EVOLUTION OF COMPONENTS OF GENE REGULATION IN *DROSOPHILA*

AND MAMMALS

A Thesis in

Biology

by

Emmanouil T. Dermitzakis

Copyright 2001 Emmanouil T. Dermitzakis

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

December 2001
We approve the thesis of Emmanouil T. Dermitzakis

Andrew G. Clark
Professor of Biology
Thesis Adviser
Chair of Committee

Hiroshi Akashi
Assistant Professor of Biology

Ross C. Hardison
Professor of Biochemistry

Steven W. Schaeffer
Associate Professor of Biology

Esther Siegfried
Assistant Professor of Biology
and Biochemistry

Douglas R. Cavener
Professor
Head of the Department of Biology
ABSTRACT.

Gene regulation is a very critical property of a gene s function. This process of gene regulation is achieved through protein-DNA and protein-protein interactions. The components of gene regulatory mechanisms are the DNA sequences that contain the signals for the accurate temporal and spatial expression of the gene they regulate and the proteins that fulfill the instructions contained within regulatory sequences. Usually a large number of proteins interact in regulatory regions to differentially regulate the gene in different tissues or by different environmental and developmental stimuli.

Phenotypic variation is abundant in nature and a fraction of this variation is harbored in components of the regulation of the genes controlling the phenotype of interest. While its importance for gene function and its potential for understanding natural phenotypic variation is widely appreciated, little is known about the evolutionary dynamics of the components of gene regulation. Understanding the processes and dynamics of evolution and natural variation of components of regulatory networks will be essential in understanding the evolutionary transitions in many variable traits in nature.

In the chapters that follow I attempt to approach the study of the evolutionary dynamics of gene regulation from various directions and ultimately improve our knowledge on how components of gene regulation vary within and between species.

In chapter 1 I look at the level of polymorphism in regulatory sequences of early development genes in *Drosophila*. I have sequenced 5 regulatory regions in several lines of *Drosophila melanogaster* from Central Pennsylvania and Africa and quantify the level of nucleotide variation within these regions. I apply traditional methods of analysis of
polymorphism and divergence data to understand the potential evolutionary forces that are acting in these sequences. I also use some newly developed methods contrasting the patterns of polymorphism and divergence and apply models of binding site prediction to identify polymorphisms that could potentially have phenotypic effect. Finally I attempt to associate levels of expression of two of these early development genes with nucleotide polymorphism in the promoter regions of these genes.

Chapter 2 is an extension on chapter 1 in the sense that I look at more distant sequence comparisons within the melanogaster subgroup and quantify the level and pattern of evolution of regulatory sequences among several species of Drosophila. I introduce a novel approach for the study of such sequences by applying binding site prediction models and comparing the amount and location of predicted binding sites for transcription factors that are known to regulate the genes under study. I also introduce a novel approach for binding site prediction that appears to be more efficient in identifying binding sites of weaker effect or binding sites that have the potential to become functional. Finally, simulations of sequence divergence reveal the evolutionary potential of regulatory regions of early development genes in Drosophila.

To test whether the observations from Drosophila stand in other taxonomic groups I extended my study to mammalian regulatory regions. In chapter 3 I searched the primary literature and mapped transcription factor binding sites in regulatory regions of human, other primates and mouse regulatory sequences. I quantified the level of divergence within the binding sites relative to the remainder of the sequence and compared these divergence levels between binding sites that share function between species vs. binding sites that are species-specific. By using a combination of statistical
approaches I estimate the degree of conservation of functional binding sites and the rate of turnover between human and rodents. These results have important implications for the annotation of regulatory sequences and for the understanding of regulatory evolution.

Regulatory evolution consists of both regulatory sequences and transcription factors. Chapter 4 addresses the question of the evolution of transcription factors. I devised a new method for the study of functional divergence of duplicated genes and test this method in transcription factors of human and mouse. I compare the pattern of amino acid divergence in DNA binding vs. trans-activation domains and attempt to detect where the functional divergence of duplicated genes occurs. I obtain some interesting insights about the evolution of the two different types of domains, which have important bearing on the understanding of how regulatory regions evolve and how their evolution correlates with the evolution of transcription factors.

Good understanding of the evolution of regulatory sequences relies on good understanding of the functional properties of these sequences. Binding site prediction is one such method that improves our understanding of regulatory sequences. In chapter 5, I take one step further in binding site prediction models and test the features of the internal structure of transcription binding site sequences and how they can be used to improve prediction models. I test interactions between adjacent positions within binding sites of well-characterized transcription factors as well as simulated sets of binding sites by means of statistical tests and I incorporate these properties in binding site prediction models. These models appear to describe the binding site features more efficiently and reduce the degree of false positives.
The results presented below will contribute to the understanding of regulatory evolution and provide some additional ground for subsequent research.

TABLE OF CONTENTS.

LIST OF FIGURES ix
LIST OF TABLES xi
ACKNOWLEDGEMENTS xii

Chapter 1: PATTERNS OF NUCLEOTIDE POLYMORPHISM AND DIVERGENCE IN REGULATORY SEQUENCES OF DROSOPHILA REVEAL STRONG SELECTIVE CONSTRAINTS 1

Abstract 2
Introduction 3
Methods 7
Results 11
Discussion 17
References 21

Chapter 2: HIGH EVOLUTIONARY POTENTIAL IN REGULATORY SEQUENCES OF EARLY DEVELOPMENT GENES IN DROSOPHILA 31

Abstract 32
Introduction 33
Chapter 3: EVOLUTION OF TRANSCRIPTION FACTOR BINDING SITES IN MAMMALIAN REGULATORY REGIONS: CONSERVATION AND TURNOVER.

Abstract
Introduction
Methods
Results
Discussion
References

Chapter 4: DIFFERENTIAL SELECTION AFTER DUPLICATION IN MAMMALIAN DEVELOPMENTAL GENES

Abstract
Introduction
Methods
Results/Discussion
References
Chapter 5: EVIDENCE FOR INTERACTION OF NUCLEOTIDE POSITIONS WITHIN TRANSCRIPTION FACTOR BINDING SITES: APPLICATIONS FOR BINDING SITE PREDICTION

Abstract
Introduction
Methods
Results
Discussion
References

Appendix A: Supplementary information for Chapter 1
Appendix B: Supplementary information for Chapter 2
Appendix C: Supplementary information for Chapter 3
LIST OF FIGURES:

Chapter 1:
Figure 1.1: Nucleotide variation in regulatory regions 27
Figure 1.2: Model 28
Figure 1.3: Binding site polymorphisms 29
Figure 1.4: Phenotypic variation 30

Chapter 2:
Figure 2.1: Regions studied 62
Figure 2.2: Binding site alignments 63
Figure 2.3: Distribution of divergence 64
Figure 2.4: Empirical null distributions of L-values 65
Figure 2.5: BCD binding sites predicted 66
Figure 2.6: Phylogenetic history of predicted binding sites 67
Figure 2.7: Counts of predicted sites in simulated data 69

Chapter 3:
Figure 3.1: Regulatory regions with available functional data 87
Figure 3.2: Divergence and variance in human-macaque comparisons 89
Figure 3.3: Divergence in human-mouse comparisons 91
Chapter 4:

Figure 4.1: Graphical description of the test 111
Figure 4.2: Hypothetical tree structure for two domains 112
Figure 4.3: Power of the test 113
Figure 4.4: Plot of D-values 114

Chapter 5:

Figure 5.1: Position weight matrices 134
Figure 5.2: Graphical description of the calculation of probability 135
Figure 5.3: Regression 136
Figure 5.4: Pattern of predicted binding sites 137
LIST OF TABLES.

Chapter 1:
Table 1.1: Summary of polymorphism data 26

Chapter 2:
Table 2.1: Data used in this study 59
Table 2.2: Pairwise divergence 60
Table 2.3: Comparison of binding site prediction methods 61

Chapter 3:
Table 3.1: Summary of the data used in this study 86

Chapter 4:
Table 4.1: Data of all pairwise comparisons 110

Chapter 1:
Table 5.1: Transcription factor binding site information 133
ACKNOWLEDGEMENTS.

Many people have made the completion of this thesis possible. First of all I would like to thank my thesis advisor Andy Clark for the very inspiring and exciting moments we shared. He offered me every possible way of support to accomplish this work. He was always a great guidance and constructive challenge for me and our discussions led to most of the ideas presented in this thesis. I also want to thank him for always being understanding and helpful in many personal moments. He will always be an example and guide in my career.

I would also like to thank my committee members, Hiroshi Akashi, Esther Siegfried, Steve Schaeffer, Ross Hardison and a former member Tom Whittam for their continuous support and inspiration during this work.

Many thanks to the past and present Clark lab members and most of all to Kristi Montooth, Brian Lazzaro, Beto Civetta, Malia Fullerton, Sarah Tishkoff. I had great moments with all of the and very exciting collaborations.

I also want to thank Ken Weiss and Webb Miller for many constructive discussions and ideas that made this work more complete and Casey Bergman for providing some of his binding site data.

I am also indebted to the biology department staff for helping me in various ways during my stay in the department and for providing me with all the necessary tools to accomplish my research.

Financial support for this project was offered by a National Science Foundation Doctoral Dissertation Improvement Grant, by a Penn State Life Sciences Consortium Innovative Research Grant and by a Life Science Consortium graduate fellowship.
I also want to thank my Masters advisor Lefteri Zouros for always being an inspiring figure in my career and for his help to come to Penn State.

Finally, I want to thank my girlfriend Rita for her love and support in every possible way she could and for being next to me during this exciting time of my life. And of course my parents Niki and Fefos and my sister Sophia who have always been next to me, although a few thousand miles away, and have fully supported me and encouraged me in every step of my life.
To: Rita

To: Niki
Fefos
Sophia
CHAPTER 1: Patterns of nucleotide polymorphism and divergence in regulatory sequences of *Drosophila* reveal strong selective constraints
1.1 Abstract.

Nucleotide variation has been studied extensively in *Drosophila*. However, very few studies have quantified nucleotide variation in regulatory regions of genes. In this report we surveyed five promoter regions of genes involved in the early development of *Drosophila melanogaster*, *bicoid*, *nanos*, *engrailed*, *fushi tarazu* and *hunchback* distal promoter. We sequenced alleles from a Pennsylvania population and compared their pattern of polymorphism to African alleles and *D. simulans*. Our results agree with previous studies and suggest that there are strong selective constraints within these regions with the exception of *nanos* promoter, which seems to be under different selective pressures. We investigate the pattern of polymorphism and divergence and propose some ways for the study of polymorphism within regulatory regions, which generally have less well-defined rules of evolution. We also quantified the expression levels of *bicoid* and *nanos* in 56 3rd chromosome extracted lines from Pennsylvania but none of the polymorphic sites scored in the promoters of these genes appears to be associated with variation in gene expression. We discuss the implications of our results for the understanding of the pattern of nucleotide and phenotypic variation in regulatory regions.
1.2 Introduction.

Variation in regulatory networks is a possible source of variation for many traits in \textit{Drosophila}. Studies have demonstrated that DNA regions other than the coding regions of genes could be responsible for a large fraction of this phenotypic variation (Long et al. 1998; Gibson and Hogness 1996; Zimmerman et al. 2000). Such regions could be the regulatory regions of the gene which could be anywhere (upstream or downstream) from the coding region (Blackwood and Kadonaga 1998) or regions of the transcript that are not coding but are responsible for its stability and processing before translation (\textit{bicoid}, \textit{nanos}; Rivera-Pomar and Jackle 1996; Wharton and Struhl 1991). For quantitative traits the phenotypic measurement is the amount rather than presence or absence of the trait, and therefore it is expected that regulatory regions are contributing to phenotypic variation since different levels of expression of the genes controlling the character can sometimes be responsible for such quantitative variation.

