......................SLSGTEDEEGSTPCQ...............
Six6_like_Petromyzon3     .VGLREDE......................EDDDDDDDDEEDDDE...............
SIX1_Rhincodon            .SPINRSK......................NLMSSSDEELSPPQS...............
SIX2_Rhincodon            .HSMNGNK......................NMLASSDEEKTPSQT...............
SIX3_Rhincodon            .VCTLSEP......................GCPAHSAAESPSTAA...............
SIX4_Rhincodon            .PRPLSNP......................SDGMNHSNVHSPPEG...............
SIX6_Rhincodon            ..............................VRSLSEEDAAVEPLV...............
SIX1b_Spotted             .SPLDGGK......................SLMSSSEDEFSPPQS...............
SIX2_Spotted              .SSMNGNK......................TVLGSSDDDKTPSGT...............
SIX3_Spotted              .MRSLSES......................GCTPHSSAESPSTAA...............
SIX4_Spotted              .SSYLQAPSNILFNGLNLGGSQPITLNALRPTNSLVSSDSANGET...............
SIX5_Spotted              PSSEAGESAGADLPSVEYGGSLAVQESGNLKTAISEPGTSSPSTA...............
SIX6_Spotted              .VRSLAEE......................DSAVDRLGAASSPEA...............
SIX7_Spotted              ............................................................
Spotted_Gar_SIX1_like     .QRSPADLG.....................PGYPSSDDDLSPPQS...............
SIX1_Callorhinchus        .SPINRNK......................NLMSSSDEELSPPQS...............
SIX2_Callorhinchus        .PPFAGEG......................ACGSVNNSARFPPNS...............
SIX3_Callorhinchus        ............................................................
SIX4a_Callorhinchus       .PRPLSNP......................PDGMNHSSIQAHSEG.........MFIQHI
SIX6_Callorhinchus        ..............................VRSLSEEEAAVGPMV...............
SIX1_Erpetoichthys        .SPLDGAK......................SLMSSSEDDFSPPQS...............
SIX2a_Erpetoichthys       .SSMNGNK......................TVLGSSDDEKTPSGT...............
SIX3_Erpetoichthys        .MRSLSES......................GCTPHSSAESPSTAA...............
SIX4_Erpetoichthys        .LNMGGTQ......................AVALNPLRNSNPIVN...............
SIX5_Erpetoichthys        PFLFNGGS..LVQASNGRVIINGLTFSDGQTITLSPVSSNPPLLVTGSTVIGSKPVTGTQ
SIX6_Erpetoichthys        .VRSLTED......................EAAVDRLGAASSPEA...............
SIX7_Erpetoichthys        ..............................SSSVSLDCSNRDCHP...............
SIX1_HUMAN                .SPLEGGK......................PLMSSSEEEFSPPQS...............
SIX2_HUMAN                .PLNGSGK......................SVLGSSEDEKTPSGT...............
SIX3_HUMAN                .MRSLAEP......................GCPTHGSAESPSTAA...............
SIX4_HUMAN                .SLVPASTSPVFLNGNSFIQGPSGVILNGLNVGNTQAVALNPPKM...............
SIX5_HUMAN                LAASGSPAVLLNGGPVIINGLALGEASSLGPLLLTGGGGAPPPQP...............
SIX6_HUMAN                ..............................GRALRAEGDGTPEVL...............
```

```
SINE_Amphimedon           ............................................................
SciSixB                   ..........................................................SS
SciSixc                   .................................................PMDQLMLTEHM
LCOSixB                   ............................................................
LCOSixc                   ........................IDHHPLRRSTSSDLPSMEELMLVDMMTMTSQDQEVN
Nvec_Six1                 ............................................................
Nvec_Six3                 ............................................................
Nvec_Six4                 ............................................................
Cwil_SIXC                 ............................................................
Cwil_SIXA                 ............................................................
Cwil_SIXB                 ............................................................
Tri_SIX3                  ............................................................
Tri_SIX1A                 ............................................................
C_elegans_ceh_33          ............................................................
C_elegans_ceh_32          ............................................................
C_elegans_ceh_34          ............................................................
C_elegans_unc_39          ............................................................
Sine_Schmidtea            ............................................................
Six3_Schmidtea            ............................................................
SIX2_Brachionus           ............................................................
SIX3_Brachionus           ............................................................
SIX1_Brachionus           ............................................................
SIX1_Pomacea              ............................................................
SIX6_Pomacea              ............................................................
SIX4_Pomacea              ............................................................
SINE_Capitella_teleta     ............................................................
Optix_Capitella_teleta    ............................................................
SIX4_Capitella_teleta     ............................................................
Sine_Drosophila           ............................................................
Optix_Drosophila          ............................................................
Six4_Drosophila           ............................................................
So_H_erato                ............................................................
Optix_H_erato             ............................................................
SIX4_H_mel                ............................................................
SINE_Daphnia              ............................................................
SIX4_Daphnia              ............................................................
OPTIX_Daphnia             .................VGDMTSAGALGSAGNDAASILKSSIGGDMGRGGQPSNPLTMST
SIX1_Strongylocentrotus   ............................................................
SIX6_Strongylocentrotus   ............................................................
SIX4_Strongylocentrotus   ............................................................
Six1/2_Halocynthia        ............................................................
Six3/6_Halocynthia        ............................................................
Six4/5_Halocynthia        ............................................................
Six1/2_Branchiostoma      ............................................................
Six4/5_Branchiostoma      ............................................................
Six1_Petromyzon           ............................................................
SIX6_Petromyzon           ............................................................
SIX4_Petromyzon           ADAAAAAAAQTHAAAATATTGKSGRLAGHVAAGAANPTLRFKTACADNNATFFSTAQVVG
Six1_like_PetromyzonX1    ............................................................
Six1_like_Petromyzon      ............................................................
Six2_like_PetromyzonX1    ............................................................
Six6_like_Petromyzon      ............................................................
Six6_like_Petromyzon2     ............................................................
Six2_like_Petromyzon      ............................................................
Six6_like_Petromyzon3     ............................................................
SIX1_Rhincodon            ............................................................
SIX2_Rhincodon            ............................................................
SIX3_Rhincodon            ............................................................
SIX4_Rhincodon            ............................................................
SIX6_Rhincodon            ............................................................
SIX1b_Spotted             ............................................................
SIX2_Spotted              ............................................................
SIX3_Spotted              ............................................................
SIX4_Spotted              ...................................VLHSSDEKDYKVLNGPVTNSAVPYN
SIX5_Spotted              ..............PSLVLAQPAPSCTADAQLPLAPVSEAEPTLQQQAARQLSQDAQVSS
SIX6_Spotted              ............................................................
SIX7_Spotted              ............................................................
Spotted_Gar_SIX1_like     ............................................................
SIX1_Callorhinchus        ............................................................
SIX2_Callorhinchus        ............................................................
SIX3_Callorhinchus        ............................................................
SIX4a_Callorhinchus       GDLKSTPSSSGILLNGNLVTTNGTPVFCNGSSFIQGPSGVLVNGLPLGNAQTISLSPVGT
SIX6_Callorhinchus        ............................................................
SIX1_Erpetoichthys        ............................................................
SIX2a_Erpetoichthys       ............................................................
SIX3_Erpetoichthys        ............................................................
SIX4_Erpetoichthys        ....................................SRAENGELVLQDKDMKIHGPNSIM
SIX5_Erpetoichthys        QANNIEQKDPAVASILPTIIVNTGSTTLSLPAVGDGVGKSEEDEQSMPTSLVYGNTSNLG
SIX6_Erpetoichthys        ............................................................
SIX7_Erpetoichthys        ............................................................
SIX1_HUMAN                ............................................................
SIX2_HUMAN                ............................................................
SIX3_HUMAN                ............................................................
SIX4_HUMAN                .......................SSNIVSNGISMTDILGSTSQDVKEFKVLQSSANSATT
SIX5_HUMAN                ................................SPQGASETKTSLVLDPQTGEVRLEEAQS
SIX6_HUMAN                ............................................................
```

```
                                                          290       300       310
                                                           .         .         .
SINE_Amphimedon             ..............................LTEVYRSAEDSFITPESIGQTDSSSIAVSP
SciSixB                     HNNAHHNHQQQQQHNGYSSSYSPAASNGGGPACGYPAANDNSYTQLHTGHHIHGRPASNS
SciSixc                     LRGQDSEVDELIHRPSDSTLPTVKSRLNSTDSAVNVGDSPFEISSILASSYQPRSNAAYP
LCOSixB                     ..........AHIYEDSYSSTTSGQYNNHFSQMSSSLRVPMQSDPACSADYAPEAKTYPV
LCOSixc                     DYILHAKQARAGDHKARLPSVDSAVHVIGSPVEFSTIPQEYFHQSSTTGAPYGHHTGLAS
Nvec_Six1                   ............................................................
Nvec_Six3                   ...........................................................P
Nvec_Six4                   ....................................................LPSLMKTE
Cwil_SIXC                   .....................................................ETPQFIM
Cwil_SIXA                   ............................................................
Cwil_SIXB                   ................................FIIRSDQIHQINLVQPHIIMSQQSYMGI
Tri_SIX3                    ............................................................
Tri_SIX1A                   ............................................................
C_elegans_ceh_33            ............................................................
C_elegans_ceh_32            ....................................EPKDLSKSHIPKLSPTLLPKMATP
C_elegans_ceh_34            ............................................................
C_elegans_unc_39            ............................................................
Sine_Schmidtea              .....................................TTAASIQGGYCSPYIYNSNPAAN
Six3_Schmidtea              ...................................SKASSIISTNWSTVDNADDTNINNI
SIX2_Brachionus             .........................AAFGMHQQHQSNTSSSSSSSSNNNYQFLNINDSYA
SIX3_Brachionus             ............................................................
SIX1_Brachionus             ............................................................
SIX1_Pomacea                ....................................................INGQGKGE
SIX6_Pomacea                .........................................................EPD
SIX4_Pomacea                ............................................................
SINE_Capitella_teleta       ........................................IRSGMFSDMPKQSHAPQAAM
Optix_Capitella_teleta      ............................................................
SIX4_Capitella_teleta       ......................HYSDADLCRKYNFVAGSSPQKGMMTHDDPRSAGLGDMT
Sine_Drosophila             ....................................QHSPGNSSGNNNGLHQQQLQHVAA
Optix_Drosophila            ..................................SLQLQHSPGSTSNGANDREESLSVDD
Six4_Drosophila             ............................................................
So_H_erato                  ............................................................
Optix_H_erato               ............................................................
SIX4_H_mel                  ............................................................
SINE_Daphnia                ..........................................ALLDTPSPPPPHHIQHHH
SIX4_Daphnia                .............................KMLFDDVVHNPSGSSGLMAASRSMSSLMASP
OPTIX_Daphnia               GGLLSDPIHHSRDREQHHAAAMALSSAGLMAPAPPSHHHHHHPSAFTFGHHPHHHPSFGA
SIX1_Strongylocentrotus     ...........................................GHVGHHHPHHTHGHQHS
SIX6_Strongylocentrotus     ............................................................
SIX4_Strongylocentrotus     ...................PLDENRGVPLNMHEPPDLANLTDATGLQAFKFVPEMLLPQM
Six1/2_Halocynthia          .........................................QQQIMQHHSDQMVQQQVHH
Six3/6_Halocynthia          ..................................VGTRSASSPELDTSSTMSMSPIPYTH
Six4/5_Halocynthia          ...................................DANKMMNESDAFSKIQMNKSLSSSE
Six1/2_Branchiostoma        .....................................................KQNIFTS
Six4/5_Branchiostoma        ..............................................QLQHHHSAHAHHMI
Six1_Petromyzon             ...............................PDHGGVSPPLLLQAPHHHHPHHQHHQSLS
SIX6_Petromyzon             ...........................................................R
SIX4_Petromyzon             GGGGVAFQPLTQFASQTVTPVASLLSMSPAGPMYLPPLQVTGATHHQPVQFVPYSPVQMA
Six1_like_PetromyzonX1      .........................................................AHS
Six1_like_Petromyzon        ........................................PDRAHVSSGSSHHAAVGPGS
Six2_like_PetromyzonX1      ............SWSWSGAGPRPPPWDPRAAPSPGPAKDGGEGGWPLARGSAGVAAGDAA
Six6_like_Petromyzon        ............................................................
Six6_like_Petromyzon2       ........................................................WGPR
Six2_like_Petromyzon        .............................................SPLSHREAAGPFHSR
Six6_like_Petromyzon3       ...........................................................E
SIX1_Rhincodon              ............................................................
SIX2_Rhincodon              ............................................................
SIX3_Rhincodon              ............................................................
SIX4_Rhincodon              ................................MFMHHIGELKPSQNSSGILLNGNLVTTN
SIX6_Rhincodon              ............................................................
SIX1b_Spotted               ............................................................
SIX2_Spotted                ............................................................
SIX3_Spotted                ............................................................
SIX4_Spotted                MSTLGSSFPATIHASEVKMEGLQTLASQDGSSLVTFSTSNGPLHLSQYSLVHIPTADTNG
SIX5_Spotted                SPQVLSLPQVVPSLQNIPVSQIVQAHASQVSACPQIVPISQLPQTSISHLPAQSFQVAPR
SIX6_Spotted                ............................................................
SIX7_Spotted                ............................................................
Spotted_Gar_SIX1_like       ............................................................
SIX1_Callorhinchus          ............................................................
SIX2_Callorhinchus          ............................................................
SIX3_Callorhinchus          ............................................................
SIX4a_Callorhinchus         TPSVLVNNISNGSLGATELKTESIHMLNSQDLGSSGSVDNGPVQINQYGLVHLHNSENNV
SIX6_Callorhinchus          ............................................................
SIX1_Erpetoichthys          ............................................................
SIX2a_Erpetoichthys         ............................................................
SIX3_Erpetoichthys          ............................................................
SIX4_Erpetoichthys          SYNVSLPGAPFPVSVSSSEIKIGSTQTVSSQDGASVLTFTTCNGPLQVNQYSVVQLPTCD
SIX5_Erpetoichthys          ASNGMDLKIESLQPTMETCPTPSTTSSVVFNQQGTTLTIPGVPGEFRVDEQQALHSGSPS
SIX6_Erpetoichthys          ............................................................
SIX7_Erpetoichthys          ............................................................
SIX1_HUMAN                  ............................................................
SIX2_HUMAN                  ............................................................
SIX3_HUMAN                  ............................................................
SIX4_HUMAN                  TSYSPSVPVSFPGLIPSTEVKREGIQTVASQDGGSVVTFTTPVQINQYGIVQIPNSGANS
SIX5_HUMAN                  EAPETKGAQVAAPGPALGEEVLGPLAQVVPGPPTAATFPLPPGPVPAVAAPQVVPLSPPP
SIX6_HUMAN                  ............................................................
```

```
                           320       330       340       350
                            .         .         .         .
SINE_Amphimedon            DNTKYYPHQSGTYPSTSDASIASSYYVSYNGTS..........................Y
SciSixB                    TTDYHQNEVKRAYVPETNAASSAAAPMVNHRFY..........................T
SciSixc                    MHSCGTAPVYPHYQSLASLACMSHGNYTTSGHS..........................M
LCOSixB                    EQPPATTPLANHHSYAEDQRTDSSLLTPSGQHMVLSS......................A
LCOSixc                    SQGAVSSTGSGMLGQTHGRGLTPTSFHLTSAQP..........................P
Nvec_Six1                  ......SHCIAQNIPVNVKMEVDDPTLMAVGHG..........................D
Nvec_Six3                  DELSLDDECSDLRELKNTPVSSPNITKLSEQRP...........................
Nvec_Six4                  ESLMNASLFTPVSLTDPSVPVPSQIIVCSAEHS..........................I
Cwil_SIXC                  LEPPVVSQLPQLIMAPHPPETEATDPSNEKQRQ..........................A
Cwil_SIXA                  ....TYTQFPTHNTVLAPSLMPHSPFLSYSEPI...........................
Cwil_SIXB                  LNQHLLHPVFTHEQRNSTDIDHAMRNSYHGSGY..........................R
Tri_SIX3                   ...........................................................T
Tri_SIX1A                  ............................................................
C_elegans_ceh_33           ..AAAAATFPGFYMNYNDMMIGAGTSYQSL..............................
C_elegans_ceh_32           FDMFAAAANPLMMLNLNPALYMQFHNFFNTMRN..........................P
C_elegans_ceh_34           ...........IDSPAPPPLPTSFDLQPY..............................Y
C_elegans_unc_39           .....GSDFLPIITPQSFNLAAAPFNMNMIYGT..........................L
Sine_Schmidtea             YISYFSSGLHNLPISGQYPNILSRESMANTDTS..........................Y
Six3_Schmidtea             INNNETDKNKPIAVIRNLVKDSRKTFNVADILD..........................T
SIX2_Brachionus            YTAQYQAPFYPSSSPLQHNQFGSYYTANYGQHHYG........................N
SIX3_Brachionus            ...........................................................S
SIX1_Brachionus            .NCFGYSSSSSAYNYGNVAKRSRLSMFNDDHV............................
SIX1_Pomacea               DTSPVQLHHPSVYTEMQKAASAAMQGMPGHGSM..........................V
SIX6_Pomacea               IDLEDDDSFSDLDSLTSSPPPHDSHHDDDRVQL..........................S
SIX4_Pomacea               ............................................................
SINE_Capitella_teleta      HPHHHGNSIPPPMLLQHPSSSLHHSQHPPGGHH..........................S
Optix_Capitella_teleta     ............................................................
SIX4_Capitella_teleta      MNLVKTEALTPPYMYTGMGAEHHHGHHQAEAHGG.........................I
Sine_Drosophila            EQGLQHHPHQPHPASNIANVAATKSSGGGGGGG..........................V
Optix_Drosophila           DKPRDLSGSLPLPLSLPLPLASPTHTPPQLPPG..........................Y
Six4_Drosophila            ...........................................................F
So_H_erato                 ..GAHAAPLYPLYE.............................................H
Optix_H_erato              ............................................................
SIX4_H_mel                 ...GSQGGLAQSFIPNAYYKLQDASHYLHGNPP...........................
SINE_Daphnia               HHHHQQQQQQQINNMAAALANSLPHHNSQDSL...........................I
SIX4_Daphnia               FVGGGKDSHVATSSLLHLPSTAFASYHPSGGHG..........................Y
OPTIX_Daphnia              NSLMAAAAGLHFNLAASLGNAPLTPFGAFGSLS..........................N
SIX1_Strongylocentrotus    HIHPHPHPHPHPHSHSHSHSHPHAHSHSHGPAG..........................L
SIX6_Strongylocentrotus    GATCLLSPHPDDHSPSPRSLADSPPLDSP..............................S
SIX4_Strongylocentrotus    LPLALASSKGDINMNHAAQLATMSMIMNPALLS..........................N
Six1/2_Halocynthia         QQHSHTPPQAHHHYPQGPSPPSIPNSMIMGTMQ..........................D
Six3/6_Halocynthia         HPSPFVSGSFGPISSEPYFTASSYFSNTFPFFA..........................R
Six4/5_Halocynthia         YLACSYSPSQPTIKQNPLSLHLNSNHGILFSHA..........................S
Six1/2_Branchiostoma       MQNPGQNSVTPTDMNTPMGPLAPSGPGPAGPSA..........................V
Six4/5_Branchiostoma       GLDYPGDVETKSCVPTPHELLGSCSAVETSVLC..........................T
Six1_Petromyzon            HQHHQHQALAQHQHQSHMSQHGGGALSYAGLPP..........................N
SIX6_Petromyzon            GYGAGGEGRQLSESPGAVASPGATSLSSLSDRP..........................A
SIX4_Petromyzon            YPAVVNGGAMPVVSGMGLPTIQLPSISPLQTAAGNILVSNTLAGGDLLNGNTMATSTTGI
Six1_like_PetromyzonX1     TPNDTPPRLLGHHPAAAAAAAGHGEHHQHHHHQ..........................Q
Six1_like_Petromyzon       ISFPGLSGLGAAASPGTAAVLLSGADVLHQHHH..........................H
Six2_like_PetromyzonX1     FATAMHGPGPTLSPTASIAVLPPPLLGLSPGRA..........................V
Six6_like_Petromyzon       .SPGLALSCSPAGSSSSSCRAQRSLTAPAT.............................S
Six6_like_Petromyzon2      GGGGGVGGGGSAVSPRPSPRASLSGLSERDEAL..........................P
Six2_like_Petromyzon       GHHDGLAAAAAALEREGYALMAASGHEGDGHGG..........................G
Six6_like_Petromyzon3      EEEAGAGPGRGLRSSSGAAAGSPGAASLSSLGE..........................P
SIX1_Rhincodon             ..PDHCSGSPVLLLPGSVNQSVDSTFSLHGLSS..........................S
SIX2_Rhincodon             ..PDHTSSSPALLLGSLGQSSNTSLQPLHTLAA..........................P
SIX3_Rhincodon             ......SPTTSVSSMTERAETATS...................................I
SIX4_Rhincodon             GSPVFYSGSSFIQGPNGVLVNGLSLGNAQTISL..........................S
SIX6_Rhincodon             ...GASSPAASLCSVSERSKAATS...................................A
SIX1b_Spotted              .....PDQNSVLLLQGNMTHPGNSSYPLTGLSA..........................S
SIX2_Spotted               ..PDHTSSSPALLLTSNSGLPPLHGLAPPPGPSA.........................I
SIX3_Spotted               ......SPTTSVSSMTERVETGTS...................................I
SIX4_Spotted               PLVNGNIGLPQLQMPPVSTAPSHGNVLLHNATG..........................A
SIX5_Spotted               MPQPQQGLSLQLGEPLSPAPTLQTLPPAQTLQVSGTQIIPISPPTQVVPLSQPGQVSPVA
SIX6_Spotted               ........................SLSGKAVAS..........................A
SIX7_Spotted               ............................................................
Spotted_Gar_SIX1_like      .........PRLLYPSLGPPPPHRQHLPP..............................A
SIX1_Callorhinchus         ..PDHCSGSPVLLLPGSLNQSVDSTFSLHGLSS..........................S
SIX2_Callorhinchus         .EDDETPSRASSSSPGGGMMVVAGPFEPGSLQSL.........................V
SIX3_Callorhinchus         ............................................................
SIX4a_Callorhinchus        QLVNGNIGLTSLQLPSVSAASSQGNILITNTSD..........GGTLLSGTAATLQQGKV
SIX6_Callorhinchus         ...GASSPAASLCSVSERSKAATS...................................A
SIX1_Erpetoichthys         .....PDQNSVLLLQGNMNHSGGSTYPMSSLQG..........................H
SIX2a_Erpetoichthys        ..PDHTSPSPALLLTTNSGLQTLHGLAPPQGPSA.........................I
SIX3_Erpetoichthys         ......SPTTSVSSMTERVETGAS...................................I
SIX4_Erpetoichthys         TNGQQVNGLPSFLMPSITTAPPQGNPLMNSTHPEQTEAFSSGQSSTPNLQHGKLLLSPLQ
SIX5_Erpetoichthys         TPTLQTSQVVPLPSQQQATVLTSSANQLASAPQVLALPQVVPS................I
SIX6_Erpetoichthys         ........................SLSSKAAAS..........................A
SIX7_Erpetoichthys         ....HLQTGSPSHPGSPETSDCSMSTEQRETST..........................P
SIX1_HUMAN                 .....PDQNSVLLLQGNMGHARSSNYSLPGLTA..........................S
SIX2_HUMAN                 ..PDHSSSSPALLLSPPPPGLPSLHSLGHPPGPSA........................V
SIX3_HUMAN                 ......SPTTSVSSLTERADTGTS...................................I
SIX4_HUMAN                 QFLNGSIGFSPLQLPPVSVAASQGNISVSSSTSDGSTFTSESTTVQQGKVFLSSLAPSAV
SIX5_HUMAN                 GYPTGLSPTSPLLNLPQVVPTSQVVTLPQAVGPLQLLAA....................G
SIX6_HUMAN                 ....GVATSPAASLSSKAATS......................................A
```

```
                               360       370
                                 .         .
SINE_Amphimedon            PTAAATPAAYNSTSLPNP......................................DMYS
SciSixB                    PDTGSAPHHSGDLMSPAPTAVPLEVAATPDPDADHYVQIVPENVGRFEPLKHRSPADAPY
SciSixc                    AAFPLNNSTQPPPMYPSPPSTQMYPQRMTSPLFNASWSSYMNHMPQQSTSMSQNTSPPPF
LCOSixB                    PAVVPNDPESDRYVRIVHPSAGRFSPLLTHSGADTPAYPYPSLSEPVYHMTAQTPTHVHG
LCOSixc                    AYPSAAAPAMHQLYQQRMSSPLFNVSWASYANSYHGTGHAEYQSSHYSSSLKYPSLSHEF
Nvec_Six1                  SSVSASELSQGLQD..............................................
Nvec_Six3                  ............................................................
Nvec_Six4                  PSSDNLAQESGHGVVKKEEILQ......................................
Cwil_SIXC                  SNGSQNEMDLKPSDM.............................................
Cwil_SIXA                  ............................................................
Cwil_SIXB                  KMIFEQGMGNHDSYFQNI..........................................
Tri_SIX3                   IDDDPRRK....................................................
Tri_SIX1A                  ............................................................
C_elegans_ceh_33           ............................................................
C_elegans_ceh_32           QIDEEENSETTVEVEADI....................................EPPKKR
C_elegans_ceh_34           PSPYTFAPHCDFSYIQNL..........................................
C_elegans_unc_39           RDSQSDNDQFTFNP..............................................
Sine_Schmidtea             MNPAKSEISSKALKLNQV...................................SECEDIE
Six3_Schmidtea             RKGEFLERKSEFGFNPIS..........................................
SIX2_Brachionus            LNGNGADLFQQNQYFSSV...................................DEQDYRL
SIX3_Brachionus            PELDPDVTD...................................................
SIX1_Brachionus            ............................................................
SIX1_Pomacea               PHGMMQSQGHGGHHQGVG..........................................
SIX6_Pomacea               PTPSNSSHHGPRPTS.............................................
SIX4_Pomacea               ............................................................
SINE_Capitella_teleta      PNVPGGGGGAHLGLQGMA..........................................
Optix_Capitella_teleta     ............................................................
SIX4_Capitella_teleta      PSNQTELKDWKSLSYDEIHSERDIRRKFERLYKK......................TGQE
Sine_Drosophila            SAAAAAQMQMPPLTAAVA..........................................
Optix_Drosophila           GGGAGAGPGGPLTGPGCLPPFKLDAATSLFSAGCYLQSFSNLKEMSQQFPIQPIVLRPHP
Six4_Drosophila            PRMFNAPSYYPETIFNGQ..........................................
So_H_erato                 PLAHLQYHHT..................................................
Optix_H_erato              ............................................................
SIX4_H_mel                 ............................................................
SINE_Daphnia               GGGGSIGDLYPDLKLASV......................................YHAT
SIX4_Daphnia               DSGSVATLANLHVAHQHV..........................................
OPTIX_Daphnia              PFAAAVAPFGSRFNGGSNRPNGSSGGGKPPTSFDMSLAETLGHHHHAAAMAAAAAAAELA
SIX1_Strongylocentrotus    MMSELSKTGYGITNLPPG..........................................
SIX6_Strongylocentrotus    PAGYHDDDDDDDL...............................................
SIX4_Strongylocentrotus    PALNLSTANNNNNIMPGMASIASLQQHQQHHHHHHPELHHQQQQQQHEQQDMSVHNVATT
Six1/2_Halocynthia         PHSVQHNLHSAAMSGHYT.....................................MASSA
Six3/6_Halocynthia         PHHDVTQNVCDVTIKSEP..........................................
Six4/5_Halocynthia         PAMDANQLIAVDALQHLH..........................................
Six1/2_Branchiostoma       AVSNSDVLNPQSMQ..............................................
Six4/5_Branchiostoma       PVSTIVKGEPPQ................................................
Six1_Petromyzon            PGVPAPSGGAGTGLQQQQ..........................................
SIX6_Petromyzon            ATACTTSPACSESDCEA...........................................
SIX4_Petromyzon            LITNALAPSTMLCSMPHVSLMTVITQDGSLALTPVFNIPSGIYPDGISTSNVIPCAGNYL
Six1_like_PetromyzonX1     QQQQHHQHHHHQQQQQQH....................................HASAYQ
Six1_like_Petromyzon       HHQQQQHHHQHHHHHQQQ..........................................
Six2_like_PetromyzonX1     PLVSNPGPFLGPACVPAS.......................................QLH
Six6_like_Petromyzon       ASVTDSDSECDVGIGDY...........................................
Six6_like_Petromyzon2      SAASRADSDSSEPDI.............................................
Six2_like_Petromyzon       PLDEGDPQGPHPPPLSCM......................................AMLS
Six6_like_Petromyzon3      PAHPASSDGSDSEA..............................................
SIX1_Rhincodon             PGGHAGPVHPHGLQ..............................................
SIX2_Rhincodon             PGPSAVAVPNSDVIHHHA..........................................
SIX3_Rhincodon             LSVTSSDSECDV................................................
SIX4_Rhincodon             PIGTTSSVLVNGISNGSA......................................GGNE
SIX6_Rhincodon             ISITSSDSDCDV................................................
SIX1b_Spotted              QSVHNIQGHPNQLQ..............................................
SIX2_Spotted               PVPNTDPVHHHTLH..............................................
SIX3_Spotted               LSVTSSDSECDV................................................