The early development of \textit{Drosophila} is controlled by a rather complex regulatory network that initiates with genes maternally expressed, such as \textit{bicoid}, \textit{hunchback}, \textit{nanos}, \textit{caudal} etc, which subsequently regulate directly or indirectly the expression of the first wave of zygotic genes such as \textit{fushi tarazu}, \textit{engrailed}, \textit{knirps}, \textit{even skipped} (Rivera-Pomar and Jackle 1995). The maternal products are distributed in a polarized manner within the embryo of \textit{Drosophila} forming gradients either from the anterior to the posterior (\textit{bcd}, \textit{hb}) or from the posterior to the anterior (\textit{nos}, \textit{cad}) and they are called
morphogens (Tabata 2001). The relative concentration of these and other morphogens defines the set of genes that will be expressed and the degree and time at which they will be expressed (Berleth et al. 1988; Frohnofer et al. 1986; Struhl 1989; Tautz 1988). The presence of multiple morphogens results in the expression of some of the downstream genes in well-defined stripes (ftz, en, eve) or larger portions of the embryo (kni) (Frasch and Levine 1987; Nusslein-Volhard 1989; Burz et al. 1998). Models have been proposed that it is a combination of anterior-posterior and dorsal-ventral regulatory networks that generate such a well organized and defined but at the same time cumbersome pattern. It is evident that the regulatory regions of these genes play a very important role in the way the pattern is created and maintained (Tautz 1988; Rivera-Pomar and Jackle 1995). In addition, the redundancy of the system with multiple regulators and compensation mechanisms (Namba et al. 1997) may result in tolerance of variation in the expression levels of these genes due to compensatory changes (Gibson 1996).

Regulatory regions consist of short sequence elements that usually serve as transcription factor binding sites. Each transcription factor has specific sequence requirements for DNA binding. These sequence requirements vary widely at the length of the sequence as well as the strictness of fit to a consensus (Stormo 1998; Claverie 1994). This latter property, the redundancy of sequence requirements, results in the tolerance of a large number of mutations that although they alter the primary sequence do not influence the binding ability of the factor to the sequence (Chapter 2). The fact that different factors have different degrees of redundancy should result in different rates of conservative mutations (mutations that do not alter binding efficiency) for different transcription factor binding sites, although this has not been formally tested yet.
Properties of binding sequence requirements are usually represented with a position weight matrix, which describes the probability of occurrence of each of the four nucleotides in each nucleotide position of the binding site (Berg and von Hippel 1987; Stormo 1998). Under a model of independence of positions across the binding site, one can assume that many sequences can have the same probability of occurrence under the weight matrix probabilities. In another study (Chapter 5) we have shown that this is not always true and that a simple position weight matrix is not always adequate to describe all the sequence properties of the binding sites of a given transcription factor. We have devised a more complex model, incorporating the interactions between adjacent positions, which appears to be more appropriate for many transcription factors. Other studies have also incorporated cooperative binding and the relative distance from the transcription initiation site as parameters that improve the identification of transcription factor binding sites in silico (GuhaThakurta and Stromo 2001; Liu et al. 2001; Benos et al. 2001).

All the above characteristics of regulatory elements make the study of regulatory regions very complex and the interpretation of the pattern of polymorphism and divergence cumbersome. In addition to the study of polymorphism within the already experimentally verified regulatory elements, which are only part of the regulatory landscape of the gene, one has to use models of binding site prediction to improve the understanding of whether some nucleotide changes could contribute to a change in the pattern or level of expression of the gene under study (Chapter 2). We have already presented analysis of the divergence pattern within the melanogaster subgroup showing that there is an abundance of potentially functional binding sites for known regulators
within regulatory regions of five genes involved in the early *Drosophila* development (Chapter 2).

Nucleotide polymorphism in regulatory regions has not been thoroughly studied and only two studies have described levels of nucleotide variation in regulatory regions of early development genes in *Drosophila*: the *eve* stripe-2 enhancer (Ludwig and Kreitman 1995) and *hunchback* proximal promoter (Tautz and Nigro1998). They observed low levels of polymorphism within these regions and concluded that these regions are likely under strong purifying selection. However, if one looks at the pattern of divergence there is clearly a high degree of sequence rearrangement, which in many cases results in gains or losses of some of the functional regulatory elements, experimentally verified in *D. melanogaster* (Ludwig et al. 1998). One of the most intriguing cases is the study by Ludwig and colleagues (2000), which neatly demonstrated that although the *eve* stripe-2 enhancer has retained identical function in both *D. melanogaster* and *D. pseudoobscura* as a whole, the internal organization has changed substantially suggesting compensatory evolutionary changes of regulatory elements.

This contrasting pattern between polymorphism and divergence raises questions about the process of evolution of regulatory regions, especially those of early development in *Drosophila*. We surveyed nucleotide sequence variation in five regulatory regions, *hunchback* distal promoter, *bicoid, nanos, engrailed* and *fushi tarazu*. We also obtained the *D. simulans* sequence for the same regions to contrast patterns of nucleotide polymorphism with divergence in. We present analysis of the pattern of polymorphism in these five regions, we compare our results to the previous studies and
we discuss the implications of our study for the interpretation and dynamics of evolution of early development regulatory sequences.

1.3 Methods.

**PCR and sequencing oligonucleotides:** The oligonucleotides used for PCR and sequencing reactions are available upon request.

**DNA sequencing:** Sequencing reactions were done using the CEQ 2000 sequencing kit. Templates for sequencing reactions were PCR products that were either ethanol-precipitated or column-purified. Reactions were cleaned and processed according to the commercial protocol and were run in CEQ 2000 8-capillary automated sequencer. For each line and for the *D. simulans* sequences both strands were sequenced to verify polymorphic and divergent sites. Sequences are submitted to the Genbank and PopSet databases.

**Statistical tests and metrics of variation:** Nucleotide variation was measured using the two different metrics of average heterozygosity, one based on the number of segregating sites (\(\ell\)) and the other based on the average number of pairwise nucleotide differences (\(\pi\)). Linkage disequilibrium, tests of neutral evolution (Tajima 1989, Fu and Li 1993, Hudson et al. 1987, Fay and Wu 2000) were done using the freely available software DnaSP version 3.51 (Rozas and Rozas 1999) and a program kindly provided by Brian P. Lazzaro. Coalescent simulations were done either using the routine in DnaSP or the
program available in the web site of Prof. Richard R. Hudson (Hudson 1990) at the University of Chicago. Heterogeneity of patterns of divergence and polymorphism within the region were tested using a method described in a previous study (Dermitzakis and Clark 2001) originally used to test heterogeneity of amino acid substitutions between paralogous genes. Under neutrality there should be no difference in the distribution of polymorphic and fixed sites along the sequence. The test can be performed using both tails of the permuted distribution. By using the opposite tail from the one used in Dermitzakis and Clark (2001) we are able to test the alternative hypothesis that the interspecific substitutions and the polymorphic sites are distributed similarly. We will explain below why this hypothesis is relevant for the study of regulatory regions. Fisher s combined probability was calculated according to Sokal and Rohlf (1997)

**Binding site prediction**: Binding sites were predicted using the method described in Chapter 2 using the same weight matrices and the same thresholds of significance for BCD and HB. Binding site prediction was done for each allele independently.

**Expression assays**: Total RNA was extracted from 56 3\textsuperscript{rd} chromosome extracted lines from Central Pennsylvania that had been backcrossed for five generations to the TM3/TM6 *D. melanogaster* balancer stock in order to make the remainder of the genome isogenic. Total RNA was extracted from whole fly abdomens from 5 flies per extraction using the RNAqueous kit by Ambion. Two total RNA extractions were done for each line. After each extractions we performed reverse transcription using a (dT)$_{15}$ oligonucleotide, in order to obtain cDNA for every transcript with a polyA tail. Levels of
bcd and nos transcripts were assayed using a Taqman real time PCR 5700. Primers for the amplification of the transcripts were designed to span introns so that distinction of genomic DNA to cDNA was apparent in agarose gels. In addition to assaying the cDNA levels of bcd and nos we also assayed the cDNA levels of the ubiquitous transcript rp49 and the caudal transcript as control of mRNA extraction and reverse transcription efficiency (rp49) or as control for the levels of total RNA within the ovaries (caudal) since it has been reported that there is segregating natural variation in ovariole number in Drosophila (Wayne and Mackay 1998), which would also reflect in the levels of maternal transcripts. The real time PCR reactions were performed for each RNA extraction separately, in different PCR plates to account for variation in the performance of the ABI 5700 system. Levels of PCR product were measured by the detection of CybrGreen bound to double-stranded DNA. We assigned a threshold line that crossed the linear phase of the exponential increase of the PCR product and obtained the cycle number that corresponds to the point the threshold crosses the linear phase. This is denoted as the C_t value. The threshold was assigned at the same CybrGreen detection level in each of the plates scored.

Markers scored in 3rd chromosome replacement lines: We scored polymorphisms in the promoters of bicoid and nanos in all 56 lines scored for bcd and nos expression levels. Polymorphisms were scored by obtaining a single stranded read of the sequence and by calling the state of most of the polymorphic sites detected in the survey among the 14 lines that were sequenced reliably from both strands (we avoided singletons). Six polymorphic sites were scored for the bcd promoter and eleven for the nos promoter.
Statistical tests for gene expression: Real time PCR C\textsubscript{T} measurements (cycle at which the pre-assigned horizontal line crosses the linear phase of the PCR product increase) were log-transformed to approximate a normal distribution. The logarithm of the measurements for \textit{bcd} and \textit{nos} cDNA levels was regressed to the logarithm of the measurements for \textit{rp49} and \textit{caudal} and the residuals from this regression were used as measurements to associate with nucleotide variation. Analysis of variance, non-parametric tests, such as Mann-Whitney, and regression analysis were used to test for association of individual markers with expression levels in \textit{bcd} and \textit{nos}. In addition, haplotype tests and cladistic analyses were done to test for the association of a set of polymorphic sites (haplotypes) rather than a single site with levels of \textit{bcd} and \textit{nos} expression (Templeton et al. 1987).
1.4 Results.

Pattern of nucleotide polymorphism and divergence: Nucleotide polymorphism and divergence was generally low in most of the regulatory regions surveyed with the exception of nos, where the polymorphism was higher and there are many intermediate frequency polymorphic sites (Table 1). Low nucleotide variation is in accordance with the pattern previously observed in eve stripe-2 enhancer and the hunchback proximal promoter (Ludwig and Kreitman 1995; Tautz and Nigro 1998). Calculation of the normalized differences between heterozygosity estimated average pairwise differences (\(d_{T}\)) and heterozygosity from the number of segregating sites (\(d_{S}\)), \(D\) (Tajima 1989) (Table 1), show clearly that there is a tendency for excess of rare polymorphisms, which could either be due to selective sweeps or due to strong selective constraints. When we applied Fisher’s combined probability (Sokal and Rohlf 1997) for Tajima’s \(D\) statistic \(p\)-values (Tajima 1989) from all 5 regions surveyed (obtained from coalescent simulations) the result was highly significant \((P = 0.002)\), suggesting a universal strong tendency for an excess of rare polymorphisms (mostly singletons) within regulatory regions. This observation combined with the fact that divergence in these five regulatory regions \((d = 0.064 \pm 0.048)\) is lower than the average \(D. melanogaster-D. simulans\) divergence \((d = 0.093 \pm 0.039,\) Moriyama and Powell 1997) supports the notion that these sequences are highly constrained. Linkage disequilibrium patterns are shown in Figure 1. Generally there is little linkage disequilibrium with the exception of nanos, which will be discussed separately below. The pattern of low variation is probably due to strong purifying
selection to maintain the correct structure of regulatory elements within these regulatory sequences.