SIX4_Spotted               PGDSFSSSSASVPQHDKLVLAPLHPSTVLYTLPCAPPAPAPTAIKQEPAEGGFSFPPGMH
SIX5_Spotted               PAPQIVPLSLPQLVPMSPAVASQSSLSLPQVVPGSTALPLSSGSFQILTTTSNISSAPGS
SIX6_Spotted               ISITSSDSECDI................................................
SIX7_Spotted               ............................................................
Spotted_Gar_SIX1_like      PLLGVDLGDLGDP...............................................
SIX1_Callorhinchus         PGGHSGPVHPHGLQ..............................................
SIX2_Callorhinchus         NGCNGGGVGFESVNVEHQ..........................................
SIX3_Callorhinchus         ............................................................
SIX4a_Callorhinchus        FLTTTLPPSAVMCTIPNSGQAVAQVKHDALEGGLVFSQLMSHNQLNVNASNGNQSGTVLN
SIX6_Callorhinchus         ISITSSDSDCDI................................................
SIX1_Erpetoichthys         PHQLQDS.....................................................
SIX2a_Erpetoichthys        PVPTSDPMHPHTLH..............................................
SIX3_Erpetoichthys         LSVTSSDSECDV................................................
SIX4_Erpetoichthys         PSPVLYSPPSMQQMVASIKQEPPEGGFTFSHLMPVDQNGQISVSTSSVNVSGMAIDTLSP
SIX5_Erpetoichthys         PGIPVSQVIQAPPSQGTTCPQLVPVSPVTTQLNQTAPAFQLPQGIPQQQVVTSAALPHIS
SIX6_Erpetoichthys         ISITSSDSECDI................................................
SIX7_Erpetoichthys         DISVCSDSDFEP................................................
SIX1_HUMAN                 QPSHGLQTHQHQLQ..............................................
SIX2_HUMAN                 PVPVPGGGGADPLQHHHG..........................................
SIX3_HUMAN                 LSVTSSDSECDV................................................
SIX4_HUMAN                 VYTVPNTGQTIGSVKQEGLERSLVFSQLMPVNQNAQVNANLSSENISGSGLHPLASSLVN
SIX5_HUMAN                 PGSPVKVAAAAGPANVHLINSGVGVTALQLPSATAPGNFLLANPVSGSPIVTGVALQQGK
SIX6_HUMAN                 ISITSSDSECDI................................................
```

```
                               380       390       400       410       420       430
                                 .         .         .         .         .         .
SINE_Amphimedon             TPAGSCLSPSYLSSYQAATGKGYTWPTAPNGVGYSAFGMTPDMYQYQAQTYQQMAASRGT
SciSixB                     PYQSANERVYHINGTQPSPMATQNGHESQVSGAPVRQLGPEQQIDHHGYRSSDAAEPGRY
SciSixc                     LPSMTYQPQGMSDFLVGGANPSTDLIDEPGKFCPTSQSELHSVLSPPTSSTHDVCTAMYP
LCOSixB                     SPAQTPLMPGPEQVDGPAFKPPSLVDSRLRDDSYETYHKSSYAQPMGTPTPTGSGAVDVA
LCOSixc                     LSAGVSPPVGLESQALPVTADPGECLVSMPTSIAFSSAALADGLHETSALSAACTPAEVP
Nvec_Six1                   ........................................SHMLTTLTNTLVGL......
Nvec_Six3                   ............................................................
Nvec_Six4                   ............................................................
Cwil_SIXC                   .........................................DMEVPIYTTLVNT......
Cwil_SIXA                   ............................................................
Cwil_SIXB                   .....DLSRELGDMPEPSTSASDLRPGPQTTPWSGRLPGQPGSSQDETTLHRETGKW...
Tri_SIX3                    ............................................................
Tri_SIX1A                   ............................................................
C_elegans_ceh_33            ............................................................
C_elegans_ceh_32            SKLSIDEILNIKSEVSPSQCSPCSNESLSPKRAVKTEEVKKEDDEAAEEDSRSVKSETSE
C_elegans_ceh_34            ............................................................
C_elegans_unc_39            ............................................................
Sine_Schmidtea              PADQTTNLLNQQSYETSNSDYHSMSSTNSSYLNYEKIFNTNITENPLSRTNTNSSTNSSS
Six3_Schmidtea              ............................HPNIPLYLLPHFNNFLQNFANSLPESRNPFVS
SIX2_Brachionus             QMAQANVNFSQANKKSNESMSNSGSPSPAQSISNSSSTSSSPLNLNASNNSLNHQNLGAS
SIX3_Brachionus             ............................................................
SIX1_Brachionus             ............................................................
SIX1_Pomacea                .........................................PMGGVANGASSDVGQLLSD
SIX6_Pomacea                ............................................................
SIX4_Pomacea                ............................................................
SINE_Capitella_teleta       .......................................SSAHVQDMNGLLSDYQSL...
Optix_Capitella_teleta      ............................................................
SIX4_Capitella_teleta       VHREMLQQQESKVAGMIDDAKSRHYSERITCANCKETFTIHAVGMKIRHELDNVQQLHIA
Sine_Drosophila             ...................................YSHLHSVMGAMPMTAMYDMGEYQHL
Optix_Drosophila            QLPQSLALNGASGGPPLHHPAYAAAYSVECVPGGHGPPHPPPKLRINSPEKLNSTAVAAA
Six4_Drosophila             ............................................................
So_H_erato                  ............................................................
Optix_H_erato               ............................................................
SIX4_H_mel                  ............................................................
SINE_Daphnia                SHQLQMASAAAAAAAAAAAAAAFGSPHHVTSHHHPHADLAHSLLQQPTSQSHDYNPHVSS
SIX4_Daphnia                .....................................VQPSSLYHGHRDQAFSHA.....
OPTIX_Daphnia               AYHHAAQQHQQQQQQQHQQHQQATTQTTSIVSPTIEPLKLKSDLIITSSKSNHSGSSRSS
SIX1_Strongylocentrotus     ................AAEAAAQALAPVNPGDSVLQNSIQNSMLPHMANNLVDMG.....
SIX6_Strongylocentrotus     ............................................................
SIX4_Strongylocentrotus     LSSIPAVVSASSALSAVLSQQAATAGAGFPSNFAGRLPSANEMLNSAAMALSTLSANANG
Six1/2_Halocynthia          HASHGMQQQHVPDMHFQQQSHIPHHHQAAAIHNPSVTLHESMMHHPMTTTMLDMGS....
Six3/6_Halocynthia          .......................................QTLKTRRIDDVIRYLGCE...
Six4/5_Halocynthia          ........................................VPQHMASSKTLPNLRHL...
Six1/2_Branchiostoma        ........................................GAMMATMASDLVSLGP....
Six4/5_Branchiostoma        ............................................................
Six1_Petromyzon             ....................QQQQQQQLHQHQQQQQHALQDSMMHSISSSLVNIG.....
SIX6_Petromyzon             ............................................................
SIX4_Petromyzon             SLTNNNISPLMVQGIPGNGIVGMGGLTPSLPTVSQTADAIAGQMGSATLFSRTVSTALMP
Six1_like_PetromyzonX1      QQYVGGTLMPLPVLVSNSAGGALPGPGSLQQQQQQHALQDSSLLQPLSSSLVNIG.....
Six1_like_Petromyzon        ................QHNHHHHHQQQQHHQQQHLHHGLQDSLLSPMSSGIVNIGS....
Six2_like_PetromyzonX1      VSGTSDRRRGAYQEEAGGRAGGRVGPTVAAAPGRVGATAAPFLAGPTGPTAMAVPPRTAD
Six6_like_Petromyzon        ............................................................
Six6_like_Petromyzon2       ............................................................
Six2_like_Petromyzon        PPPPLPPLPHAQQHGGPDLAFGDGALNGQHSYADLFPPSTESLLYSLSANLLELGS....
Six6_like_Petromyzon3       ............................................................
SIX1_Rhincodon              ........................................DSLLGTLTSSLVDLGS....
SIX2_Rhincodon              ......................................LQDSMLNPMSSNLVDLGS....
SIX3_Rhincodon              ............................................................
SIX4_Rhincodon              LKTESVHMLTSEEAVSSGSAEISRGPPQVNQYSLVQLQNTENNIQLVNGNIGLPPLQLPS
SIX6_Rhincodon              ............................................................
SIX1b_Spotted               ........................................DSLLGPLTSSLVDLGS....
SIX2_Spotted                ........................................DTILNPMSSNLVDLGS....
SIX3_Spotted                ............................................................
SIX4_Spotted                LDQSGQLSLGSTHLSASASSPLSAEAALNNAYAPSVLGPSDPLNSGNAAGLSPPPSSQAS
SIX5_Spotted                FRLNQLGALQISGTQGVGTAGSSSTGVQILNSSIFQLPSASPGNILLTNPAGGSTILTGV
SIX6_Spotted                ............................................................
SIX7_Spotted                ............................................................
Spotted_Gar_SIX1_like       ............................................................
SIX1_Callorhinchus          ........................................DSLLGTLTSSLVDLGS....
SIX2_Callorhinchus          ....................................YPIHDPILNPASPNLLQLGS....
SIX3_Callorhinchus          ............................................................
SIX4a_Callorhinchus         TPSSSLISTGPPQNLPLTSSNVLNGSGTLSFSLTVSLPLSTATHATLDQNVSTAIGDSLP
SIX6_Callorhinchus          ............................................................
SIX1_Erpetoichthys          ..........................................LLGPLTSSLVDLGS....
SIX2a_Erpetoichthys         ........................................EAILNPMSSNLVDLGS....
SIX3_Erpetoichthys          ............................................................
SIX4_Erpetoichthys          SFSVISTPGLGSSGSLNSSSSGSMCSTQSSQPTSPADSSPSSEAAGNSNGYAALHDTPLT
SIX5_Erpetoichthys          QVGTSPAPPLTVQQVIQTPQGLQTVQTLSQLPAPQIIPISSPTQVVPLSQPGQASPNATP
SIX6_Erpetoichthys          ............................................................
SIX7_Erpetoichthys          ............................................................
SIX1_HUMAN                  ........................................DSLLGPLTSSLVDLGS....
SIX2_HUMAN                  ......................................LQDSILNPMSANLVDLGS....
SIX3_HUMAN                  ............................................................
SIX4_HUMAN                  VSPTHNFSLSPSTLLNPTELNRDIADSQPMSAPVASKSTVTSVSNTNYATLQNCSLITGQ
SIX5_HUMAN                  IILTATFPTSMLVSQVLPPAPGLALPLKPETAISVPEGGLPVAPSPALPEAHALGTLSAQ
SIX6_HUMAN                  ............................................................
```

```
                            440       450
                              .         .
SINE_Amphimedon          YQTPAGAASYFTGQTASLPNTMS.....................................
SciSixB                  TAAGYNNSVSFSQPAVPTQASTPMDVSNGYHGNNGYHHLQQGYR................
SciSixc                  EQPAACVPATAAAHSPTGGLLDTSMPATVPALSAASTTSAPQPTTAVASVSCETEM....
LCOSixB                  NGTYQGYSMSQPVPQQFHRNTNYSQHASSAVQSPTYS.......................
LCOSixc                  TPPTLAAADVMPARTIACSGSSTAVSTNSPVEQCSTTSV.....................
Nvec_Six1                ............................................................
Nvec_Six3                ............................................................
Nvec_Six4                ............................................................
Cwil_SIXC                ............................................................
Cwil_SIXA                ............................................................
Cwil_SIXB                ............................................................
Tri_SIX3                 ............................................................
Tri_SIX1A                ............................................................
C_elegans_ceh_33         ............................................................
C_elegans_ceh_32         DPKHSSPKSTTSQSE.............................................
C_elegans_ceh_34         ............................................................
C_elegans_unc_39         ............................................................
Sine_Schmidtea           YNPYFSQNTYGNFHNPDYDTPIYSYN..................................
Six3_Schmidtea           YGLM........................................................
SIX2_Brachionus          TAPFHSL.....................................................
SIX3_Brachionus          ............................................................
SIX1_Brachionus          ............................................................
SIX1_Pomacea             YQTL........................................................
SIX6_Pomacea             ............................................................
SIX4_Pomacea             ............................................................
SINE_Capitella_teleta    ............................................................
Optix_Capitella_teleta   ............................................................
SIX4_Capitella_teleta    DIQCVLLSVASL................................................
Sine_Drosophila          ............................................................
Optix_Drosophila         ASVGGGGGNQHHEPTTTGYHHSGQLMLHRPFSTSPELKHSAPEIT...............
Six4_Drosophila          ............................................................
So_H_erato               ............................................................
Optix_H_erato            ............................................................
SIX4_H_mel               ............................................................
SINE_Daphnia             ............................................................
SIX4_Daphnia             ............................................................
OPTIX_Daphnia            RSSNSSRHSGRSTPPTPPTPPQKTSILHQHEMVAETGTQAS...................
SIX1_Strongylocentrotus  ............................................................
SIX6_Strongylocentrotus  ............................................................
SIX4_Strongylocentrotus  GGPGHYASE...................................................
Six1/2_Halocynthia       ............................................................
Six3/6_Halocynthia       ............................................................
Six4/5_Halocynthia       ............................................................
Six1/2_Branchiostoma     ............................................................
Six4/5_Branchiostoma     ............................................................
Six1_Petromyzon          ............................................................
SIX6_Petromyzon          ............................................................
SIX4_Petromyzon          TTLNTNALAMQQTYLYNHAGGLDTAAVQVVYAEAPPVYTPMPVGAPAESAVYKSAN....
Six1_like_PetromyzonX1   ............................................................
Six1_like_Petromyzon     ............................................................
Six2_like_PetromyzonX1   EMERRTEEGEEERGEGPREEGNGAPAVLATLCSVLFEEQLP...................
Six6_like_Petromyzon     ............................................................
Six6_like_Petromyzon2    ............................................................
Six2_like_Petromyzon     ............................................................
Six6_like_Petromyzon3    ............................................................
SIX1_Rhincodon           ............................................................
SIX2_Rhincodon           ............................................................
SIX3_Rhincodon           ............................................................
SIX4_Rhincodon           VSAASSQ.....................................................
SIX6_Rhincodon           ............................................................
SIX1b_Spotted            ............................................................
SIX2_Spotted             ............................................................
SIX3_Spotted             ............................................................
SIX4_Spotted             PVTIISSSSAEPVGSASYTTLTVAASPGAQAAHHQGMGAGQGGGAISADYNGHRVP....
SIX5_Spotted             TFQQGKLILTATFPASMQLASLPLKPDAESAPANGGIVLTPVISVGSAGGGGAPQG....
SIX6_Spotted             ............................................................
SIX7_Spotted             ............................................................
Spotted_Gar_SIX1_like    ............................................................
SIX1_Callorhinchus       ............................................................
SIX2_Callorhinchus       ............................................................
SIX3_Callorhinchus       ............................................................
SIX4a_Callorhinchus      IVSMGNSVYATPQNCGLLSNSTQENVCESVVVSVSGDKPRNDLTEGHQEYEHVSVL....
SIX6_Callorhinchus       ............................................................
SIX1_Erpetoichthys       ............................................................
SIX2a_Erpetoichthys      ............................................................
SIX3_Erpetoichthys       ............................................................
SIX4_Erpetoichthys       TTAGQGTFTVFSTQQGLARESGRGTDDDRTNQHLPHDYSHQQSTIQQLLPALKGNF....
SIX5_Erpetoichthys       TAQIVPLSAQVVSPCPPPLPQVVPSSQTVPGTFQILTSVNNGNNSAVKYSQPNTIQFPSS
SIX6_Erpetoichthys       ............................................................
SIX7_Erpetoichthys       ............................................................
SIX1_HUMAN               ............................................................
SIX2_HUMAN               ............................................................
SIX3_HUMAN               ............................................................
SIX4_HUMAN               DLLSVPMTQAALGEIVPTAEDQVGHPSPAVHQDFVQEHRLVLQSVANMKENFLSNS....
SIX5_HUMAN               QPPPAAATTSSTSLPFSPDSPGLLPNFPAPPPEGLMLSPAAVPVWSAGLELSAGTE....
SIX6_HUMAN               ............................................................
```

```
SINE_Amphimedon            ............................................................
SciSixB                    ............................................................
SciSixc                    ............................................................
LCOSixB                    ............................................................
LCOSixc                    ............................................................
Nvec_Six1                  ............................................................
Nvec_Six3                  ............................................................
Nvec_Six4                  ............................................................
Cwil_SIXC                  ............................................................
Cwil_SIXA                  ............................................................
Cwil_SIXB                  ............................................................
Tri_SIX3                   ............................................................
Tri_SIX1A                  ............................................................
C_elegans_ceh_33           ............................................................
C_elegans_ceh_32           ............................................................
C_elegans_ceh_34           ............................................................
C_elegans_unc_39           ............................................................
Sine_Schmidtea             ............................................................
Six3_Schmidtea             ............................................................
SIX2_Brachionus            ............................................................
SIX3_Brachionus            ............................................................
SIX1_Brachionus            ............................................................
SIX1_Pomacea               ............................................................
SIX6_Pomacea               ............................................................
SIX4_Pomacea               ............................................................
SINE_Capitella_teleta      ............................................................
Optix_Capitella_teleta     ............................................................
SIX4_Capitella_teleta      ............................................................
Sine_Drosophila            ............................................................
Optix_Drosophila           ............................................................
Six4_Drosophila            ............................................................
So_H_erato                 ............................................................
Optix_H_erato              ............................................................
SIX4_H_mel                 ............................................................
SINE_Daphnia               ............................................................
SIX4_Daphnia               ............................................................
OPTIX_Daphnia              ............................................................
SIX1_Strongylocentrotus    ............................................................
SIX6_Strongylocentrotus    ............................................................
SIX4_Strongylocentrotus    ............................................................
Six1/2_Halocynthia         ............................................................
Six3/6_Halocynthia         ............................................................
Six4/5_Halocynthia         ............................................................
Six1/2_Branchiostoma       ............................................................
Six4/5_Branchiostoma       ............................................................
Six1_Petromyzon            ............................................................
SIX6_Petromyzon            ............................................................
SIX4_Petromyzon            ............................................................
Six1_like_PetromyzonX1     ............................................................
Six1_like_Petromyzon       ............................................................
Six2_like_PetromyzonX1     ............................................................
Six6_like_Petromyzon       ............................................................
Six6_like_Petromyzon2      ............................................................
Six2_like_Petromyzon       ............................................................
Six6_like_Petromyzon3      ............................................................
SIX1_Rhincodon             ............................................................
SIX2_Rhincodon             ............................................................
SIX3_Rhincodon             ............................................................
SIX4_Rhincodon             ............................................................
SIX6_Rhincodon             ............................................................
SIX1b_Spotted              ............................................................
SIX2_Spotted               ............................................................
SIX3_Spotted               ............................................................
SIX4_Spotted               ............................................................
SIX5_Spotted               ............................................................
SIX6_Spotted               ............................................................
SIX7_Spotted               ............................................................
Spotted_Gar_SIX1_like      ............................................................
SIX1_Callorhinchus         ............................................................
SIX2_Callorhinchus         ............................................................
SIX3_Callorhinchus         ............................................................
SIX4a_Callorhinchus        ............................................................
SIX6_Callorhinchus         ............................................................
SIX1_Erpetoichthys         ............................................................
SIX2a_Erpetoichthys        ............................................................
SIX3_Erpetoichthys         ............................................................
SIX4_Erpetoichthys         ............................................................
SIX5_Erpetoichthys         SPVAQNGGSIGCTTAGVQLINSGGIFQLPAAAPGNLILTNPAGGSTLLTFQQGKLILTAT
SIX6_Erpetoichthys         ............................................................
SIX7_Erpetoichthys         ............................................................
SIX1_HUMAN                 ............................................................
SIX2_HUMAN                 ............................................................
SIX3_HUMAN                 ............................................................
SIX4_HUMAN                 ............................................................
SIX5_HUMAN                 ............................................................
SIX6_HUMAN                 ............................................................
```

```
SINE_Amphimedon           ............................................................
SciSixB                   ............................................................
SciSixc                   ............................................................
LCOSixB                   ............................................................
LCOSixc                   ............................................................
Nvec_Six1                 ............................................................
Nvec_Six3                 ............................................................
Nvec_Six4                 ............................................................
Cwil_SIXC                 ............................................................
Cwil_SIXA                 ............................................................
Cwil_SIXB                 ............................................................
Tri_SIX3                  ............................................................
Tri_SIX1A                 ............................................................
C_elegans_ceh_33          ............................................................
C_elegans_ceh_32          ............................................................
C_elegans_ceh_34          ............................................................
C_elegans_unc_39          ............................................................
Sine_Schmidtea            ............................................................
Six3_Schmidtea            ............................................................
SIX2_Brachionus           ............................................................
SIX3_Brachionus           ............................................................
SIX1_Brachionus           ............................................................
SIX1_Pomacea              ............................................................
SIX6_Pomacea              ............................................................
SIX4_Pomacea              ............................................................
SINE_Capitella_teleta     ............................................................
Optix_Capitella_teleta    ............................................................
SIX4_Capitella_teleta     ............................................................
Sine_Drosophila           ............................................................
Optix_Drosophila          ............................................................
Six4_Drosophila           ............................................................
So_H_erato                ............................................................
Optix_H_erato             ............................................................
SIX4_H_mel                ............................................................
SINE_Daphnia              ............................................................
SIX4_Daphnia              ............................................................
OPTIX_Daphnia             ............................................................
SIX1_Strongylocentrotus   ............................................................
SIX6_Strongylocentrotus   ............................................................
SIX4_Strongylocentrotus   ............................................................
Six1/2_Halocynthia        ............................................................
Six3/6_Halocynthia        ............................................................
Six4/5_Halocynthia        ............................................................
Six1/2_Branchiostoma      ............................................................
Six4/5_Branchiostoma      ............................................................
Six1_Petromyzon           ............................................................
SIX6_Petromyzon           ............................................................
SIX4_Petromyzon           ............................................................
Six1_like_PetromyzonX1    ............................................................
Six1_like_Petromyzon      ............................................................
Six2_like_PetromyzonX1    ............................................................
Six6_like_Petromyzon      ............................................................
Six6_like_Petromyzon2     ............................................................
Six2_like_Petromyzon      ............................................................
Six6_like_Petromyzon3     ............................................................
SIX1_Rhincodon            ............................................................
SIX2_Rhincodon            ............................................................
SIX3_Rhincodon            ............................................................
SIX4_Rhincodon            ............................................................
SIX6_Rhincodon            ............................................................
SIX1b_Spotted             ............................................................
SIX2_Spotted              ............................................................
SIX3_Spotted              ............................................................
SIX4_Spotted              ............................................................
SIX5_Spotted              ............................................................
SIX6_Spotted              ............................................................
SIX7_Spotted              ............................................................
Spotted_Gar_SIX1_like     ............................................................
SIX1_Callorhinchus        ............................................................
SIX2_Callorhinchus        ............................................................
SIX3_Callorhinchus        ............................................................
SIX4a_Callorhinchus       ............................................................
SIX6_Callorhinchus        ............................................................
SIX1_Erpetoichthys        ............................................................
SIX2a_Erpetoichthys       ............................................................
SIX3_Erpetoichthys        ............................................................
SIX4_Erpetoichthys        ............................................................
SIX5_Erpetoichthys        FPASMLVSSSATSLANLPLKTESEYSQNGAGTGIVLTPIISVGPGQQNCPVNNISPSGTS
SIX6_Erpetoichthys        ............................................................
SIX7_Erpetoichthys        ............................................................
SIX1_HUMAN                ............................................................
SIX2_HUMAN                ............................................................
SIX3_HUMAN                ............................................................
SIX4_HUMAN                ............................................................
SIX5_HUMAN                ............................................................
SIX6_HUMAN                ............................................................
```

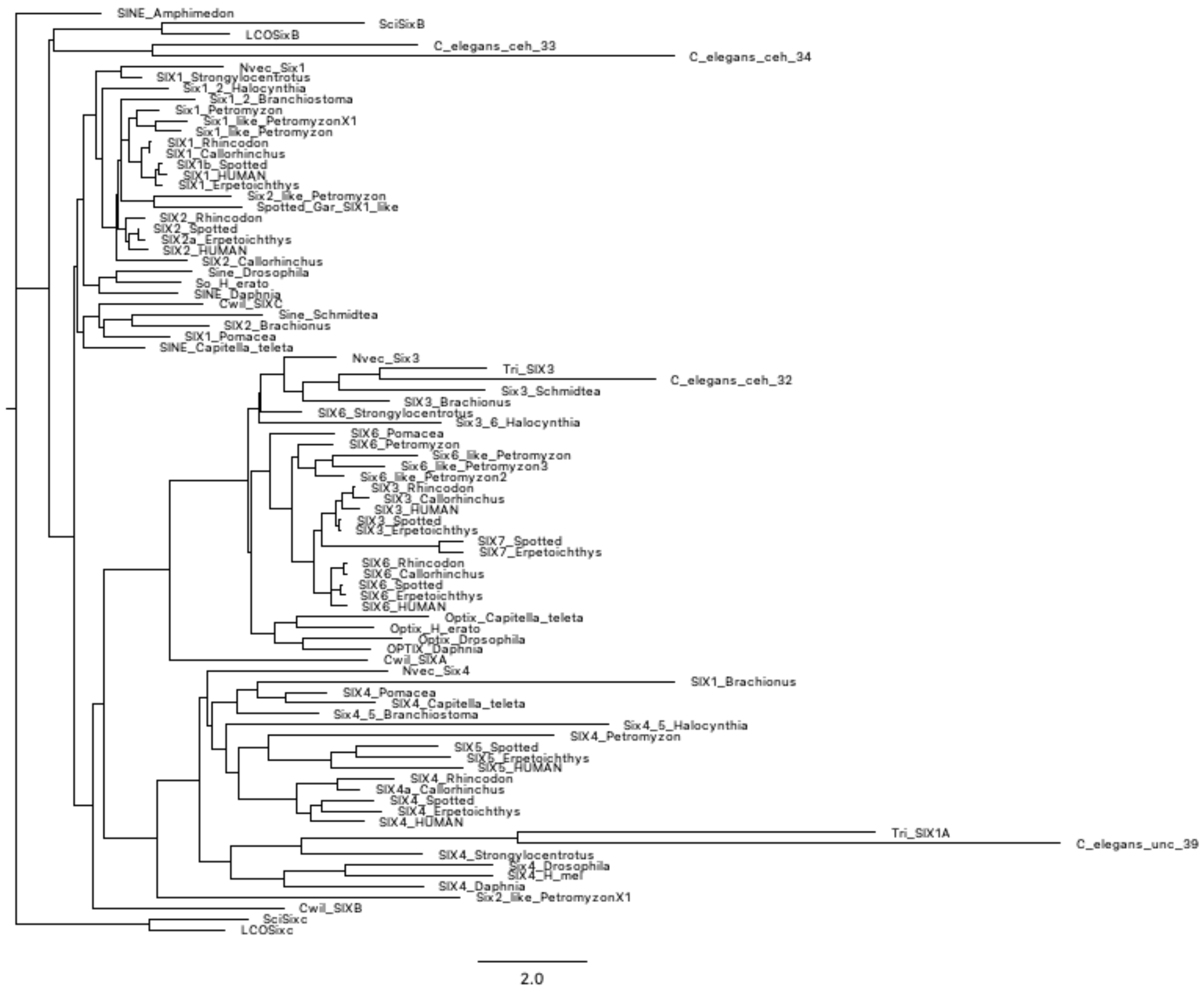
```
SINE_Amphimedon           ............................................................
SciSixB                   ............................................................
SciSixc                   ............................................................
LCOSixB                   ............................................................
LCOSixc                   ............................................................
Nvec_Six1                 ............................................................
Nvec_Six3                 ............................................................
Nvec_Six4                 ............................................................
Cwil_SIXC                 ............................................................
Cwil_SIXA                 ............................................................
Cwil_SIXB                 ............................................................
Tri_SIX3                  ............................................................
Tri_SIX1A                 ............................................................
C_elegans_ceh_33          ............................................................
C_elegans_ceh_32          ............................................................
C_elegans_ceh_34          ............................................................
C_elegans_unc_39          ............................................................
Sine_Schmidtea            ............................................................
Six3_Schmidtea            ............................................................
SIX2_Brachionus           ............................................................
SIX3_Brachionus           ............................................................
SIX1_Brachionus           ............................................................
SIX1_Pomacea              ............................................................
SIX6_Pomacea              ............................................................
SIX4_Pomacea              ............................................................
SINE_Capitella_teleta     ............................................................
Optix_Capitella_teleta    ............................................................
SIX4_Capitella_teleta     ............................................................
Sine_Drosophila           ............................................................
Optix_Drosophila          ............................................................
Six4_Drosophila           ............................................................
So_H_erato                ............................................................
Optix_H_erato             ............................................................
SIX4_H_mel                ............................................................
SINE_Daphnia              ............................................................
SIX4_Daphnia              ............................................................
OPTIX_Daphnia             ............................................................
SIX1_Strongylocentrotus   ............................................................
SIX6_Strongylocentrotus   ............................................................
SIX4_Strongylocentrotus   ............................................................
Six1/2_Halocynthia        ............................................................
Six3/6_Halocynthia        ............................................................
Six4/5_Halocynthia        ............................................................
Six1/2_Branchiostoma      ............................................................
Six4/5_Branchiostoma      ............................................................
Six1_Petromyzon           ............................................................
SIX6_Petromyzon           ............................................................
SIX4_Petromyzon           ............................................................
Six1_like_PetromyzonX1    ............................................................
Six1_like_Petromyzon      ............................................................
Six2_like_PetromyzonX1    ............................................................
Six6_like_Petromyzon      ............................................................