Comparison of the distribution of divergence and polymorphism along the sequence may provide some clue about the difference, if any, in the selective constraints within and between species. We applied the test described in Dermitzakis and Clark (2001), initially designed to compare the patterns of amino acid substitutions between duplicated genes. Under neutrality the expectation is that the two patterns will be random and not significantly different from each other.

We applied the test in all five genes surveyed here as well as the eve stripe-2 enhancer and the proximal hunchback promoter. The initial result is that in no case were the patterns of substitution and polymorphism significantly different. We noticed however that the $p$-values generated from our random permutations were very high sometimes close to 1. We decided therefore to use the other tail of the permuted null distribution and consider the opposite alternative hypothesis that the polymorphic sites are significantly close to the interspecific substitutions. Using this alternative hypothesis, two of the 7 genes (eve stripe-2 and engrailed) showed significant proximity of substitutions and polymorphisms. When we applied Fisher’s combined probability (Sokal and Rohlf 1997) for all seven genes the result was significant with $P = 0.0296$. Note that from the other five genes that did not have significant clustering three were close to a significant clustering (bcd, ftz, proximal hunchback) and two were essentially random (distal hunchback and nanos). The significance of the clustering of polymorphic and divergent sites can be explained by the fact that these sequences contain many functional regulatory elements. If the vast majority of the sequence is constrained due to functional
importance, then the remainder will be only a small fraction of the sequence and that is where there is freedom for the accumulation of mutations and substitutions. When this pattern is compared to a completely random pattern assuming no constraint (as our permutations assume) it will show significant clustering of the two types of nucleotide changes (Figure 2).

Polymorphism in experimentally verified binding sites: The argument that there is strong purifying selection within these regions is also supported by the fact that there are very few polymorphic sites within experimentally verified transcription factor binding sites. Out of a total of 24 binding sites 5 are polymorphic 4 of which have insertion/deletion polymorphisms (Figure 3).

Variation in predicted binding sites: Previous analysis has suggested that besides the experimentally verified binding sites, which are shown to be essential for the regulation of the gene, there may be other weak binding sites that serve as a backup or as secondary support to the essential ones. We attempted to quantify variation in the number and location of such potential binding sites. We used binding site prediction models described in Chapter 2 to determine whether there was any variation in the amount and location of potential transcription factor binding sites for factors already known to regulate the gene of interest. Our analysis revealed that the pattern is generally the same among alleles, but there is slight variation in the number and location of predicted binding sites. Whether this variation contributes to actual variation in the expression levels of the gene remains to be tested. We believe however that such difference in the pattern of potential binding
sites may contribute to the evolutionary potential of the region, allowing evolutionary transitions within these regulatory regions such as the one observed between *D. melanogaster* and *D. pseudoobscura* (Ludwig et al. 2000).

**Nucleotide variation and divergence in the nanos promoter:** Among the five regions surveyed in this study and the total of seven available, nanos appears to be different in terms of the pattern and level of nucleotide variation. It has much higher _ and ,, values than the other 6 regions (Table 1) and a strong pattern of linkage disequilibrium (Figure 1). There are 9 intermediate frequency polymorphic sites in almost absolute linkage disequilibrium. We wanted to test whether these two haplotypes defined by these 5 sites constitute ancestral segregating polymorphism or one of them is recently derived and has increased in frequency. First of all, one of the two haplotypes is defined by all the ancestral states in the 9 sites mentioned above as this is define by the African allele and the *D. simulans* homologous sequence. We split the alleles in two groups, each defined by the two haplotypes, and calculated _ , ,, and Tajima s D statistic (Tajima 1989). If both haplotypes were old we expect no difference in those metrics between the two groups. Contrary to this expectation we observed that the group defined by the ancestral haplotype has high _ and ,, values (_ = 0.0258, _ = 0.02295), large number of segregating sites (s = 27) and almost neutral Tajima s D (D = -0.596). The other group defined by the derived haplotype has small number of segregating sites (s = 9), all polymorphisms are singletons, and has low _ and ,, values (_ = 0.00513, _ = 0.0085) and significant negative Tajima s D (-1.7232, P<0.05). This pattern is consistent, although not direct evidence, with the derived haplotype recently and rapidly having increased in frequency in the
population. It remains to be seen whether this pattern is present in other populations (e.g., European) and whether it has phenotypic consequences. When we applied the Fay and Wu test (2000) the result was not significant, probably because the frequency of the derived allele has to be very high for the test to be significant. From our initial results on nos expression there seems to be no association of this haplotype structure with differential expression levels (see below). It is possible that the functional polymorphism selected is in the coding region and has caused a whole set of polymorphisms to sweep to high frequency. It is interesting that nanos is involved in germ cell development (Verrotti and Wharton 2001) so it is possible that it undergoes similar selective pressures to reproductive genes which generally show accelerated evolution. Comparison of the coding regions of the same alleles as well as the D. simulans sequence may clarify the reasons for the unusual pattern we observe.

**Variation at cDNA levels of bcd and nos:** In order to test whether nucleotide variation in the regulatory regions we surveyed may be affecting gene expression levels, we quantified mRNA levels of two of the genes that showed the highest level of nucleotide variation, bcd and nos. Total mRNA was reverse transcribed to cDNA and cDNA levels of four genes, bcd, nos, rp49 and caudal were measured in a well controlled design. For the association with nucleotide variation all Ct values (see methods) were log-transformed to approximate the normal distribution, and the residuals of the regression of the log-transformed bcd or nos with log-transformed rp49 and cad were obtained and they were used as phenotypic measurements. The effects of both rp49 and cad were significant and that is why they were regressed from the bcd and nos measurements and
the residuals from these regressions were used as the phenotypic measurement. Significant variation in the residuals among lines was detected for *bicoid* ($F = 1.81, P = 0.015$) but not for *nanos* ($F = 0.64, P = 0.952$) (Figure 4).

Several statistical methods were used to associate nucleotide variation in *bcd* and *nos* with cDNA levels (see methods). None of the methods showed any significant or close to significant association of cDNA levels (as expressed by the residuals described above) with nucleotide variation. This can be a result of a combination of reasons. Levels of variation were generally low and the number of lines was small enough that we could not probably detect nucleotide polymorphisms with small effect. In addition, *nanos* did not have significant differences among lines and therefore variation in cDNA levels was not high enough to detect even larger effects. What remains to be calculated is the power of detection of genetic effects in gene expression in these two genes, given the sample size, the levels of cDNA and nucleotide variation and the measurement error with the real time PCR.
1.5 Discussion.

In this study we looked at levels of nucleotide and mRNA variation in order to understand the evolutionary processes that regulatory regions undergo within and between species. We looked at genes involved in the early development of *Drosophila*, which have generally well characterized expression and functional profiles. We believe that this system is a good representative of regulatory networks and will provide good insights for regulatory evolution.

The first significant observation is that regulatory regions have generally low levels of nucleotide variation and divergence. This observation is consistent with the previous studies reporting low variation in well-characterized regulatory regions (Ludwig and Kreitman 1995; Tautz and Nigro 1998). Therefore such regulatory regions undergo strong selective constraints to maintain the structure of regulatory elements. However, patterns of sequence rearrangements between distant Drosophila species have shown that regulatory evolution is possible sometimes with radical changes (Ludwig et al. 2000).

Why aren’t these changes present within species? It is likely that there is underlying structure within the regulatory regions as we have demonstrated in a previous study. There are sequences with potential for regulatory function, which are not being used in the current organization of the regulatory region. In cases where extreme mutations occur that cause a major sequence rearrangement such potential elements may be brought to light and become readily available. This new structure may be equally or more efficient than the previous one for the appropriate regulation of the gene and thus
could be maintained in the population. Such a process may be threshold dependent, in other words it may require many changes are required for a new pattern to be functional. Many of these changes may be neutral until they are recruited when a significant amount of them have occurred. The requirement for many changes could be the reason why major rearrangements occur in distant species where there is enough time of separation for the accumulation of many changes.

Our efforts to associate nucleotide variation with levels of gene expression were not successful. As we discussed earlier this could have been because of small sample size, low level of variation or high enough measurement error that would mask the association. It is also possible that there is no variation in gene expression that could be associated with nucleotide variation in the promoter regions. This scenario is consistent with the generally low levels of nucleotide polymorphism and divergence in regulatory regions. In addition, the phenotypes that these genes control are crucial for the correct development of the fly and therefore any distortion of the expression pattern may be detrimental for the developing embryo. Such mutations will not be maintained in the population and therefore we may never see them.

One of the regulatory regions however showed a different pattern of nucleotide polymorphism, which is worth considering separately. The *nanos* promoter had a large abundance of intermediate frequency polymorphic sites relative to the other regions, which were in strong linkage disequilibrium and one of the two distinct haplotypes was completely shared with an African line, which is generally considered ancestral (Begun and Aquadro 1993) and the *D. simulans* homologous sequence. In addition, the ancestral group had much larger level of variation than the derived group. This pattern is indicative
of one of the derived haplotype increasing in frequency recently and rapidly due to a potential selective advantage. There is not much known about the functional elements within the nanos promoter so we cannot say whether these polymorphic sites fall within functional sequences and whether this selective advantage could be driven by the regulatory region or by adjacent regions such as the coding sequence. It remains therefore to be tested whether we are looking at a step of regulatory evolution or at promoter haplotype structure driven by positive selection acting on another region.

An interesting observation that also suggests strong selective constraint within regulatory regions is that patterns of interspecific substitutions and polymorphisms significantly coincide. As we explained before if only a small fraction of the regulatory sequence is free to accumulate mutations then both intra- and inter-specific differences will tend to accumulate in these small regions and therefore there will create a pattern of significant proximity of these two types of nucleotide changes. We believe that such methods like the one we used to identify this significant pattern (Dermitzakis and Clark 2001) could be helpful in annotating regulatory regions without the need of very distant species comparisons. Moreover it may be able to identify sequences that are functionally conserved across small distances (D. melanogaster-D. simulans) but not conserved between more distant species (D. melanogaster-D. pseudoobscura). If it is true what has been suggested that regulatory regions evolve with big transitions rather than steady evolutionary rate then conservation between distant species will not be informative for the identification of the landscape of functional elements. Individual functional studies will be necessary to verify function and many regulatory elements will be species-specific (Chapter 3).
Regulatory evolution is far from being well understood and despite the efforts that are being done there is still much to learn about the pattern and the rules under which such sequences evolve. It is immediately apparent that more complicated models than the ones used in coding sequences have to be adopted. In this study we proposed methods for the study and interpretation of the pattern of polymorphism and divergence. We presented data suggesting that regulatory sequences be under strong selective constraints and concluded that evolution of regulatory sequences, at least in the early development genes in *Drosophila*, is occurring through large and rare transitions rather than a steady state sequence change.
1.6 References.

Drosophila melanogaster are very different at the DNA level. Nature 7: 548-550.


*J. Mol. Biol.* 193: 723-750.