Six6_like_Petromyzon2     ............................................................
Six2_like_Petromyzon      ............................................................
Six6_like_Petromyzon3     ............................................................
SIX1_Rhincodon            ............................................................
SIX2_Rhincodon            ............................................................
SIX3_Rhincodon            ............................................................
SIX4_Rhincodon            ............................................................
SIX6_Rhincodon            ............................................................
SIX1b_Spotted             ............................................................
SIX2_Spotted              ............................................................
SIX3_Spotted              ............................................................
SIX4_Spotted              ............................................................
SIX5_Spotted              ....................................................TGGPAALA
SIX6_Spotted              ............................................................
SIX7_Spotted              ............................................................
Spotted_Gar_SIX1_like     ............................................................
SIX1_Callorhinchus        ............................................................
SIX2_Callorhinchus        ............................................................
SIX3_Callorhinchus        ............................................................
SIX4a_Callorhinchus       ............................................................
SIX6_Callorhinchus        ............................................................
SIX1_Erpetoichthys        ............................................................
SIX2a_Erpetoichthys       ............................................................
SIX3_Erpetoichthys        ............................................................
SIX4_Erpetoichthys        ............................................................
SIX5_Erpetoichthys        LPPISAVIPISPTTTTSSSSVSLPQEGTITSSQQSYQTDSNITFINPGGFYPNTAPGGDN
SIX6_Erpetoichthys        ............................................................
SIX7_Erpetoichthys        ............................................................
SIX1_HUMAN                ............................................................
SIX2_HUMAN                ............................................................
SIX3_HUMAN                ............................................................
SIX4_HUMAN                ............................................................
SIX5_HUMAN                ............................................................
SIX6_HUMAN                ............................................................
```

```
SINE_Amphimedon           ............................................................
SciSixB                   ............................................................
SciSixc                   ............................................................
LCOSixB                   ............................................................
LCOSixc                   ............................................................
Nvec_Six1                 ............................................................
Nvec_Six3                 ............................................................
Nvec_Six4                 ............................................................
Cwil_SIXC                 ............................................................
Cwil_SIXA                 ............................................................
Cwil_SIXB                 ............................................................
Tri_SIX3                  ............................................................
Tri_SIX1A                 ............................................................
C_elegans_ceh_33          ............................................................
C_elegans_ceh_32          ............................................................
C_elegans_ceh_34          ............................................................
C_elegans_unc_39          ............................................................
Sine_Schmidtea            ............................................................
Six3_Schmidtea            ............................................................
SIX2_Brachionus           ............................................................
SIX3_Brachionus           ............................................................
SIX1_Brachionus           ............................................................
SIX1_Pomacea              ............................................................
SIX6_Pomacea              ............................................................
SIX4_Pomacea              ............................................................
SINE_Capitella_teleta     ............................................................
Optix_Capitella_teleta    ............................................................
SIX4_Capitella_teleta     ............................................................
Sine_Drosophila           ............................................................
Optix_Drosophila          ............................................................
Six4_Drosophila           ............................................................
So_H_erato                ............................................................
Optix_H_erato             ............................................................
SIX4_H_mel                ............................................................
SINE_Daphnia              ............................................................
SIX4_Daphnia              ............................................................
OPTIX_Daphnia             ............................................................
SIX1_Strongylocentrotus   ............................................................
SIX6_Strongylocentrotus   ............................................................
SIX4_Strongylocentrotus   ............................................................
Six1/2_Halocynthia        ............................................................
Six3/6_Halocynthia        ............................................................
Six4/5_Halocynthia        ............................................................
Six1/2_Branchiostoma      ............................................................
Six4/5_Branchiostoma      ............................................................
Six1_Petromyzon           ............................................................
SIX6_Petromyzon           ............................................................
SIX4_Petromyzon           ..........................................................EA
Six1_like_PetromyzonX1    ............................................................
Six1_like_Petromyzon      ............................................................
Six2_like_PetromyzonX1    ............................................................
Six6_like_Petromyzon      ............................................................
Six6_like_Petromyzon2     ............................................................
Six2_like_Petromyzon      ............................................................
Six6_like_Petromyzon3     ............................................................
SIX1_Rhincodon            ............................................................
SIX2_Rhincodon            ............................................................
SIX3_Rhincodon            ............................................................
SIX4_Rhincodon            ............................................................
SIX6_Rhincodon            ............................................................
SIX1b_Spotted             ............................................................
SIX2_Spotted              ............................................................
SIX3_Spotted              ............................................................
SIX4_Spotted              ............................................................
SIX5_Spotted              MASPTTSSSSFQSDSSLSFVSPAGLYPAPAPEAVSSTAMLHVAPPSTVGGLSPQVSKLGD
SIX6_Spotted              ............................................................
SIX7_Spotted              ............................................................
Spotted_Gar_SIX1_like     ............................................................
SIX1_Callorhinchus        ............................................................
SIX2_Callorhinchus        ............................................................
SIX3_Callorhinchus        ............................................................
SIX4a_Callorhinchus       ............................................................
SIX6_Callorhinchus        ............................................................
SIX1_Erpetoichthys        ............................................................
SIX2a_Erpetoichthys       ............................................................
SIX3_Erpetoichthys        ............................................................
SIX4_Erpetoichthys        ............................................................
SIX5_Erpetoichthys        HAMLSSVSLTPAASVGLTCSDVIGGLNTSMSMGGVGANAGHLSSANVTSSAMSNLAQVVW
SIX6_Erpetoichthys        ............................................................
SIX7_Erpetoichthys        ............................................................
SIX1_HUMAN                ............................................................
SIX2_HUMAN                ............................................................
SIX3_HUMAN                ............................................................
SIX4_HUMAN                ............................................................
SIX5_HUMAN                ............................................................
SIX6_HUMAN                ............................................................
```

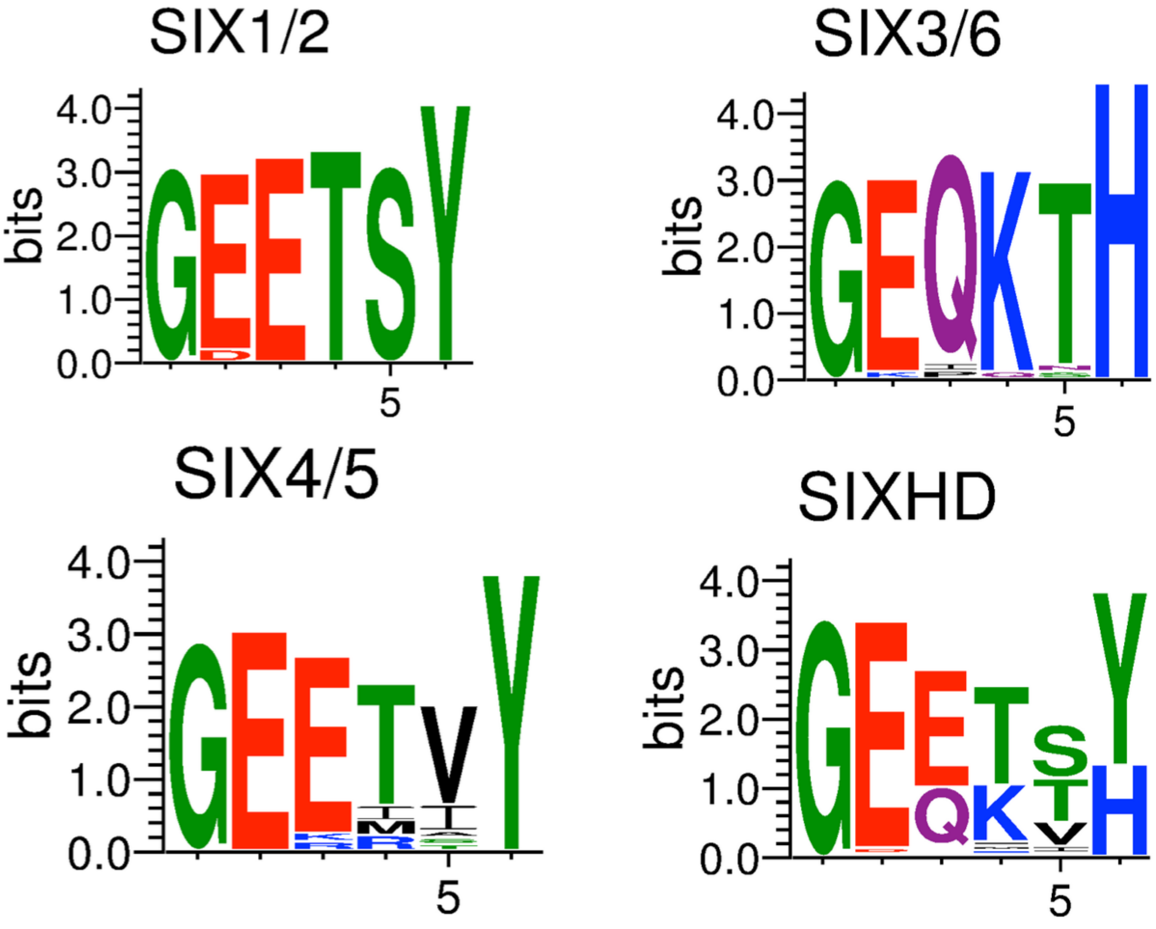
```
SINE_Amphimedon              ............................................................
SciSixB                      ............................................................
SciSixc                      ............................................................
LCOSixB                      ............................................................
LCOSixc                      ............................................................
Nvec_Six1                    ............................................................
Nvec_Six3                    ............................................................
Nvec_Six4                    ............................................................
Cwil_SIXC                    ............................................................
Cwil_SIXA                    ............................................................
Cwil_SIXB                    ............................................................
Tri_SIX3                     ............................................................
Tri_SIX1A                    ............................................................
C_elegans_ceh_33             ............................................................
C_elegans_ceh_32             ............................................................
C_elegans_ceh_34             ............................................................
C_elegans_unc_39             ............................................................
Sine_Schmidtea               ............................................................
Six3_Schmidtea               ............................................................
SIX2_Brachionus              ............................................................
SIX3_Brachionus              ............................................................
SIX1_Brachionus              ............................................................
SIX1_Pomacea                 ............................................................
SIX6_Pomacea                 ............................................................
SIX4_Pomacea                 ............................................................
SINE_Capitella_teleta        ............................................................
Optix_Capitella_teleta       ............................................................
SIX4_Capitella_teleta        ............................................................
Sine_Drosophila              ............................................................
Optix_Drosophila             ............................................................
Six4_Drosophila              ............................................................
So_H_erato                   ............................................................
Optix_H_erato                ............................................................
SIX4_H_mel                   ............................................................
SINE_Daphnia                 ............................................................
SIX4_Daphnia                 ............................................................
OPTIX_Daphnia                ............................................................
SIX1_Strongylocentrotus      ............................................................
SIX6_Strongylocentrotus      ............................................................
SIX4_Strongylocentrotus      ............................................................
Six1/2_Halocynthia           ............................................................
Six3/6_Halocynthia           ............................................................
Six4/5_Halocynthia           ............................................................
Six1/2_Branchiostoma         ............................................................
Six4/5_Branchiostoma         ............................................................
Six1_Petromyzon              ............................................................
SIX6_Petromyzon              ............................................................
SIX4_Petromyzon              ISAIADCCYEDLNTSSSVGSDSTDNKSEADSLLGSPLDLSGSMGAVGNAENEGKEEGSEM
Six1_like_PetromyzonX1       ............................................................
Six1_like_Petromyzon         ............................................................
Six2_like_PetromyzonX1       ............................................................
Six6_like_Petromyzon         ............................................................
Six6_like_Petromyzon2        ............................................................
Six2_like_Petromyzon         ............................................................
Six6_like_Petromyzon3        ............................................................
SIX1_Rhincodon               ............................................................
SIX2_Rhincodon               ............................................................
SIX3_Rhincodon               ............................................................
SIX4_Rhincodon               ............................................................
SIX6_Rhincodon               ............................................................
SIX1b_Spotted                ............................................................
SIX2_Spotted                 ............................................................
SIX3_Spotted                 ............................................................
SIX4_Spotted                 .........................TLAHASGLKGNFLSIADSKPRAENLLLRTKPGISD
SIX5_Spotted                 AHLSLAMSTPVSGQAAVWSPGLFDVRKGDLPEEEAHQGLLGLTGGDGLLLGAPSPGPHGE
SIX6_Spotted                 ............................................................
SIX7_Spotted                 ............................................................
Spotted_Gar_SIX1_like        ............................................................
SIX1_Callorhinchus           ............................................................
SIX2_Callorhinchus           ............................................................
SIX3_Callorhinchus           ............................................................
SIX4a_Callorhinchus          ...........................HPTVKETYIAVSENKCSNHMMMMDSKSKYVIHD
SIX6_Callorhinchus           ............................................................
SIX1_Erpetoichthys           ............................................................
SIX2a_Erpetoichthys          ............................................................
SIX3_Erpetoichthys           ............................................................
SIX4_Erpetoichthys           ...................................LSISEGRSSEDMIILESKSKCAVSE
SIX5_Erpetoichthys           SPALNPTSAVSTGLVLPLGLRKEDRLLPDDGVDHRSLLALPGGESLLLGTAPEVRGQQLE
SIX6_Erpetoichthys           ............................................................
SIX7_Erpetoichthys           ............................................................
SIX1_HUMAN                   ............................................................
SIX2_HUMAN                   ............................................................
SIX3_HUMAN                   ............................................................
SIX4_HUMAN                   .......................................ESKATSSLMMLDSKSKYVLDG
SIX5_HUMAN                   ............................GLLEAEKGLGTQAPHTVLRLPDPDPEGLLLGA
SIX6_HUMAN                   ............................................................
```

```
SINE_Amphimedon            ...............................
SciSixB                    ...............................
SciSixc                    ...............................
LCOSixB                    ...............................
LCOSixc                    ...............................
Nvec_Six1                  ...............................
Nvec_Six3                  ...............................
Nvec_Six4                  ...............................
Cwil_SIXC                  ...............................
Cwil_SIXA                  ...............................
Cwil_SIXB                  ...............................
Tri_SIX3                   ...............................
Tri_SIX1A                  ...............................
C_elegans_ceh_33           ...............................
C_elegans_ceh_32           ...............................
C_elegans_ceh_34           ...............................
C_elegans_unc_39           ...............................
Sine_Schmidtea             ...............................
Six3_Schmidtea             ...............................
SIX2_Brachionus            ...............................
SIX3_Brachionus            ...............................
SIX1_Brachionus            ...............................
SIX1_Pomacea               ...............................
SIX6_Pomacea               ...............................
SIX4_Pomacea               ...............................
SINE_Capitella_teleta      ...............................
Optix_Capitella_teleta     ...............................
SIX4_Capitella_teleta      ...............................
Sine_Drosophila            ...............................
Optix_Drosophila           ...............................
Six4_Drosophila            ...............................
So_H_erato                 ...............................
Optix_H_erato              ...............................
SIX4_H_mel                 ...............................
SINE_Daphnia               ...............................
SIX4_Daphnia               ...............................
OPTIX_Daphnia              ...............................
SIX1_Strongylocentrotus    ...............................
SIX6_Strongylocentrotus    ...............................
SIX4_Strongylocentrotus    ...............................
Six1/2_Halocynthia         ...............................
Six3/6_Halocynthia         ...............................
Six4/5_Halocynthia         ...............................
Six1/2_Branchiostoma       ...............................
Six4/5_Branchiostoma       ...............................
Six1_Petromyzon            ...............................
SIX6_Petromyzon            ...............................
SIX4_Petromyzon            TSDGHEDFVHGLLPKMTSAPDDDNFYDFDDF
Six1_like_PetromyzonX1     ...............................
Six1_like_Petromyzon       ...............................
Six2_like_PetromyzonX1     ...............................
Six6_like_Petromyzon       ...............................
Six6_like_Petromyzon2      ...............................
Six2_like_Petromyzon       ...............................
Six6_like_Petromyzon3      ...............................
SIX1_Rhincodon             ...............................
SIX2_Rhincodon             ...............................
SIX3_Rhincodon             ...............................
SIX4_Rhincodon             ...............................
SIX6_Rhincodon             ...............................
SIX1b_Spotted              ...............................
SIX2_Spotted               ...............................
SIX3_Spotted               ...............................
SIX4_Spotted               MVRVICGEMETEEKELAKLQNVQMEEDMNDL
SIX5_Spotted               QAQLEDPEDMDGDPKVLTQLQSVPVDEDLGL
SIX6_Spotted               ...............................
SIX7_Spotted               ...............................
Spotted_Gar_SIX1_like      ...............................
SIX1_Callorhinchus         ...............................
SIX2_Callorhinchus         ...............................
SIX3_Callorhinchus         ...............................
SIX4a_Callorhinchus        VVNSVCKELETEEKELAKLQNVPMDEDMCDI
SIX6_Callorhinchus         ...............................
SIX1_Erpetoichthys         ...............................
SIX2a_Erpetoichthys        ...............................
SIX3_Erpetoichthys         ...............................
SIX4_Erpetoichthys         MVRVICGQMENEDKQLAKLQNVQMEEDISEI
SIX5_Erpetoichthys         EGPNMDSDDLESDGKVLTQLQSVPVDEDLGM
SIX6_Erpetoichthys         ...............................
SIX7_Erpetoichthys         ...............................
SIX1_HUMAN                 ...............................
SIX2_HUMAN                 ...............................
SIX3_HUMAN                 ...............................
SIX4_HUMAN                 MVDTVCEDLETDKKELAKLQTVQLDEDMQDL
SIX5_HUMAN                 TAGGEVDEGLEAEAKVLTQLQSVPVEEPLEL
SIX6_HUMAN                 ...............................
```

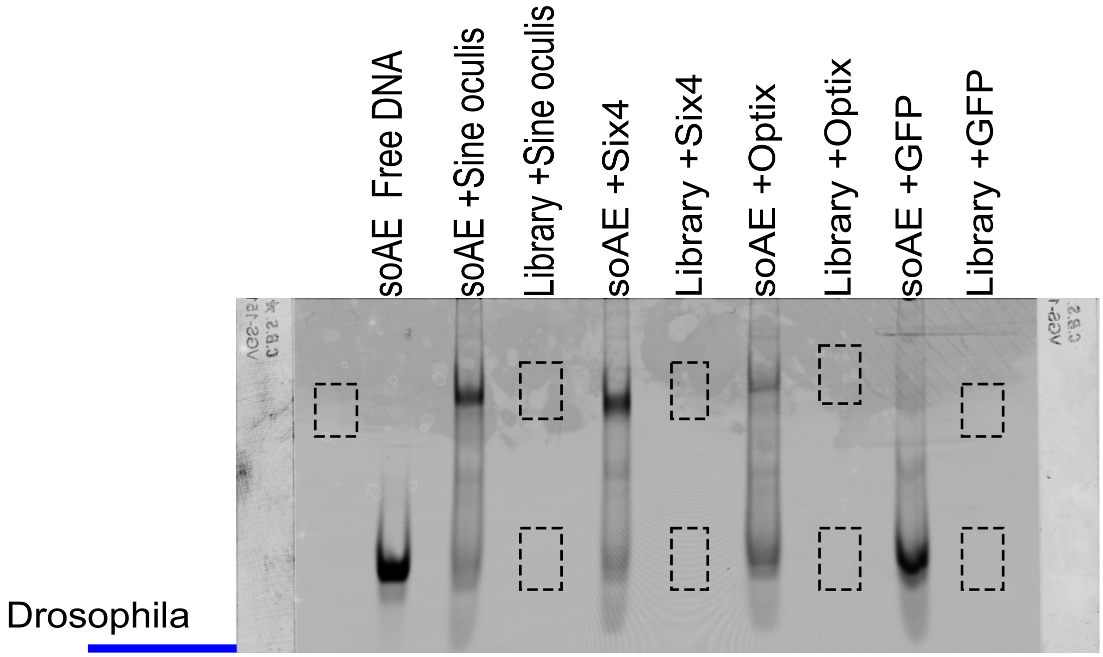
Supplementary Image 2: SIX phylogenetic tree before Dendroscope

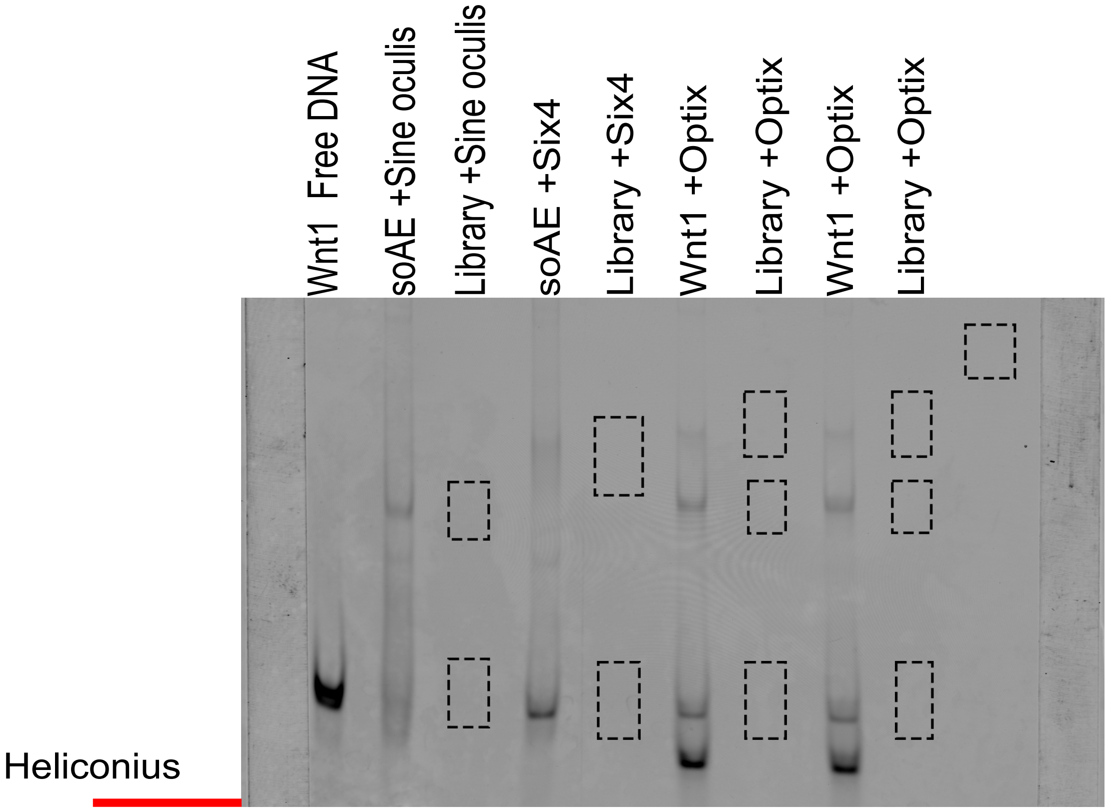

SINE_Amphimedon
SciSixB
LCOSixB
C_elegans_ceh_33
C_elegans_ceh_34
Nvec_Six1
SIX1_Strongylocentrotus
Six1_2_Halocynthia
Six1_2_Branchiostoma
Six1_Petromyzon
Six1_like_PetromyzonX1
Six1_like_Petromyzon
SIX1_Rhincodon
SIX1_Callorhinchus
SIX1b_Spotted
SIX1_HUMAN
SIX1_Erpetoichthys
Six2_like_Petromyzon
Spotted_Gar_SIX1_like
SIX2_Rhincodon
SIX2_Spotted
SIX2a_Erpetoichthys
SIX2_HUMAN
SIX2_Callorhinchus
Sine_Drosophila
So_H_erato
SINE_Daphnia
Cwil_SIXC
Sine_Schmidtea
SIX2_Brachionus
SIX1_Pomacea
SINE_Capitella_teleta
Nvec_Six3
Tri_SIX3
C_elegans_ceh_32
Six3_Schmidtea
SIX3_Brachionus
SIX6_Strongylocentrotus
Six3_6_Halocynthia
SIX6_Pomacea
SIX6_Petromyzon
Six6_like_Petromyzon
Six6_like_Petromyzon3
Six6_like_Petromyzon2
SIX3_Rhincodon
SIX3_Callorhinchus
SIX3_HUMAN
SIX3_Spotted
SIX3_Erpetoichthys
SIX7_Spotted
SIX7_Erpetoichthys
SIX6_Rhincodon
SIX6_Callorhinchus
SIX6_Spotted
SIX6_Erpetoichthys
SIX6_HUMAN
Optix_Capitella_teleta
Optix_H_erato
Optix_Drosophila
OPTIX_Daphnia
Cwil_SIXA
Nvec_Six4
SIX1_Brachionus
SIX4_Pomacea
SIX4_Capitella_teleta
Six4_5_Branchiostoma
Six4_5_Halocynthia
SIX4_Petromyzon
SIX5_Spotted
SIX5_Erpetoichthys
SIX5_HUMAN
SIX4_Rhincodon
SIX4a_Callorhinchus
SIX4_Spotted
SIX4_Erpetoichthys
SIX4_HUMAN
Tri_SIX1A
C_elegans_unc_39
SIX4_Strongylocentrotus
Six4_Drosophila
SIX4_H_mel
SIX4_Daphnia
Six2_like_PetromyzonX1
Cwil_SIXB
SciSixc
LCOSixc
2.0

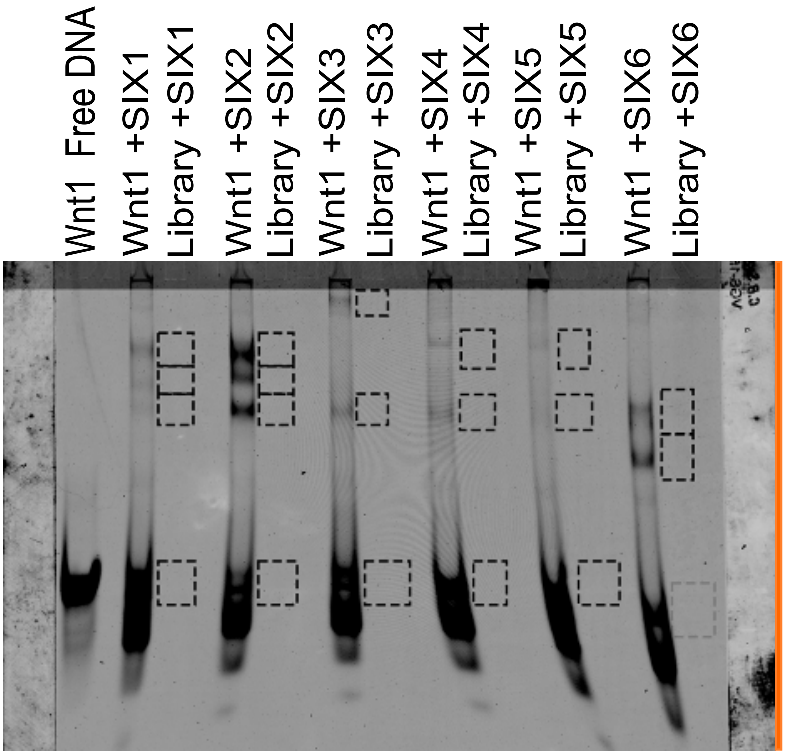
Supplementary Image 3: SIX HD hexapeptide for each subgroup and HD overall


SIX1/2
bits
4.0
3.0
2.0
1.0
0.0
5
GEETSY
SIX3/6
bits
4.0
3.0
2.0
1.0
0.0
5
GEQKTH
SIX4/5
bits
4.0
3.0
2.0
1.0
0.0
5
GEETVY
SIXHD
bits
4.0
3.0
2.0
1.0
0.0
5
GEETSY
QKTH

Supplementary Image 4: SELEX-seq gels for all proteins

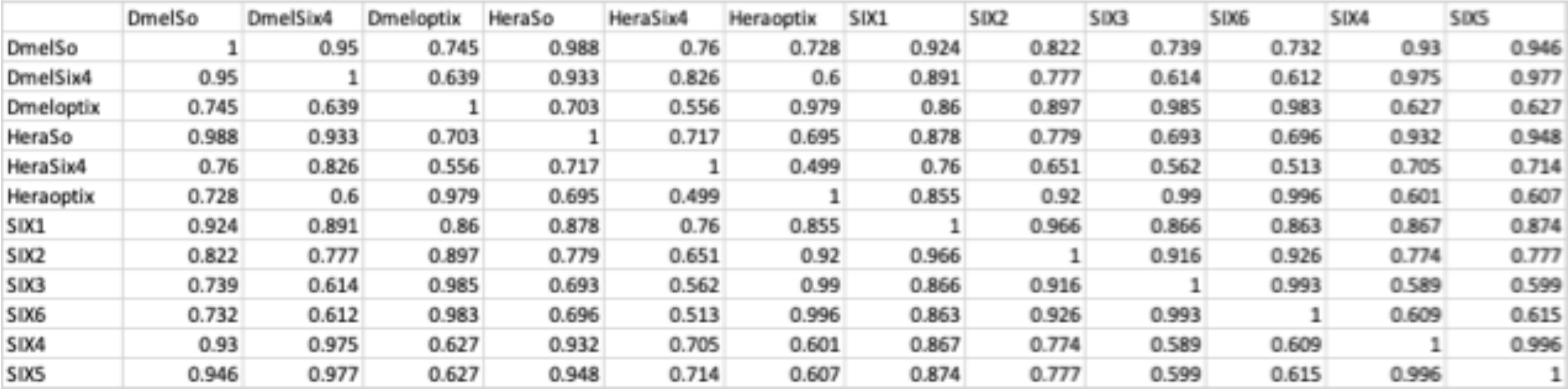

soAE Free DNA
soAE +Sine oculis
Library +Sine oculis
soAE +Six4
Library +Six4
soAE +Optix
Library +Optix
soAE +GFP
Library +GFP
Drosophila


Wnt1 Free DNA
soAE +Sine oculis
Library +Sine oculis
soAE +Six4
Library +Six4
Wnt1 +Optix
Library +Optix
Wnt1 +Optix
Library +Optix
Heliconius



Homo sapiens

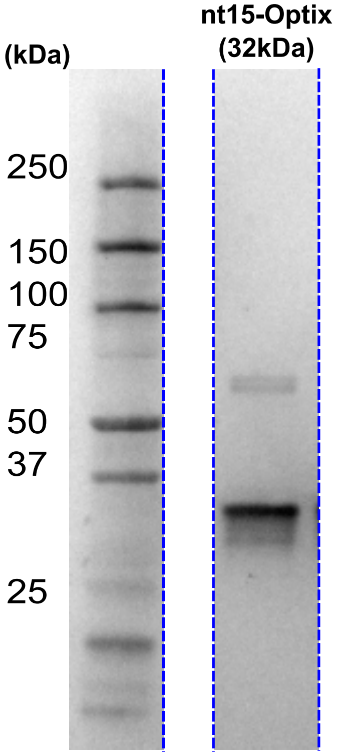
Supplementary information 5: Binding Matrices Correlations



	DmelSo	DmelSix4	Dmeloptix	HeraSo	HeraSix4	Heraoptix	SIX1	SIX2	SIX3	SIX6	SIX4	SIX5
DmelSo	1	0.95	0.745	0.988	0.76	0.728	0.924	0.822	0.739	0.732	0.93	0.946
DmelSix4	0.95	1	0.639	0.933	0.826	0.6	0.891	0.777	0.614	0.612	0.975	0.977
Dmeloptix	0.745	0.639	1	0.703	0.556	0.979	0.86	0.897	0.985	0.983	0.627	0.627
HeraSo	0.988	0.933	0.703	1	0.717	0.695	0.878	0.779	0.693	0.696	0.932	0.948
HeraSix4	0.76	0.826	0.556	0.717	1	0.499	0.76	0.651	0.562	0.513	0.705	0.714
Heraoptix	0.728	0.6	0.979	0.695	0.499	1	0.855	0.92	0.99	0.996	0.601	0.607
SIX1	0.924	0.891	0.86	0.878	0.76	0.855	1	0.966	0.866	0.863	0.867	0.874
SIX2	0.822	0.777	0.897	0.779	0.651	0.92	0.966	1	0.916	0.926	0.774	0.777
SIX3	0.739	0.614	0.985	0.693	0.562	0.99	0.866	0.916	1	0.993	0.589	0.599
SIX6	0.732	0.612	0.983	0.696	0.513	0.996	0.863	0.926	0.993	1	0.609	0.615
SIX4	0.93	0.975	0.627	0.932	0.705	0.601	0.867	0.774	0.589	0.609	1	0.996
SIX5	0.946	0.977	0.627	0.948	0.714	0.607	0.874	0.777	0.599	0.615	0.996	1

Supplementary Image 6: Anti-optix Western Blot to nt15-optix

Anti-optix Western Blot to nt15-optix



Bibliography

1. Charoensawan V, Wilson D, Teichmann SA. Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res*. 2010;38(21):7364–7377.