Bergman, C. M. & Kreitman, M. 2001. Analysis of conserved non-coding DNA in
Drosophila reveals similar structural and evolutionary properties of intergenic and

Berleth, T., Burri, M., Thoma, G., Bopp, D., Richstein, S., Frigerio, G., Noll, M. &
Nusslein-Volhard, C. 1988. The role of localization of bicoid RNA in organizing


by Bicoid provides a mechanism for threshold-dependent gene activation in the


Table 1.1: Summary of the polymorphism data in the regulatory regions surveyed

<table>
<thead>
<tr>
<th>Gene region</th>
<th>length</th>
<th># polym. sites</th>
<th>_</th>
<th>&quot;</th>
<th>Tajima s D</th>
<th>p-value *</th>
</tr>
</thead>
<tbody>
<tr>
<td>bcd</td>
<td>926</td>
<td>22</td>
<td>0.00936</td>
<td>0.00540</td>
<td>-1.7822</td>
<td>0.018</td>
</tr>
<tr>
<td>en</td>
<td>1039</td>
<td>9</td>
<td>0.00354</td>
<td>0.00223</td>
<td>-1.50018</td>
<td>0.053</td>
</tr>
<tr>
<td>ftz</td>
<td>550</td>
<td>6</td>
<td>0.00377</td>
<td>0.00255</td>
<td>-1.19518</td>
<td>0.109</td>
</tr>
<tr>
<td>hb distal</td>
<td>960</td>
<td>14</td>
<td>0.00491</td>
<td>0.00280</td>
<td>-1.67573</td>
<td>0.041</td>
</tr>
<tr>
<td>nanos</td>
<td>480</td>
<td>36</td>
<td>0.02412</td>
<td>0.01746</td>
<td>-1.6620</td>
<td>0.077</td>
</tr>
</tbody>
</table>

*Calculated by coalescent simulations*
Figure 1.1: The figure shows the regions sequenced, the number of segregating sites (vertical lines), and each square indicates the result of Fisher’s exact test for linkage disequilibrium. Gray squares are significant associations with Fisher’s exact test and black squares are significant associations with Bonferoni correction.
**Figure 1.2**: model to explain the significant proximity of polymorphic sites and fixed differences. Gray boxes indicate hypothetical regulatory elements (functionally constrained). Arrows indicate polymorphisms (below) and divergent sites (above).
en (EN, EVE, PRD etc)
TTAAATGTCAATTAAATATCAATCAATT
..........................A

en (EVE, ZEN, OCT2)
CATTTA.AAT
......T....

ftz (TTK)
ACATTTC.GCCG
.......T....

hb distal (HB)
TTTTTTTTTTT-AGTCC
.........T.....

hb distal (HB)
TTGGGAAAAAAAA-TTGG
..........A....

**Figure 1.3:** Polymorphic sites within experimentally verified binding sites.
Figure 1.4: Box plots representing variation in the residuals of the logarithm of *bcd* and *nos* Ct values from two independent measurements, after regressing out the effect of the logarithm of the two controls (*rp49* and *cad*).
CHAPTER 2: High evolutionary potential in regulatory sequences of early development genes in *Drosophila*
2.1 Abstract.

The evolutionary tuning of the regulation of gene expression is a primary means by which a gene's function adapts to the needs of the organism. Regulatory regions usually consist of a set of short sequences where transcription factors bind and regulate the dynamics of transcription complex assembly and subsequent expression of nearby genes. In this study we use alignments of known transcription factor binding sites to construct a statistical model for site prediction, and we apply it to promoter regions of related species to characterize the sequence divergence in these binding sites. We focus on a well-characterized system and use sequence data from five regulatory regions of five *Drosophila* species early development genes for which functional binding sites have been identified in *Drosophila melanogaster*. Although functional elements generally exhibit conservation of sequence, functional elements are sometimes gained by sequence changes to neighboring sites that previously were too divergent from a binding site to be functional. We introduce a method for the prediction of transcription factor binding sites using recurrent sampling of weight matrices that is independent of the background sequence. The approach allows inference of confidence in transcription factor binding site prediction avoiding the influence of potential nucleotide biases in the background sequence, which could reduce the power of prediction. The integration of bioinformatics tools into the study of evolutionary processes provides a powerful approach to quantify evolutionary processes in regions such as gene regulatory regions that evolve under less well-understood constraints.
2.2 Introduction.

Theoretical studies have shown that the patterning of the Drosophila embryo is fluid and auto-regulatory (Gibson 1996), so that some perturbations can easily be corrected at later developmental stages (Namba et al. 1997). The organization is precise and progresses from the control of maternal morphogens such as Bicoid (BCD), Hunchback (HB), Nanos (NOS), Pumilio (PUM), and Caudal (CAD) to a finer network that localizes the expression of downstream genes in stripes with very well defined boundaries (Rivera-Pomar and Jackle 1996). It is the relative concentration of the initial morphogens in each segment of the Drosophila embryo that determines which set of genes will be expressed, and this combined expression defines the fate of the specific group of cells (Berleth et al. 1988; Frohnofer et al. 1986; Struhl 1989; Tautz 1988). These genes include even-skipped, fushi tarazu, knirps, engrailed, and the zygotic transcript of hunchback. One of the most important properties of the system is how, where and when these genes are expressed (Tautz 1988; Rivera-Pomar and Jackle 1996).

Although gene regulation has been studied extensively in this system (Driever and Nusslein-Volhard 1989; Burz et al. 1998; Tautz 1988), little is known about the way these regulatory sequences evolve (but see Ludwig et al. 1998). As genomic data are accumulating and the need to discover gene function is increasing, it is becoming a necessity to understand the functional elements of a gene's regulatory sequences, the level of nucleotide variation that these sequences harbor and how these sequences differ between species. One of the biggest challenges will be to determine how regulatory sequence variation contributes to phenotypic variation and how much of this variation segregates in natural populations (Gibson and Hogness 1996).
Regulatory sequences have properties that make their study more difficult than that of coding sequences (Bergman and Kreitman 2001; Leung et al. 2000; Wasserman et al. 2000; see also Chapter 3). We do not know much about the rules that define the position, amount and strength of functional regulatory elements. There are no straightforward properties in regulatory sequences analogous to the open reading frame and codons in coding sequences, and although studies have shown that the spatial information between regulatory elements is important, it is not clear what attributes of this spatial distribution are essential (Bergman and Kreitman 2001). In addition, the model of cis-regulation is not a simple one of activation or suppression by transcription factors but includes competitive binding of proteins (Small et al. 1991), co-operative binding (Zhao et al. 2000; Burz et al. 1998), chromatin bending and other molecular interactions that are not always reflected in the nucleotide sequence. As a first approximation these interactions are not taken into account and a model closer to an additive contribution of transcription factors has been adopted in most studies.

Several methods, such as phylogenetic footprinting, have been developed that make use of the conservation of regulatory sequences to infer potential function (reviewed in Hardison 2000). Phylogenetic footprinting is considered a powerful method and in some cases it has proven useful in the identification of the fine structure of some regulatory regions (Wasserman et al. 2000; Hardison et al. 1997; Leung et al. 2000). However, it may be problematic to assume that all functional sequences are conserved and all non-functional sequences have diverged. We have already shown that sequence comparisons of closely related species produces many false positives and comparison of distantly related species has low power to detect functional elements in the species
compared, due to species-specific loss and gain of transcription factor binding sites
(Chapter 3).

Another useful approach is to use motif extraction methods and information from verified functional elements to derive useful models that predict functional patterns. Some of these models make use of probability weight matrices (Yada et al. 1998; Eddy 1998; Durbin et al. 1998). Weight matrices are built using the common sequence features of the binding sites of a given transcription factor. Such methods appear to be useful for the quantification of divergence of sequence of binding sites. One can use the
probabilities produced by the model to estimate whether a substitution within the binding site has a significant effect in the binding ability of the transcription factor. It is generally assumed that weight matrices express the real properties of the binding sites of a transcription factor. In some cases it has been shown that the probabilities produced by the matrix correlate with the binding affinity of the transcription factor to the binding sites under study (Berg and von Hippel 1987). Most methods for binding site prediction consider the background sequence as null and attempt to identify a short sequence that has an outstanding match to the weight matrix (Durbin et al. 1998; Eddy 1998). One of the main problems with such an approach is that the criteria for distinguishing a binding site from the background sequence are not obvious and may be misleading. For example, if the model uses nucleotide composition to generate the null hypothesis and both the weight matrix and the sequence under study have the same nucleotide composition bias (e.g. it is A-T rich) it will be difficult to obtain a significant match to the weight matrix since such a motif will have a high probability to occur by chance in the given sequence. Comparison of short sequences with the background sequence is not biologically
meaningful because binding does not depend on how different a binding site is from the background sequence, and the same sequence segment can be functional in many different sequence environments. Therefore, under certain conditions weighing significance by the properties of the background sequence will not be powerful enough to predict functional binding sites.

In this report we describe quantitatively the evolutionary changes that occur in the regulatory sequences of segmentation genes in five *Drosophila* species. We focus on regulatory regions of five well-studied early development genes in *Drosophila*, the *even-skipped* stripe-2 enhancer, the proximal promoter of *hunchback*, the enhancer of *knirps*, the zebra element of *fushi tarazu* and the proximal promoter of *engrailed*. We initially present a qualitative description of the pattern of divergence in the binding site sequences and the background sequence. Then we introduce an approach for binding site prediction that does not take into account the background sequence, using probability weight matrices. This approach appears to be more biologically meaningful than several methods described until now and it is more flexible for evolutionary analysis. Finally, we use the above approach to obtain quantitative measures of divergence of the above regulatory regions and describe the evolutionary potential of these regions. From this we draw conclusions about the mode of evolution of the early development regulatory network in *Drosophila*.
2.3 Results.

Pattern of divergence in regulatory regions

We have used data from 5 species of *Drosophila* to measure divergence within 5 regulatory regions (Figure 1). Divergence in these 5 regulatory regions is low (Table 2) compared to the synonymous divergence rates described in previous studies (Bunn et al. 2001). This suggests that purifying selection is likely to play a significant role in the maintenance of function of the regulatory region. Several indels are present in the alignments. In addition, slippage events are observed and extension (or retraction) of stretches of nucleotides repeats is common.

Divergence within binding sites

Many binding sites have been identified in these regions in *Drosophila melanogaster*, some of which contain substitutions or indels in the other species (see Figure 2 for binding site alignments). The distribution of divergence values of these binding sites in other species from the *D. melanogaster* sequence is shown in Figure 3. There are also several indels within binding sites. In some cases the mutational events are so radical that the binding sites are most likely not functional in other species. As previously observed by Ludwig and colleagues (1998) a BCD site (BCD_3) in the eve stripe 2 enhancer has so many substitutions that it is predicted not to be recognized by BCD in *D. yakuba* and *D. erecta*. This is potentially a gain of a BCD site in the melanogaster subgroup. In addition, one of the HB sites (HB_1) appears to be deleted in *D. erecta* and since it is present in *D. yakuba* it is probably a loss of this site in the *D. erecta* lineage. Finally, one of the CAD sites in the *ftz* zebra element has obtained a two-
base insertion in *D. simulans* and it is possible this site is not as functional as the sites in other species, which is a probable loss of an essential site (ref: nature ftz paper) in the *D. simulans* lineage.

Evolutionary events like this show that gain and loss of sites is happening. However, the general trend within this range of divergence is towards conservation of function across species. In more distant species some major sequence rearrangements have occurred that are either compensatory in terms of function (Ludwig et al. 2000) or lead to new patterns of expression and regulation (ref). We are attempting to trace hidden structure in these sequences that facilitates these functional evolutionary transitions (evolutionary potential).

**Evaluation of the significance of predicted binding sites**

Our approach for binding site prediction is to initially assume that a short sequence is binding site and test whether the \( L \) value (see methods) it obtains falls within the 95% confidence interval of known functional sites (Figure 4) of the same transcription factor. The rationale behind the null hypothesis of this approach is:

1. We wanted to avoid using the background sequence as a standard for reasons mentioned in the Introduction so we had to rely just on the properties of known functional sites of the factors we are studying (BCD and HB).

2. There is not much known about the properties of short sequences that are not functional binding sites of a given transcription factor. Therefore it is not possible to construct models that are independent of the background sequence and have as null hypothesis that a short sequence is not a binding site.
3. Some transcription factors are stricter in their sequence requirements for binding than others and a null hypothesis that a short sequence is a binding site can more efficiently accommodate the special properties of each transcription factor. Even if we had good information on what the features of non-binding sites of a given transcription factor are, as sequence requirements for binding become more flexible the properties of short sequences that are not functional binding sites become unclear.

4. We are interested in strong as well as weak binding sites and our model is more likely to identify such sites since they are generally less pronounced in terms of how well they fit a consensus, but are still within the range of the binding requirements of the transcription factor of interest.

We chose to generate models for BCD and HB because these proteins play an important role in the regulation of the genes examined here. Many experimentally verified binding sites of these two transcription factors have been identified in their regulatory regions. As a first step, we wanted to obtain an estimate of the power of our method to predict transcription factor binding sites. We approached this by performing a jackknife, each time excluding one of the BCD or HB sites, generating the matrix and the thresholds from the remaining sites and testing whether the excluded site was predicted as functional. This method predicted as functional 92.7% of the observed BCD sites (47 out of 51) and 94.84% of the observed HB sites (92 out of 97).

These power calculations are not typical because our test is not designed to reject the hypothesis that a sequence is not a site, but rather it is designed to reject the null hypothesis that a sequence is a binding site (see Methods for justification). In essence, the above calculations are calculations of type-I error, the probability that we will reject the
hypothesis that a site is functional when in fact it is. The reassuring result is that these calculations produce values very close to 95%, which is what we expect given the way the test is performed.

Another issue is the frequency of false positives our prediction method can generate. This is the real value of the test since it will show how many times we include non-functional site in the functional range, in other words how frequently we fail to reject that a site is functional when it is not functional. This is a difficult issue to address because we do not have any data on what is the structure of a non-functional BCD or HB binding site. There have been no systematic experiments showing what are the types of sequences that BCD or HB do not bind and therefore we cannot generate models assuming non functionality.

To obtain an idea of how high our false positive rate is, we compared our results to those obtained using a binding site prediction method that tests the null hypothesis that a sequence is not a binding site. We used the method developed by Yada et al. (1997) and is available under the web based software YEBIS (http://www-scc.jst.go.jp:8080/sankichi/). We use the D. melanogaster sequences and the BCD and HB weight matrices obtained from our binding site data. Two thresholds (LOD score) were used for the binding site prediction with YEBIS. When the cut-off LOD score was set to 3, YEBIS predicted 26% less BCD binding sites and 35% more HB binding sites than our method but all the YEBIS sites were also predicted with our method. For a cut-off LOD score of 2, the YEBIS results were almost identical to our method with slight variations across genes for BCD but more than twice as many more HB sites were predicted relative to our method (see Table 3 for summary of the data). This demonstrates
one of the useful properties of our method. The weight matrix for HB is more degenerate than that of BCD and we see that a conventional method predicts many more HB sites than our method even with a strict threshold indicating the increased power of our method for degenerate weight matrices. Since methods such as YEBIS with null hypothesis of no function produce very similar results to our method, which has the opposite null hypothesis, it is safe to assume that the degree of false positives produced by our method is low. In addition, our method is more biologically meaningful because we take into account the distribution of $L$ values derived from functional binding and ignore the properties of the background sequence, which as we said in the introduction can mask perfectly functional sites. It is also important to point out that for very flexible weight matrices (high nucleotide degeneracy in the binding requirements of a transcription factor) the power to detect binding sites with a null hypothesis of no function will decrease while the power of our test to predict the same sites will not decrease since the weight matrix degeneracy is taken into account in obtaining the cut-off values.

**Prediction of the pattern of binding site evolution**

Using the above matrices of BCD and HB binding sites we searched the regulatory regions in the different species for putative BCD and HB binding sites. The objective of this analysis was to determine whether the nucleotide substitutions that have occurred within the experimentally verified BCD or HB binding sites have potentially eliminated or altered function, and whether nucleotide changes outside of the known binding sites provide any evidence of gain and loss of BCD and HB binding sites. In
Figure 5 we show the pattern of predicted BCD binding sites, in regions that are regulated by BCD, in all 5 *Drosophila* species studied. An important aspect of these data is that all but one of the known functional BCD binding sites (BCD_3 in eve stripe 2 enhancer in *D. yakuba* and *D. erecta*) and all but one of the HB binding sites (HB_1 in eve stripe 2 enhancer in *D. erecta*) are predicted as functional in all species. This suggests that within the range of divergence between *D. melanogaster* and *D. yakuba/D. orena/D. erecta* purifying selection was strong enough to maintain most sites that are essential for gene regulation.

On the other hand, many of the putative sites are not shared between species, which suggests that several potential binding sites with less essential role can be gained and lost independently without a strong negative effect on the regulation of the gene. In Figure 6 we present data describing the potential process of gain and loss of putative BCD and HB binding sites. It appears from this figure that the process of gain and loss is continuous and there are many such events even in short evolutionary time. Sites that turn over would also be more tolerant to mutations, since the essential ones are already present, and may even be responsible for variation in expression levels of the genes they regulate. An analysis of the pattern of polymorphism for several of these sequences and variation in gene expression levels will be presented in another study (Chapter 1). An interesting observation is that although the lineages of *D. yakuba* and *D. orena/D. erecta* appear to have diverged at about the same rate from *D. melanogaster*, the change in the L values caused by substitutions within binding sites in *D. yakuba* is higher than that in *D. orena/D. erecta*, which may indicate a unique mode of evolution of these regions in *D. yakuba*. In addition, the level of nucleotide divergence (Table 2) is generally lower than
the proportion of not shared predicted sites (Figure 6). This is a result of a high proportion of indels, which seem to provide new sequences that increase the density of putative BCD and HB binding sites.

Another aspect of binding site divergence is whether the nucleotide substitutions between species are likely to affect quantitatively the ability of the transcription factor to bind the sequence. We chose experimentally verified BCD sites that have substitutions in *D. yakuba* and *D. orena / D. erecta* relative to the *D. melanogaster* sequence. For each site, we counted the number of nucleotide changes within the binding site sequence that differentiate the *D. melanogaster* sequence from the sequence of *D. yakuba* or *D. orena/D. erecta* and then generated all the possible sequences that can be obtained by applying the same number of changes randomly to the *D. melanogaster* binding site sequence. We then scanned the mutated sites with the BCD weight matrix and obtained for each original site the distribution of $L$ values of these randomly mutated binding site sequences. Finally, we compared the values from the random distribution to the $L$ values the sites obtain in their observed substituted form in the other species. Out of 5 BCD sites with substitutions, 3 sites showed changes significantly towards the conservative (look more like a BCD site) extreme, 1 showed changes within the random range and 1 showed changes significantly towards the radical (looks less like a BCD site) extreme. This result suggests that although most of the nucleotide changes are maintaining the properties of the binding site, there are still changes that are not constrained and can potentially result in the loss of the site.
Simulation of the evolutionary potential of regulatory regions

One of the hypotheses of this study is that elements within regulatory regions undergo turnover, which results in the continuous gain and loss of functional elements, but conserves the function of the regulatory region as a whole on the expression pattern of the gene (Ludwig et al. 2000). If this is true, we expect to find traces of intermediate steps of this turnover process within these regions. These traces could either be lost binding sites due to mutational events or sites that are close to becoming a functional site. The latter is likely because the binding site sequences are short and the probability of obtaining a functional binding site by chance is relatively high given enough divergence time to accumulate mutations. It is also possible that some regulatory regions have properties such as nucleotide composition bias, repetitive sequences etc, that allow the emergence of new binding sites for some factors with high probability.

We simulated a range of divergence rates in the regulatory regions of *knirps* enhancer and *hunchback* proximal promoter and asked whether there are regions where a new BCD site, not predicted in the original *D. melanogaster* sequence, appears consistently at the same position in the simulated diverged sequences. We did 1000 simulations of divergence of the *D. melanogaster* sequences with rates 10%, 20%, 30% and 40%. We then searched the simulated diverged sequences with the weight matrix for putative BCD binding sites and obtained positions that produce an $L$ value higher than the threshold for BCD. We then counted how many out of the 1000 diverged sequences had a new BCD site predicted in the same position that was not predicted in the original *D. melanogaster* sequence.
In several cases a BCD site was consistently generated at the same position with occurrence between 10% -30% (Figure 7). This indicates that there are sequences within the regulatory regions that are close enough to becoming a functional site that a few substitutions can make then potentially functional. The original sequences (before divergence) can initially obtain substitutions and become weak sites, which may then be favorably selected if another functional BCD site is lost or if the addition of the new binding site is advantageous. If favorable selection for new binding sites operates, the rate at which some of these are fixed will be higher. We also observed that, from the known functional sites, some are more robust to substitutions than others and therefore the rate of loss of function is different for different sites. This is a result of the degeneracy of BCD binding requirements and the sequence evolutionary starting point. Another interesting observation is that at 10% divergence the occurrence of a new site at the same position was lower than at 20% divergence but then the occurrence reached a plateau at higher divergence values. This suggests that with divergence 10% there were not enough substitutions to generate the site but above 20% divergence the number of substitutions was high enough to reach the maximum probability of gaining a site.
2.4 Discussion.

Our findings suggest that there are many putative BCD and HB binding sites within regulatory regions, which may be potentially functional for the regulation of the gene. We also demonstrate that a high proportion of these potential sites are gained and lost in the evolutionary history of the Drosophila species, indicating substantial turnover of potential binding sites. In addition, nucleotide changes occurring within the known and essential functional binding sites sometimes result in radical changes in the binding potential of the sites as this is described by the magnitude of the $L$ value for BCD and HB sites or by the magnitude of the mutational events. Finally, by simulating divergence from the D. melanogaster sequence, we showed that regulatory sequences have a high potential of obtaining new functional binding sites. The above results suggest an underlying dynamic process in the evolution of regulatory sequences, where there is an evolving background of potential binding sites ready to acquire essential role when selected.

The initial impression from past studies of divergence in regulatory regions is that most of the functional elements are highly conserved within the range of divergence we considered in his study (Tautz 1998; Ludwig et al. 1998). Many studies have demonstrated high sequence conservation and conserved function of regulatory regions across distant Drosophila species (Langeland and Carroll 1993; Lukowitz et al. 1994; Bonneton et al. 1997). However, recent studies indicate that the system is more complicated and fluid, with regulatory regions having an underlying pattern of evolution not directly visible from simple sequence comparisons (Ludwig et al. 2000; see Tautz 2000 for review). One of the reasons for such a contradiction is that the rules that govern
the evolution of regulatory elements are not well known. Therefore it is difficult to
determine which changes are contributing to functional differences. One of the problems
is that we do not know exactly how selection operates on regulatory elements.
Furthermore, it is difficult to quantify the effect of a nucleotide change just with sequence
comparisons even within known functional sequences because different changes will alter
the properties of a binding site in different ways.