2. Inukai S, Kock KH, Bulyk ML. Transcription factor–DNA binding: beyond binding site motifs. *Curr Opin Genet Dev*. 2017;43:110–119.
3. Lambert SA, Yang AWH, Sasse A, et al. Similarity regression predicts evolution of transcription factor sequence specificity. *Nat Genet*. 2019;51(6):981–989.
4. DavidsonEH_GenomicContProc_2015_Book.
5. T.R. Hughes. A Handbook of Transcription Factors.
6. Stormo GD, Zhao Y. Determining the specificity of protein-DNA interactions. *Nat Rev Genet*. 2010;11(11):751–760.
7. Lambert SA, Jolma A, Campitelli LF, et al. The Human Transcription Factors. *Cell*. 2018;172(4):650–665.
8. Li Y, Chen C yu, Kaye AM, Wasserman WW. The identification of cis-regulatory elements: A review from a machine learning perspective. *BioSystems*. 2015;138:6–17.
9. Wittkopp PJ, Kalay G. Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet*. 2012;13(1):59–69.
10. Zeitlinger J. Seven myths of how transcription factors read the cis-regulatory code. *Curr Opin Syst Biol*. 2020;23:22–31.
11. Kolovos P, Knoch TA, Grosveld FG, Cook PR, Papantonis A. Enhancers and silencers: An integrated and simple model for their function. *Epigenetics Chromatin*. 2012;5(1):.
12. Brasset E, Vaury C. Insulators are fundamental components of the eukaryotic genomes. *Heredity (Edinb)*. 2005;94(6):571–576.
13. Stewart AJ, Plotkin JB. Why transcription factor binding sites are ten nucleotides long. *Genetics*. 2012;192(3):973–985.
14. Rohs R, Jin X, West SM, et al. Origins of specificity in protein-DNA recognition. *Annu Rev Biochem*. 2010;79:233–269.
15. Crocker J, Preger-Ben Noon E, Stern DL. The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution. *Curr Top Dev Biol*. 2016;117:455–469.
16. Stormo GD, Zhao Y. Determining the specificity of protein–DNA interactions. *Nat Rev Genet*. 2010;11(November):
17. Jolma A, Yin Y, Nitta KR, et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*. 2015;527(7578):384–388.
18. Rogers JM, Bulyk ML. Diversification of transcription factor–DNA interactions and the evolution of gene regulatory networks. *Wiley Interdiscip Rev Syst Biol Med*. 2018;10(5):.
19. Park PJ. ChIP-seq: Advantages and challenges of a maturing technology. *Nat Rev Genet*. 2009;10(10):669–680.
20. Slattery M, Zhou T, Yang L, et al. Absence of a simple code: How transcription factors read the genome. *Trends Biochem Sci*. 2014;39(9):381–399.
21. Levo M, Segal E. In pursuit of design principles of regulatory sequences. *Nat Rev Genet*. 2014;15(7):453–468.
22. Weirauch MT, Yang A, Albu M, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. 2014;158(6):1431–1443.

23. Jolma A, Kivioja T, Toivonen J, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*. 2010;20(6):861–873.
24. Slattery M, Riley T, Liu P, et al. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*. 2011;147(6):1270–82.
25. Riley TR, Slattery M, Abe N, et al. SELEX-seq: A Method for Characterizing the Complete Repertoire of Binding Site Preferences for Transcription Factor Complexes. 2014;255–278.
26. Nitta KR, Jolma A, Yin Y, et al. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution.
27. Jolma A, Yan J, Whitington T, et al. DNA-binding specificities of human transcription factors. *Cell*. 2013;152(1–2):327–339.
28. Stormo GD. Modeling the specificity of protein-DNA interactions. *Quantitative Biology*. 2013;1(2):115–130.
29. Blackwell TK, Kretzner L, Blackwood EM, Eisenman RN, Weintraub H. Sequence-specific DNA binding by the c-Myc protein. *Science (1979)*. 1990;250(494):1149–1151.
30. Marmorstein R, Fitzgerald MX. Modulation of DNA-binding domains for sequence-specific DNA recognition.
31. Nakagawa S, Gisselbrecht SS, Rogers JM, Hartl DL, Bulyk ML. DNA-binding specificity changes in the evolution of forkhead transcription factors.
32. Rogers JM, Waters CT, Seegar TCM, et al. Bispecific Forkhead Transcription Factor FoxN3 Recognizes Two Distinct Motifs with Different DNA Shapes. *Mol Cell*. 2019;74(2):245-253.e6.
33. Cheatle Jarvela AM, Brubaker L, Vedenko A, et al. Modular evolution of DNA-binding preference of a tbrain transcription factor provides a mechanism for modifying gene regulatory networks. *Mol Biol Evol*. 2014;31(10):2672–2688.
34. Noyes MB, Christensen RG, Wakabayashi A, et al. Analysis of Homeodomain Specificities Allows the Family-wide Prediction of Preferred Recognition Sites. *Cell*. 2008;133(7):1277–1289.
35. Bürglin TR, Affolter M. Homeodomain proteins: an update. *Chromosoma*. 2016;125(3):497–521.
36. Gehring W 1, Aff M, Burglin T. HOMEODOMAIN PROTEINS. 1994.
37. Scott MP, Weinert AJ. Structural relationships among genes that control development: Sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila (homoeosis/protein domains/gene evolution). 1984.
38. Carrasco AE, Mcginnis W, Gehring WJ, de Robertis EM. Cloning of an X. laevis Gene Expressed during Early Embryogenesis Coding for a Peptide Region Homologous to Drosophila Homeotic Genes. 1984.
39. Banerjee-Basu S, Baxevanis AD. Molecular evolution of the homeodomain family of transcription factors. 2001.
40. Berger MF, Badis G, Gehrke AR, et al. Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. *Cell*. 2008;133(7):1266–1276.
41. Milani R. Two new eye-shape mutant alleles in D. meanogaster. 1941;

42. Cheyette BNR, Green PJ, Martin K, et al. The Drosophila sine oculis locus Encodes a Homeodomain-Containing Protein Required for the Development of the Entire Visual System. 1994.
43. Guillermo Oliver, Wher Roland, Jenkins Nancy, Copeland Neal. Homeobox genes and connective tissue patterning. *Developlmental*. 1995;693–705.
44. Oliver G, Mailhos A, Wehr R, et al. Six3, a murine homologue of the sine oculis gene, demarcates the most anterior border of the developing neural plate and is expressed during eye development. *Development*. 1995;121(12):4045–55.
45. Oliver ’,~ G, L_A~osli B,~ F, K6ster R, Wittbrodt J, Gruss P. Ectopic lens induction in fish in response to the murine homeobox gene Six3. 1996.
46. Kawakami K, Ohto H, Ikeda K, Roeder RG. Structure, function and expression of a murine homeobox protein AREC3, a homologue of Drosophila sine oculis gene product, and implication in development. 1996.
47. Niiya A, Ohto H, Kawakami K, Araki M. Localization of Six4/AREC3 in the Developing Mouse Retina ; Implications in Mammalian Retinal Development †. Article Number; 1998.
48. Kawakamp K, Ohto H, Takizawa T, Saito T. Identification and expression of six family genes in mouse retina. 1996.
49. Toy J, Yang J, Leppert GS, Sundin OH. The Optx2 homeobox gene is expressed in early precursors of the eye and activates retina-specific genes. 1998.
50. Dominique Jean, Gilbert Bernier, Gruss Peter. Six6(Optx2) is a novel murineSix3-related homeobox gene thatdemarcates the presumptive pituitary/hypothalamic axisand the ventral optic stalk. *Mechanisms of Development* . 1999;(84):31–40.
51. Ogawa Y, Shiraki T, Kojima D, Fukada Y. Homeobox transcription factor Six7 governs expression of green opsin genes in zebrafish. *Proceedings of the Royal Society B: Biological Sciences*. 2015;282(1812):.
52. Ogawa Y, Shiraki T, Asano Y, et al. Six6 and Six7 coordinately regulate expression of middle-wavelength opsins in zebrafish. *Proc Natl Acad Sci U S A*. 2019;116(10):4651–4660.
53. Inbal A, Kim SH, Shin J, Solnica-Krezel L. Six3 Represses Nodal Activity to Establish Early Brain Asymmetry in Zebrafish. *Neuron*. 2007;55(3):407–415.
54. Seo H-C, Curtiss J, Mlodzik M, Fjose A. Six class homeobox genes in Drosophila belong to three distinct families and are involved in head development.
55. Patrick AN, Cabrera JH, Smith AL, et al. Structure-function analyses of the human SIX1-EYA2 complex reveal insights into metastasis and BOR syndrome. *Nat Struct Mol Biol*. 2013;20(4):447–453.
56. Kumar JP. The sine oculis homeobox (SIX) family of transcription factors as regulators of development and disease. *Cellular and Molecular Life Sciences*. 2009;66(4):565–583.
57. Patrick AN, Schiemann BJ, Yang K, Zhao R, Ford HL. Biochemical and functional characterization of six SIX1 Branchio-oto-renal syndrome mutations. *Journal of Biological Chemistry*. 2009;284(31):20781–20790.
58. Serikaku MA, O’tousa JE. sine oculis Is a Homeobox Gene Required for Drosophila Visual System Development. 1994.
59. Kawakami K. Six family genes--structure and function as transcription factors and their roles in development. *Nat Reviews Cancer*. 2016;22(7):616–626.

60. Pauli T, Seimiya M, Blanco J, Gehring WJ. Identification of functional sine oculis motifs in the autoregulatory element of its own gene, in the eyeless enhancer and in the signalling gene hedgehog. *Development*. 2005;132(12):2771–2782.

61. Santolini M, Sakakibara I, Gauthier M, et al. MyoD reprogramming requires Six1 and Six4 homeoproteins: genome-wide cis-regulatory module analysis. *Nucleic Acids Res*. 2016;44(18):8621–8640.

62. Hu S, Mamedova A, Hegde RS. DNA-binding and regulation mechanisms of the SIX family of retinal determination proteins. *Biochemistry*. 2008;47(11):3586–3594.

63. Weasner BP, Kumar JP. The non-conserved C-terminal segments of Sine Oculis Homeobox (SIX) proteins confer functional specificity. *Genesis*. 2009;47(8):514–523.

64. Redmond AK, McLysaght A. Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. *Nat Commun*. 2021;12(1):.

65. Musser JM, Schippers KJ, Nickel M, et al. Profiling cellular diversity in sponges informs animal cell type and nervous system evolution. 2021.

66. Fortunato SAV, Leininger S, Adamska M. Evolution of the Pax-Six-Eya-Dach network: The calcisponge case study. *Evodevo*. 2014;5(1):.

67. Leys SP, Cronin TW, Degnan BM, Marshall JN. Spectral sensitivity in a sponge larva. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol*. 2002;188(3):199–202.

68. Hoshiyama D, Iwabe N, Miyata T. Evolution of the gene families forming the Pax/Six regulatory network: Isolation of genes from primitive animals and molecular phylogenetic analyses. *FEBS Lett*. 2007;581(8):1639–1643.

69. Brodbeck S, Englert C. Genetic determination of nephrogenesis: The Pax/Eya/Six gene network. *Pediatric Nephrology*. 2004;19(3):249–255.

70. Furuya M, Qadota H, Chisholm AD, Sugimoto A. The C. elegans eyes absent ortholog EYA-1 is required for tissue differentiation and plays partially redundant roles with PAX-6. *Dev Biol*. 2005;286(2):452–463.

71. Amin NM, Lim SE, Shi H, Chan TL, Liu J. A conserved Six-Eya cassette acts downstream of Wnt signaling to direct non-myogenic versus myogenic fates in the C. elegans postembryonic mesoderm. *Dev Biol*. 2009;331(2):350–360.