In this study we attempted to quantify the level of divergence by adopting a more
sophisticated approach than just nucleotide sequence comparison. We devised a binding
site prediction method that uses weight matrices generated from known binding site data
for a given transcription factor and uses the distribution of the $L$ values for the known
functional sites to obtain reliable thresholds that determine whether a tested sequence is
within the functional range of $L$ values for the transcription factor of interest. The use of
such matrices provides a tool of comparison that takes into account an aspect of function
(binding) rather than just divergence. It allows for discrimination of nucleotide changes
that affect the binding of the transcription factor from changes that do not really change
the binding affinity of the protein. This method is an extension of the comparison of rates
of synonymous vs. non-synonymous changes in coding sequences but with two important
distinctions: a) it is more precise because it is not a binary value (synonymous — non-
synonymous) but rather it is a measure of the extent of deviation as described by the
difference in $L$ values in the same sequence in two different species; b) it accounts for
compensatory changes because the same likelihood value can be obtained even if several
nucleotide changes have occurred and it may not be necessarily correlated with
nucleotide divergence.
This quantification of divergence of regulatory elements allows for more sophisticated analysis of the dynamics and potential of regulatory regions. We were able to identify putative binding sites for BCD and HB in these regions that are predicted with high enough $L$ values that they are likely functional or potentially functional. What we do not know is how essential these putative binding sites are for the regulation of the gene. It is possible that these sites are used only in a subset of the tissues in which BCD and HB regulate these genes and that is the reason why they are not identified in functional studies. They may also have little contribution to the regulation of the gene, but could serve as a reserve in cases of evolutionary transitions. Last but not least, since BCD and HB are morphogens these may be weak sites that contribute to a finer sensing of the nuclear regulatory environment by the gene so that downstream pathways are activated or suppressed in a more accurate way. The latter is an important point since these genes are segmentation genes that are usually expressed in a stripe fashion and thus a very accurate sensing of the regulatory environment is necessary to define the precise boundaries of the expression domains.

Theoretical studies have predicted that the early development system of genes in *Drosophila* is highly fluid and allows for many evolutionary changes without serious distortion of the embryonic patterning (Gibson 1996; von Dassow et al. 2001). Transgenic experiments have shown that excessive expression of *bicoid* initially distorts the anterior-posterior pattern but is later corrected by fusion and cell death (Namba et al. 1997). In addition, loss of maternal *hunchback* and *nanos* simultaneously leads to an almost normal embryo, while loss of either of the two is lethal (Hulskamp et al. 1989). Studies across more distant species have demonstrated that the anterior-posterior system
in *Drosophila* has relatively recently evolved and major regulatory changes in roles and interactions have occurred (Wolff et al. 1998). It was also surprising to discover that *bicoid* is a new gene and its role in anterior determination even newer (Stauber et al. 1999 and 2000). All the studies above demonstrate that early development in Drosophila is evolving much faster than it would be anticipated based on the importance of the system for survival.

The pattern of conservation of molecular sequence data in these regions does not initially suggest such highly fluid pattern, and the requirements of the system appear stricter than one anticipates from the available phenotypic data. This discrepancy between sequence and phenotypic data may mean that the models describing the molecular interactions are oversimplified. Our analysis shows that there is possibly underlying fluidity in the system, not readily observed, as indicated by the properties of nucleotide sequence within these regulatory regions. We have shown that there are short sequences within these regions that have a high potential of becoming BCD binding sites with many different combinations of substitutions. This could be due to dead sites that were lost not long ago and still have the signal or because the probability of emergence of a binding site is increased due to a combination of nucleotide composition bias or special mutation events, such as stretches of As or Ts that may provide good background sequence for BCD and HB sites to eventually emerge.

The main conclusion from this study is that regulatory regions do not evolve in a simplified fashion by keeping conserved the functional elements and allowing everything else to diverge. There is likely a finer organization that is not sensed by most of the experimental approaches, probably due to the fact that this organization is subtle and
contributes to the evolutionary potential rather than to the phenotype at any given time. This organization seems to allow regulatory regions to reassign roles to some regulatory elements and occasionally obtain new regulatory properties. Our results are consistent with previous studies that have shown that the evolution of regulatory regions is very dynamic (Ludwig et al. 2000). We present data that suggests that there is the potential turnover of binding sites within regulatory elements and that some of these sequences have a predisposition either by chance or due to special properties of the sequence for the \textit{de novo} emergence of new functional elements. We believe that regulatory regions have a powerful evolutionary potential that allows for either conservation of the proper regulation of the gene or acquisition of a new mode of regulation when this is advantageous. This potential is fulfilled in a slow rate and possibly with big transitions rather than a continuous evolutionary process.
2.5 Methods.

*Sequencing and data from Genbank*

Sequencing of the regulatory regions for *ftz*, *en* and *kni* was done using PCR products as templates (PCR primers and conditions available on request). After the PCR reaction, products were either ethanol-precipitated or column-purified. A fraction of the PCR product was used in sequencing reactions using the commercial kit from Beckman Coulter. After the sequencing reaction, products were ethanol precipitated according to the instructions in the manual and were run in Beckman CEQ2000 8-capillary automated sequencer. Base calling was done with the Beckman software and the calls were subsequently analyzed with the option SeqMan of Lasergene. For each region we sequenced both strands at least once and in most cases more than once. Variant sites were verified with careful inspection of the trace files, after the interspecific alignment was done, to confirm that they are real. Sequences were aligned with ClustalW and divergence was calculated using *p*-distance.

Sequences will be submitted to the Genbank. We also used data for the regulatory regions of *eve* and *hb* already available in the Genbank. Table 1 summarizes the regions and the lengths of the sequence used and Figure 1 shows the position of the regulatory sequences analyzed in this paper relative to the coding sequences.

*Derivation of ancestral sequences*

Ancestral sequences were derived using the maximum likelihood program PAML (Yang 1997) based on the approach of Yang et al. (1995). This approach is not reliable for alignment gaps so we derived the ancestral sequences of alignment gaps manually applying parsimony criteria. For this approach *D. yakuba* and *D. orena*/*D. erecta* were
considered equally distant to D. melanogaster since our divergence analysis did not resolve reliably the relationships of D. melanogaster, D. yakuba and D. orena / D. erecta.

**Binding site collection for BCD and HB**

In order to obtain a reliable weight matrix that will represent the binding properties of the transcription factor of interest we have used a large sample of binding sites collected from primary literature and personal communications. The two transcription factors of interest are Bicoid (BCD) and Hunchback (HB). We have used 51 BCD binding sites (supplementary data) and 97 HB binding sites (supplementary data). The length of the consensus BCD binding site is 8 nucleotides and the length of the consensus HB binding site is 14 nucleotides. Alignments were performed with the web-based motif extraction method YEBIS (Yada et al. 1998), which uses Hidden Markov Models to extract a motif from a given set of sequences and obtain the optimal alignment.

**Binding site prediction.**

We have devised a method for the prediction of a transcription factor binding site within a query sequence. Contrary to the usual way of setting up such a test, the null hypothesis of this test is that a short sequence is a binding site and we reject that sequences are binding sites of a given factor when they deviate from a pattern consistent with the properties of the observed set of binding sites. The rationale behind the construction of this model is explained in the Results section.

Initially, we align a set of binding sites for a given transcription factor and generate a probability weight matrix, which describes the probability of each of the four
nucleotides at each position of the binding site sequence (supplementary data) as expressed by their frequency in the respective position in the alignment of the observed functional sites. We then use this matrix to generate a value $L$ that reflects how likely it is that a given sequence with the same length as the binding site is compatible with the matrix. The value $L$ is calculated using the equation below:

$$L_j = \sum_{i=1}^{n} \log(p_i)$$

where $n$ is the length of the binding site, $j$ is the first position of window $n$ on the nucleotide query sequence and $p_i$ is the matrix frequency in matrix position $i$ of the nucleotide in position $j+i$ of the query sequence.

To obtain the critical threshold we use the $L$ values that the observed sites, used to generate the matrix, acquire. A weight matrix is itself determined by a sample of binding sites, and so the confidence in the weight matrix improves as the sampling of binding sites increases in number. To determine how sensitive the weight matrix is to the particular sample of binding sites used to generate it, we drew 1000 bootstrap samples of the binding sites (drawn with replacement), and from each sample generated a corresponding weight matrix. Because this collection of weight matrices represents the sampling variability in our confidence in a determination of $L$, the distribution of $L$ values across these 1000 matrices is taken to be the null distribution of $L$ (Figure 4). We used the bootstrap in order to take into account the heterogeneity of the sample of binding sites when setting the critical threshold. Any sequence that obtains an $L$ value below the 5% cut-off of the distribution will be considered rejected to be a binding site of the respective transcription factor, since fewer than 5% of the weight matrices accept it as such. In order to obtain the predicted sites within a query sequence we calculate the $L$
values by taking a sliding window of sequence equal to the length of the binding site and stepping one nucleotide at a time from the 5' end to the 3' end of the query sequence. We also scan the query sequence with the matrix in its reverse orientation to identify binding sites in the opposite strand.

**Simulations of divergence:** The simulations of divergence are forward simulations of accumulation of substitutions. We pre-assigned a divergence rate and obtained, for every initial sequence, a set of 1000 sequences of equal length that differed from the original sequence by a proportion of nucleotides equal to the divergence rate. For simplicity we did not incorporate any mutation bias or transition/transversion bias.

All the computer programs were written in PERL and were run in a windows environment. They are available from the corresponding author on request.
2.6 References.


Gibson, G. & Hogness, D. S. 1996. Effect of polymorphism in the Drosophila regulatory

Gibson, G. 1996. Epistasis and pleiotropy as natural properties of transcriptional

Hardison, R. C. 2000. Conserved non-coding sequences are reliable guides to regulatory

reveal novel regulatory elements: a reason to sequence the mouse genome.
*Genome Res.* **7**: 959-966.

segmentation of the Drosophila embryo in the absence of a maternal posterior


Leung, J.Y., McKenzie, F.E., Uglialoro, A.M., Flores-Villanueva, P.O., Sorkin, B.C.
2000. Identification of phylogenetic footprints in primate tumor necrosis factor-


enhancer evolution in Drosophila: rules governing conservation and change.
*Development* **125**: 949-958.


<table>
<thead>
<tr>
<th>Gene</th>
<th>Length (bps)</th>
<th>Distance from transcription initiation site</th>
<th># of binding sites</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>eve stripe 2 enhancer</td>
<td>670</td>
<td>932 bps</td>
<td>17</td>
<td>mel, sim, sech, yak, ere</td>
</tr>
<tr>
<td>knirps enhancer</td>
<td>350</td>
<td>1.6 Kb</td>
<td>11</td>
<td>mel, sim, sech, yak, ore</td>
</tr>
<tr>
<td>ftz zebra element</td>
<td>600</td>
<td>0</td>
<td>12</td>
<td>mel, sim, sech, yak, ore</td>
</tr>
<tr>
<td>en promoter</td>
<td>910</td>
<td>0</td>
<td>3</td>
<td>mel, sim, sech, yak, ore</td>
</tr>
<tr>
<td>hb promoter 2</td>
<td>348</td>
<td>0</td>
<td>6</td>
<td>mel, sech, yak, ore</td>
</tr>
<tr>
<td></td>
<td>mel</td>
<td>sim/sech</td>
<td>yak</td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>----------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>sim/sech</td>
<td>0.025 ± 0.0071</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yak</td>
<td>0.063 ± 0.0102  0.064 ± 0.00846</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ore/ere</td>
<td>0.053 ± 0.0099  0.056 ± 0.00839  0.051 ± 0.00782</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Estimated from the divergence values in the five different regulatory regions
Table 2.3: Comparison of our binding site prediction method with YEBIS

<table>
<thead>
<tr>
<th>Factor</th>
<th>Gene</th>
<th>Our method</th>
<th>YEBIS, LOD =3*</th>
<th>YEBIS, LOD =2*</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCD</td>
<td>Eve stripe 2</td>
<td>30</td>
<td>20 (20)</td>
<td>28 (28)</td>
</tr>
<tr>
<td></td>
<td>Knirps enhancer</td>
<td>16</td>
<td>12 (12)</td>
<td>16 (16)</td>
</tr>
<tr>
<td></td>
<td>Hbp2</td>
<td>14</td>
<td>12 (12)</td>
<td>15 (14)</td>
</tr>
<tr>
<td>HB</td>
<td>Eve stripe 2</td>
<td>3</td>
<td>5 (3)</td>
<td>5 (3)</td>
</tr>
<tr>
<td></td>
<td>Knirps enhancer</td>
<td>2</td>
<td>2 (2)</td>
<td>2 (2)</td>
</tr>
<tr>
<td></td>
<td>Hbp2</td>
<td>1</td>
<td>1 (1)</td>
<td>2 (1)</td>
</tr>
<tr>
<td></td>
<td>Ftz</td>
<td>3</td>
<td>4 (3)</td>
<td>4 (3)</td>
</tr>
<tr>
<td></td>
<td>En</td>
<td>5</td>
<td>7 (5)</td>
<td>9 (5)</td>
</tr>
</tbody>
</table>

* In parentheses we show the number of binding sites that were predicted by both YEBIS and our method.
Figure 2.1: Schematic of the regions analyzed (in gray — not in scale). Respective coding regions are in black.
Figure 2.2: Alignments of binding sites in all species. The order of the species is *D. melanogaster* (mel), *D. simulans* (sim), *D. sechellia* (sech), *D. yakuba* (yak) and *D. orena* (ore). For alignments of binding sites in eve stripe-2 enhancer refer to Ludwig et al 1998.
Figure 2.3: Distribution of divergence within individual binding sites in: a) between *D. melanogaster* and *D. simulans/D. sechellia*; b) between *D. melanogaster* and *D. yakuba*; c) between *D. melanogaster* and *D. orena/D. erecta*. 
Figure 2.4: The empirical distribution of $L$ values used to obtain the 95% thresholds for binding site prediction
Figure 2.5: BCD Binding site predicted with the method described in this study (only sites with $L$ values higher than the threshold are shown) in all 5 species; ID refers to the location and the $L$ values obtained only for the experimentally verified BCD sites.
Figure 2.6: Phylogenetic trees describing the most likely process of gain and loss of putative/potential BCD and HB binding sites in the sequences studied. The number in the circle indicates the number of putative binding sites predicted at the ancestral sequence of a given node. The number in parentheses next to the species indicates the number of binding sites predicted in the current state of the sequence of the species. Numbers on the branches indicate the number of putative sites gained and lost on a specific branch.
**Figure 2.6:** (continued)
Figure 2.7: Counts of predicted sites at each position in 1000 simulated divergent sequences for *knirps* and *hunchback* with rates 10% and 20%. Filled bars are BCD sites that occur in more than 100 of the 1000 simulated sequences and are not predicted at the same position in the original sequence (new sites). Stripped bars are experimentally verified BCD sites in the original *D. melanogaster* sequence.
CHAPTER 3: Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover
3.1 Abstract.

Comparisons between human and rodent DNA sequences are widely used for the identification of regulatory regions (phylogenetic footprinting), and the importance of such inter-genomic comparisons for promoter annotation is expanding. The efficacy of such comparisons for identification of functional regulatory elements hinges on the evolutionary dynamics of promoter sequences. While it is widely appreciated that conservation of sequence motifs may provide a suggestion of function, it is not known what proportion of the functional binding sites in humans is conserved in distant species. In this report we present an analysis of the evolutionary dynamics of transcription factor binding sites whose function had been experimentally verified in promoters of 51 human genes, and compare their sequence to homologous sequences in other primate species and rodents. Our results show that there is extensive divergence within the nucleotide sequence of transcription factor binding sites. Using direct experimental data from functional studies in both human and rodents for 20 of the regulatory regions, we estimate that 32%-40% of the human functional sites are not functional in rodents. This is evidence that there is widespread turnover of transcription factor binding sites. These results have important implications for the efficacy of phylogenetic footprinting and the interpretation of the pattern of evolution in regulatory sequences.
3.2 Introduction.

Although regulatory regions are not under the same constraints as coding sequence, alignments of regulatory regions of human and rodent genes often reveal blocks of highly conserved sequence (Hardison et al. 1997; Wasserman et al. 2000; Jareborg et al. 1999; Leung et al. 2000). Observation of such strong sequence conservation suggests conserved function, thereby generating testable hypotheses that have often been confirmed (Wasserman et al. 2000; Leung et al 2000). However, studies in Drosophila have revealed compensatory changes in gene enhancers (Ludwig et al. 2000), illustrating that conservation of function can be maintained in the face of fluidity in the exact composition of regulatory regions. Individual binding sites may exhibit relatively little conservation, either due the degeneracy of the transcription factor binding requirements, or because their small size makes it relatively likely that a new functional site will arise by chance (Ludwig et al. 2000; Florea et al. 2000). A new site may relax the selective constraint acting on another already present site, allowing for transcription factor binding site turnover. Nucleotide variation in regulatory regions is considered an important component for disease risk (Collins et al. 1997; Risch and Merikangas 1996) because variation in binding sites may alter gene expression level and likely contribute to variation in human disease risk (McDermott et al. 1998; Werth et al. 2000; Wei and Hemmings 2000; Picketts et al. 1994). Understanding the evolutionary processes that binding sites undergo would prove valuable for inference of potential phenotypic effects and for interpretation of likely function from human-rodent sequence comparisons.
Knowledge of the distribution of divergence within functional binding sites will provide useful information for the calibration of phylogenetic footprinting methods.

In the present study we analyzed the evolution of human functional binding site sequence in 51 regulatory regions by contrasting the sequences with those of non-human primates and rodents. The sequence analysis is rooted by the direct experimental confirmation that the sites under study are functional sequences in the human promoters. For a subset of 20 of the regulatory regions we obtained comparative functional data from the primary literature for both human and rodents. By comparing regulatory regions from a series of species across a range of divergence times from humans, we capture binding sites at varying degrees of sequence divergence. Based on the functional information this analysis suggests attributes of the manner in which regulatory regions undergo evolutionary turnover.

3.3 Materials and Methods.

**Sequence data:** Human genes were selected for analysis based on the completeness of experimental assessment of identification of functional binding sites in promoter regions (see below). Sequence data were obtained from the NCBI Genbank. We used a combination of keyword and BLAST searches to identify the homologous sequences in nonhuman primate species and rodents. Some of the rodent sequences were also retrieved from the MGI database (www.informatics.jax.org). Summary of the relevant data is presented in Table 1. Species are indicated with the common or genus name. For the analysis, species within the Old World monkey lineage were pooled together, and species
from within the New World monkey lineage were separately pooled. Divergence was calculated based on the consensus sequence of the lineage. Special attention was paid to the confirmation that the sequences compared were homologous, especially for the human-rodent comparisons. A combination of BLAST searches with the coding sequence of the genes and gene annotation available in the NCBI Genbank and MGI for human and mouse was used to verify homology. The Genbank accession numbers are provided as supplementary material (see supplementary data).

Alignments: The primate sequences were aligned with ClustalW and by manual inspection. The divergence among primates was low (< 10%), making confidence in the alignments high. For the alignments of human and rodent sequences we used the web-based software PipMaker to obtain significant local alignments (Schwartz et al. 2000). PipMaker alignments were subsequently manually optimized to obtain the best possible alignment for the binding site sequences. In addition, we used the Bayes Aligner (BA) developed by Zhu et al. (1998) to compare with some of the PipMaker alignments within the binding site sequences. Alignments with BA produced essentially the same result. In the rare cases that the alignment was not the same PipMaker alignments were uniformly better (lower divergence). Therefore we used the manually optimized PipMaker alignments for our analysis.

Human functional transcription factor binding sites: The transcription factor binding sites, used in the analysis, were selected on the basis of direct experimental confirmation of binding ability (footprinting, gel shift assays) and function (promoter deletion experiments, directed mutagenesis, expression of reporter genes) in previous studies. We
identified the location of these binding sites in the human sequence by searching the primary literature and the TRANSFAC database (Wingender et al. 2000). References used for the identification of the binding sites are available in supplementary data. Divergence of binding site sequences for all the human-rodent analysis was done including alignment gaps since we are interested in how different the sequences are in the species compared and not how the substitutions occurred.

**Comparative functional analysis for human and rodents:** Data were collected from the primary literature. We restricted the analysis to studies that tested the function and binding ability of binding sites with the same criteria and methods. The criteria for the validity of the function of transcription factor binding sites were as strict as for the human collection of binding sites. From 20 genes we collected data on 64 binding sites that align between human and rodent, 33 of which share function between human and rodents, 14 that are functional in humans only (human-specific) and 17 that are rodent-specific. References and Genbank accession numbers of the regulatory region sequences are available in supplementary data.

**3.4 Results.**

We analyzed 51 gene regulatory regions in which sequence data are available for human and at least one other primate species or rodent. We used a set of binding sites in these 51 human gene regulatory regions that had strong experimental evidence for a functional role, derived from footprinting, gel-shift assays accompanied by at least one
other functional confirmation from either promoter deletion experiments, directed mutagenesis assays, or ability to drive expression in reporter genes. For each regulatory region we used interspecific sequence alignments produced by ClustalW (for primates) or PipMaker (for rodents) followed by manual optimization. Binding sites were mapped on the sequences by using reports in the primary literature or by using data available in the database TRANSFAC (Wingender et al. 2000) (see supplementary data). Summary of the genes analyzed is shown in Table 1 and Figure 1.

**Analysis of divergence within regulatory regions between human and other primates:** Nucleotide divergence within binding sites between the human sequence and the homologous sequence in other primates suggests that there is a slow process of accumulation of substitutions within binding site sequences. In particular, it appears that the divergence of binding sites between human and macaque is concentrated only in a few sites rather than being distributed homogeneously across sites (Figure 2a). We tested this hypothesis by simulating the same average level of divergence in a sample of sequences equal in length and number to the one aligned between human and macaque. We then computed the variance of divergence for each of the 1000 simulated data sets. The observed variance, when compared to the simulated distribution, fell in the right tail at \( P = 0.015 \) (Figure 2b), indicating that the substitution pattern within binding site sequences between human and macaque has significantly greater dispersion than the neutral Poisson expectation. The excess dispersion suggests heterogeneity in rates of substitution across binding sites, either due to higher flexibility of the binding properties of some of the transcription factors or due to more relaxed constraints in some binding sites.
Analysis of regulatory sequence divergence between human and rodents: Human-rodent sequence comparisons are widely used to identify regulatory elements in humans (Hardioson et al. 1997; Wasserman et al. 2000). However, it is not known what proportion of the embedded functional binding sites in human regulatory regions is conserved in rodents. This is a relevant question because non-conserved elements will not produce a strong signal of conservation and therefore will not be identified by sequence comparisons. Among comparisons of 46 regulatory regions of human-rodent homologs, 43 produced at least some significant PipMaker local alignments within the region (sry, ccr5 and myoglobin were not successfully aligned).