72. Dozier C, Kagoshima H, Niklaus G, Cassata G, Bürglin TR. The Caenorhabditis elegans Six/sine oculis class homeobox gene ceh-32 is required for head morphogenesis. *Dev Biol*. 2001;236(2):289–303.

73. Fortunato SAV, Leininger S, Adamska M. Evolution of the Pax-Six-Eya-Dach network: The calcisponge case study. *Evodevo*. 2014;5(1):.

74. Yanowitz JL, Shakir MA, Hedgecock E, et al. UNC-39, the C. elegans homolog of the human myotonic dystrophy-associated homeodomain protein Six5, regulates cell motility and differentiation. *Dev Biol*. 2004;272(2):389–402.

75. Stierwald M, Yanze N, Bamert RP, Kammermeier L, Schmid V. The Sine oculis/Six class family of homeobox genes in jellyfish with and without eyes: Development and eye regeneration. *Dev Biol*. 2004;274(1):70–81.

76. Hroudova M, Vojta P, Strnad H, et al. Diversity, phylogeny and expression patterns of pou and six homeodomain transcription factors in hydrozoan jellyfish craspedacusta sowerbyi. *PLoS One*. 2012;7(4):.

77. Pineda D, Gonzalez J, Callaerts P, et al. Searching for the prototypic eye genetic network: Sine oculis is essential for eye regeneration in planarians. 1999.

78. Xu PX. The EYA-SO/SIX complex in development and disease. *Pediatric Nephrology*. 2013;28(6):843–854.

79. Laclef C, Souil E, Demignon J, Maire P. Thymus, kidney and craniofacial abnormalities in Six1 deficient mice. *Mech Dev*. 2003;

80. O'Brien LL, Guo Q, Lee YJ, et al. Differential regulation of mouse and human nephron progenitors by the six family of transcriptional regulators. *Development (Cambridge)*. 2016;143(4):595–608.

81. Xu J, Li J, Ramakrishnan A, et al. Six1 and Six2 of the Sine Oculis Homeobox Subfamily are Not Functionally Interchangeable in Mouse Nephron Formation. *Front Cell Dev Biol*. 2022;10:.

82. Meurer L, Ferdman L, Belcher B, Camarata T. The SIX Family of Transcription Factors: Common Themes Integrating Developmental and Cancer Biology. *Front Cell Dev Biol*. 2021;9:.

83. Fujimoto Y, Tanaka SS, Yamaguchi YL, et al. Homeoproteins Six1 and Six4 Regulate Male Sex Determination and Mouse Gonadal Development. *Dev Cell*. 2013;26(4):416–430.

84. Christensen KL, Patrick AN, McCoy EL, Ford HL. Chapter 5 The Six Family of Homeobox Genes in Development and Cancer. *Adv Cancer Res*. 2008;101:93–126.

85. The Drosophila homeobox gene optix is capable of inducing ectopic eyes by an eyeless-independent mechanism(2000).

86. Domınguez-Cejudo MA, Casares F. Anteroposterior patterning of Drosophila ocelli requires an anti-repressor mechanism within the hh pathway mediated by the Six3 gene Optix. *Development (Cambridge)*. 2015;142(16):2801–2809.

87. al Khatib A, Siomava N, Iannini A, Posnien N, Casares F. Specific expression and function of the Six3 optix in Drosophila serially homologous organs. *Biol Open*. 2017;6(8):1155–1164.

88. Monteiro A. Gene regulatory networks reused to build novel traits: Co-option of an eye-related gene regulatory network in eye-like organs and red wing patches on insect wings is suggested by optix expression. *BioEssays*. 2012;34(3):181–186.

89. Reed RD, Papa R, Martin A, et al. Optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science (1979)*. 2011;333(6046):1137–1141.

90. Zhang L, Mazo-Vargas A, Reed RD. Single master regulatory gene coordinates the evolution and development of butterfly color and iridescence. *Proceedings of the National Academy of Sciences*. 2017;201709058.

91. van Belleghem SM, Lewis JJ, Rivera ES, Papa R. Heliconius butterflies: a window into the evolution and development of diversity. *Curr Opin Genet Dev*. 2021;69:72–81.

92. Geng X, Acosta S, Lagutin O, Gil HJ, Oliver G. Six3 dosage mediates the pathogenesis of holoprosencephaly. *Development (Cambridge)*. 2016;143(23):4462–4473.

93. Wallis DE, Roessler E, Hehr U, et al. Mutations in the homeodomain of the human SIX3 gene cause holoprosencephaly. 1999.

94. Pasquier L, Dubourg C, Blayau M, et al. A new mutation in the six-domain of SIX3 gene causes holoprosencephaly. 2000.

95. Turcu DC, Lillehaug JR, Seo H-C. SIX3 and SIX6 interact with GEMININ via C-terminal regions. *Biochem Biophys Rep*. 2019;20:100695.

96. Zheng Y, Zeng Y, Qiu R, et al. The homeotic protein SIX3 suppresses carcinogenesis and metastasis through recruiting the LSD1/NuRD(MTA3) complex. *Theranostics*. 2018;8(4):972–989.
97. Mohanty K, Dada R, Dada T. Identification and genotype phenotype correlation of novel mutations in SIX6 gene in primary open angle glaucoma. *Ophthalmic Genet*. 2018;39(3):366–372.
98. Kirby RJ, Hamilton GM, Finnegan DJ, Johnson KJ, Jarman AP. Drosophila homolog of the myotonic dystrophy-associated gene, SIX5, is required for muscle and gonad development. 2005.
99. Clark IBN, Boyd J, Hamilton G, Finnegan DJ, Jarman AP. D-six4 plays a key role in patterning cell identities deriving from the Drosophila mesoderm. *Dev Biol*. 2006;294(1):220–231.
100. Wang J, Liu M, Zhao L, et al. Disabling of nephrogenesis in porcine embryos via CRISPR/Cas9-mediated SIX1 and SIX4 gene targeting. *Xenotransplantation*. 2019;26(3):.
101. Tang X, Yang Y, Song X, et al. SIX4 acts as a master regulator of oncogenes that promotes tumorigenesis in non-small-cell lung cancer cells. *Biochem Biophys Res Commun*. 2019;516(3):851–857.
102. Wakimoto H, Maguire CT, Sherwood MC, et al. Characterization of Cardiac Conduction System Abnormalities in Mice with Targeted Disruption of Six5 Gene. Kluwer Academic Publishers; 2002.
103. Ohno S. Evolution by Gene Duplication. Springer Berlin Heidelberg; 1970.
104. Jones DM, Vandepoele K. Identification and evolution of gene regulatory networks: insights from comparative studies in plants. *Curr Opin Plant Biol*. 2020;54:42–48.
105. Singh NP, de Kumar B, Paulson A, et al. A six-amino-acid motif is a major determinant in functional evolution of HOX1 proteins. *Genes Dev*. 2020;34(23–24):1680–1696.
106. Sebé-Pedrós A, Ariza-Cosano A, Weirauch MT, et al. Early evolution of the T-box transcription factor family. *Proc Natl Acad Sci U S A*. 2013;110(40):16050–16055.
107. Sayou C, Monniaux M, Nanao M, et al. A Promiscuous Intermediate Underliesthe Evolution of LEAFY DNABinding Specificity. *Science (1979)*. 2014;343(6171):645–649.
108. Siddiq MA, Hochberg GK, Thornton JW. Evolution of protein specificity: insights from ancestral protein reconstruction. *Curr Opin Struct Biol*. 2017;47:113–122.
109. Zhang J. Evolution by gene duplication: An update. *Trends Ecol Evol*. 2003;18(6):292–298.
110. Hsia CC, McGinnis W. Evolution of transcription factor function. *Curr Opin Genet Dev*. 2003;13(2):199–206.
111. Dowell RD. Transcription factor binding variation in the evolution of gene regulation. *Trends in Genetics*. 2010;26(11):468–475.
112. Opazo JC, Kuraku S, Zavala K, Toloza-Villalobos J, Hoffmann FG. Evolution of nodal and nodal-related genes and the putative composition of the heterodimers that trigger the nodal pathway in vertebrates. *Evol Dev*. 2019;21(4):205–217.
113. Hoffmann FG, Opazo JC, Storz JF. Gene cooption and convergent evolution of oxygen transport hemoglobins in jawed and jawless vertebrates. *Proc Natl Acad Sci U S A*. 2010;107(32):14274–14279.
114. Hoffmann FG, Vandewege MW, Storz JF, Opazo JC. Gene turnover and diversificationof the α-and β-GlobinGene families in sauropsid vertebrates. *Genome Biol Evol*. 2018;10(1):344–358.

115. Kenny NJ, Chan KW, Nong W, et al. Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. *Heredity (Edinb)*. 2016;116(2):190–199.

116. Bebenek IG, Gates RD, Morris J, Hartenstein V, Jacobs DK. Sine oculis in basal Metazoa. *Dev Genes Evol*. 2004;214(7):342–351.

117. Ryan JF, Pang K, Comparative N, et al. The homeodomain complement of the ctenophore Mnemiopsis leidyi suggests that Ctenophora and Porifera diverged prior to the ParaHoxozoa. 2010.

118. Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. *Nucleic Acids Res*. 2022;50(D1):D988–D995.

119. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 2019;47(D1):D506–D515.

120. Sayers EW, Bolton EE, Brister JR, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2022;50(D1):D20–D26.

121. Mcclean P. BLAST Basic Local Alignment Search Tool. 2004.

122. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 2018;35(6):1547–1549.

123. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792–1797.

124. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32(1):268–274.

125. Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res*. 2016;44(W1):W232–W235.

126. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;14(6):587–589.

127. Thi Hoang D, Chernomor O, von Haeseler A, et al. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol*. 2017;35(2):518–522.

128. Huson DH, Richter DC, Rausch C, et al. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*. 2007;8:.

129. Holland PWH. Evolution of homeobox genes. *Wiley Interdiscip Rev Dev Biol*. 2013;2(1):31–45.

130. Holland PWH, Booth HAF, Bruford EA. Classification and nomenclature of all human homeobox genes. *BMC Biol*. 2007;5:.

131. Gibson DG, Young L, Chuang R-Y, et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*. 2009;6(5):343–345.

132. Jeong Y, Leskow FC, El-Jaick K, et al. Regulation of a remote Shh forebrain enhancer by the Six3 homeoprotein. *Nat Genet*. 2008;40(11):1348–1353.

133. Slattery M, Riley T, Liu P, et al. Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. *Cell*. 2011;147(6):1270–1282.

134. Jusiak B, Karandikar UC, Kwak S-J, et al. Regulation of Drosophila Eye Development by the Transcription Factor Sine oculis. *PLoS One*. 2014;9(2):e89695.

135. Kronforst MR, Papa R. The functional basis of wing patterning in Heliconius butterflies: The molecules behind mimicry. *Genetics*. 2015;200(1):1–19.

136. Hines HM, Papa R, Ruiz M, et al. Transcriptome analysis reveals novel patterning and pigmentation genes underlying Heliconius butterfly wing pattern variation. *BMC Genomics*. 2012;13(1):.

137. Kronforst MR, Papa R. The functional basis of wing patterning in Heliconius butterflies: The molecules behind mimicry. *Genetics*. 2015;200(1):1–19.

138. van Belleghem SM, Rastas P, Papanicolaou A, et al. Complex modular architecture around a simple toolkit of wing pattern genes. *Nat Ecol Evol*. 2017;1(3):.

139. Mazo-Vargas A, Concha C, Livraghi L, et al. Macroevolutionary shifts of *WntA* function potentiate butterfly wing-pattern diversity. *Proceedings of the National Academy of Sciences*. 2017;2(20):201708149.

140. Fenner J, Benson C, Rodriguez-Caro L, et al. Wnt Genes in Wing Pattern Development of Coliadinae Butterflies. *Front Ecol Evol*. 2020;8:.

141. Papa R, Kapan DD, Counterman BA, et al. Multi-Allelic Major Effect Genes Interact with Minor Effect QTLs to Control Adaptive Color Pattern Variation in Heliconius erato. *PLoS One*. 2013;8(3):.

142. Hines HM, Papa R, Ruiz M, et al. Transcriptome analysis reveals novel patterning and pigmentation genes underlying Heliconius butterfly wing pattern variation. *BMC Genomics*. 2012;13(1):.

143. James J. Lewis, Rachel C. Geltman, Patrick C. Pollak, et al. Parallel evolution of ancient, pleiotropic enhancers underlies butterfly wing pattern mimicry. *PNAS*. 2019;12:.

144. Insausti TC, Casas J. The functional morphology of color changing in a spider: Development of ommochrome pigment granules. *Journal of Experimental Biology*. 2008;211(5):780–789.

145. Ryall RL, Howells AJ. OMMOCHROME BIOSYNTHETIC PATHWAY OF DROSOPHILA MELANOGASTER: VARIATIONS IN LEVELS OF ENZYME ACTIVITIES AND INTERMEDIATES DURING ADULT DEVELOPMENT. Pergamon Press; 1974.

146. Linzen B. The Tryptophan- Ommochrome Pathway in Insects. 1974;

147. Dinwiddie A, Rachootin S. Patterning of a compound eye on an extinct dipteran wing. *Biol Lett*. 2011;7(2):281–284.

148. Jiggins CD, Wallbank RWR, Hanly JJ. Waiting in the wings: What can we learn about gene co-option from the diversification of butterfly wing patterns? *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2017;372(1713):.

149. Toivonen J, Das PK, Taipale J, et al. MODER2: First-order Markov modeling and discovery of monomeric and dimeric binding motifs. *Bioinformatics*. 2020;36(9):2690–2696.

150. Korhonen JH, Palin K, Taipale J, Ukkonen E. Fast motif matching revisited: High-order PWMs, SNPs and indels. *Bioinformatics*. 2017;33(4):514–521.

151. Pizzi C, Rastas P, Ukkonen E. Finding significant matches of position weight matrices in linear time. *IEEE/ACM Trans Comput Biol Bioinform*. 2011;8(1):69–79.

152. Korhonen J, Martinmäki P, Pizzi C, Rastas P, Ukkonen E. MOODS: Fast search for position weight matrix matches in DNA sequences. *Bioinformatics*. 2009;25(23):3181–3182.

153. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24–26.

154. Salomone J, Qin S, Fufa TD, et al. Conserved Gsx2/Ind homeodomain monomer versus homodimer DNA binding defines regulatory outcomes in flies and mice. *Genes Dev*. 2021;35(1):157–174.
155. Choi Y, Luo Y, Lee S, et al. FOXL2 and FOXA1 cooperatively assemble on the TP53 promoter in alternative dimer configurations. *Nucleic Acids Res*. 2022;50(15):8929–8946.