Average divergence of sequence in the human-rodent comparison within binding sites ($p$-distance: $d = 0.229$, SD = 0.177; Kimura 2-parameter: $d = 0.273$, SD = 0.182; Figure 3a) is lower than the average synonymous human-mouse divergence (Kimura 2-parameter: $d = 0.468$, SD = 0.169; Makalowski and Boguski 1998) but much higher than the non-synonymous human-mouse divergence (Kimura 2-parameter: $d = 0.090$, SD = 0.102; Makalowski and Boguski 1998), and the divergence of the background sequence ($p$-distance: $d = 0.310$, SD = 0.175; Kimura 2-parameter: $d = 0.399$, SD = 0.178) is very similar to the synonymous divergence. It is possible that other binding sites reside in the aligned regions and are not yet identified as functional. However, the fact that the Kimura 2-parameter estimate of divergence is not very different from the synonymous rate of substitution implies that the density of such potentially unidentified binding sites is low. Additionally, there is no correlation between amino acid sequence divergence of the genes and binding site sequence divergence ($P = 0.680$), and the amino acid divergence
in the genes compared is generally low, averaging $d = 0.269$ (SD = 0.139). Therefore the relatively high binding site divergence we observe cannot be explained by rapid overall gene divergence. In addition, there is no correlation between divergence in individual binding sites in human-rodent and human-macaque comparisons ($r = 0.001, P = 0.909$), suggesting that constraints for each site are generally independent in the two different lineages and not a property of the importance of the site for the expression of the gene. Manual inspection of expression profiles from public databases (Unigene, LocusLink, MGI, NCBI) does not suggest any major differences in expression pattern of the genes between human and rodents, but we cannot exclude the possibility that such changes have occurred. Unfortunately, data on tissue- and temporal-specific expression patterns are not unified sufficiently to allow a formal comparison of human vs. rodent expression patterns.

**Proportion of species-specific transcription factor binding sites**: In order to estimate how many binding sites exhibit species-specificity in function we need experimental data for both species. For 20 of the 43 alignable regulatory regions compared between human and rodents such data were available. A total of 64 alignable binding sites have been identified in these 20 regions, out of which 33 have shared function between human and rodents (mouse or rat), 14 are human-specific and 17 are rodent specific. First we tested whether the subset of the data for which there is functional information for both species is representative of the original sample of 43 genes (Figure 3). A Mann-Whitney test shows that there is no significant difference between the divergence values obtained from the sample of 20 genes and the divergence values from the remainder of the data ($W = 7746, P = 0.1948$). In addition, there is no difference between the divergence values of the
human-specific vs. rodent-specific binding sites (Mann-Whitney: \( W = 151, P = 0.9173 \)), so they can be pooled in one class of species-specific binding sites. There was highly a significant difference, as expected, in the divergence values in binding sites with shared function vs. the species-specific binding sites (Mann-Whitney: \( W = 628, P = 0.000 \)). Finally there was no difference between the divergence values in binding sites compared between human-mouse vs. the values in binding sites compared between human-rat (Mann-Whitney: \( W = 468, P = 0.930 \)).

Our data collection method was not biased with respect to functional conservation. Assuming that the comparative studies available in the primary literature are not biased either, we can estimate the proportion of binding sites that do not have shared function between human and rodents. An average of 15.5 sites are species specific (average of 14 human-specific and 17 rodent-specific) in a total of \( 33 + 15.5 = 48.5 \) functional sites present in each species. From this we can calculate that 32% (15.5/48.5) of the functional sites in either human or rodents are not functional in the other species. This is probably an underestimate, since observation of the primary literature suggests that most studies consider as null hypothesis the conservation in the mechanisms of regulation between human and rodents, and therefore a strong pattern of functional divergence has to be present so that it is observed and reported.

In order to bypass this bias, we used another method to estimate the proportion of species-specific binding sites, this time taking into account the distribution of divergence of each of the two functional classes of the 64 binding sites (shared function vs. species-specific function). We used these distributions to define the probability of shared function of a binding site between species given a value of divergence of the functional sequence
from the other species sequence. For each functional class we counted the number of occurrences for each interval of divergence equal to 0.1 (e.g. 0.00 to 0.10, 0.11 to 0.2, 0.21 to 0.3 etc), and calculated the proportion of values that fall within this interval for each class. We then estimated the probability that a site does not share function in the two species compared, by dividing, for each interval, the proportion of the species-specific values in this interval with the sum of proportions of species-specific and shared for the same interval. We then used the data from the other subset of the data for which there was functional information only for the human binding sites and computed the predicted number of sites with species-specific function by multiplying the probability defined above with the number of binding sites observed within the same interval of divergence. A total of 38 out of 96 binding sites were estimated to be human-specific (40%), which is higher than the estimate obtained directly from the experimental data, suggesting that the collection of comparative functional data may be slightly biased towards more conserved functions.

3.5 Discussion.

The results of the present study shed light on long-standing questions about the processes of evolution of transcription factor binding sites. The pattern of conservation of transcription factor binding sites suggests independent gain and loss in different phylogenetic lineages. The striking variation across sites in the degree of sequence conservation indicates that selective constraints are not always shared among phylogenetic lineages. Comparisons between human and rodents remain informative for
the identification of many essential regulatory regions and binding sites (Hardison et al. 1997; Wasserman et al. 2000). However, based on our analysis, a proportion between 32% and 40% of the functional human binding sites are not functional in rodents. It is possible that new binding sites have emerged in the rodent regulatory sequences that replace the function of the lost sites (Ludwig et al. 2000; Florea et al. 2000). This is very likely given the short length of binding sites and the degeneracy of sequence requirements of the binding factor. In addition, new functions or expression patterns may arise by the independent loss or gain of regulatory elements (Shasikan et al. 1998).

This pattern of evolution has important implications for the use of phylogenetic methods to identify functional regulatory elements for basic and medical research. Distant interspecific comparisons will reveal mainly highly conserved binding sites, and focusing only on those imposes an unfortunate bias in our understanding of regulatory variation. The highly conserved binding sites are those likely to have a radical effect on the expression of the gene, and nucleotide variation in these sites is likely to be associated with rare monogenic disorders. Complex disorders are likely to be mediated by common variants in less constrained binding sites (Risch and Merikangas 1996), precisely those sites that are missed in distant comparisons. On the other hand, comparisons of more closely related species are confounded by the low divergence even in non-functional sequences, which will produce many false positives. The positive aspect of our results is that 60-68% of the transcription factor binding sites are functionally conserved between human and rodents. Therefore, their nucleotide sequence is functionally constrained, and by using the appropriate parameters for calibration, which our data and analysis provides,
several methods will be able to identify them within human-rodent alignments of regulatory regions.

The small size of transcription factor binding sites and the degeneracy of binding requirements allows not only for the accumulation of conservative substitutions within binding sites but also the independent emergence of new binding sites, since many different nucleotide combinations will satisfy the binding requirements of a DNA binding protein (Berg and von Hippel 1987). These new sites may relax the evolutionary constraint in previously essential sites and lead to loss of some of them without serious phenotypic consequences (Ludwig et al. 2000). This pattern of evolution will make it difficult to identify regulatory elements that have undergone turnover. Thus a tight combination of probabilistic methods for binding site prediction, such as Hidden Markov Models (Eddy 1998; Durbin et al. 1998), study of polymorphism in promoter sequences, and extensive functional (Ren et al. 2000) and computational studies (Bussemaker et al. 2001) will be able to detect non-conserved binding sites. Detailed studies of regulatory sequence function combined with more sophisticated comparative genomics (Dubchak et al. 2000; Sumiyama et al. 2001), including comparison across multiple species of varying degrees of divergence (such as dog and rabbit), and polymorphism analysis will be informative in capturing the fluid regulatory landscape of mammalian genomes.

3.6 References.


Table 3.1: Summary of the data used in this study.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Species</th>
<th>Human binding sites (in bps)</th>
<th>Rodent binding sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>acid-labile subunit</td>
<td>Human, mouse</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>acute regulatory protein</td>
<td>Human, macaque</td>
<td>6, 9, 10, 8</td>
<td></td>
</tr>
<tr>
<td>adenosine deaminase</td>
<td>Human, mouse</td>
<td>13, 16, 31</td>
<td></td>
</tr>
<tr>
<td>adh1</td>
<td>Human, baboon, macaque, mouse</td>
<td>19, 26, 25, 44, 17, 13</td>
<td></td>
</tr>
<tr>
<td>aldolase A</td>
<td>Human, mouse</td>
<td>49, 25, 19</td>
<td></td>
</tr>
<tr>
<td>Apolipoprotein AI</td>
<td>Human, mouse</td>
<td>27, 29, 16</td>
<td></td>
</tr>
<tr>
<td>Apolipoprotein All</td>
<td>Human, mouse</td>
<td>7, 21, 14, 22, 17, 25, 8</td>
<td></td>
</tr>
<tr>
<td>Apolipoprotein CIII</td>
<td>Human, macaque, mouse</td>
<td>15, 52, 17, 22, 21, 21</td>
<td></td>
</tr>
<tr>
<td>Apolipoprotein E</td>
<td>Human, chimp, mouse</td>
<td>21, 12, 32, 21, 7, 6, 13, 19</td>
<td></td>
</tr>
<tr>
<td>App</td>
<td>Human, macaque</td>
<td>12, 12</td>
<td></td>
</tr>
<tr>
<td>atrial natriuretic factor</td>
<td>Human, mouse</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>c/ebp _</td>
<td>Human, mouse</td>
<td>16, 16, 17, 30, 16 / 21, 24</td>
<td></td>
</tr>
<tr>
<td>ccr5</td>
<td>Human, chimp, gorilla, orangutan, macaque, aotus, mouse</td>
<td>9, 10, 9, 18, 9, 16, 20, 10</td>
<td></td>
</tr>
<tr>
<td>CD68</td>
<td>Human, mouse</td>
<td>9, 13, 14</td>
<td></td>
</tr>
<tr>
<td>Cfr</td>
<td>Human, gibbon, macaque, saimiri, mouse</td>
<td>7, 18, 7, 7, 36</td>
<td></td>
</tr>
<tr>
<td>c-myb</td>
<td>Human, mouse</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>COL1A1</td>
<td>Human, mouse</td>
<td>28 / 15</td>
<td></td>
</tr>
<tr>
<td>c-reactive protein</td>
<td>Human, mouse</td>
<td>8, 11, 34</td>
<td></td>
</tr>
<tr>
<td>CYP1A1</td>
<td>Human, mouse</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>dio2</td>
<td>Human, rat</td>
<td>8, 7, 4</td>
<td></td>
</tr>
<tr>
<td>Erythropoietin</td>
<td>Human, mouse</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>factor IX</td>
<td>Human, chimp, macaque, mouse</td>
<td>22, 15, 15</td>
<td></td>
</tr>
<tr>
<td>_-interferon</td>
<td>Human, callithrix</td>
<td>13, 61, 13</td>
<td></td>
</tr>
<tr>
<td>GnrH</td>
<td>Human, rat</td>
<td>/ 8, 8, 8, 9</td>
<td></td>
</tr>
<tr>
<td>GRP78</td>
<td>Human, rat</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>growth hormone</td>
<td>Human, macaque, callithrix, mouse</td>
<td>27, 35, 28</td>
<td></td>
</tr>
<tr>
<td>Haptoglobin</td>
<td>Human, macaque, atelis, mouse</td>
<td>15, 21, 14, 8, 7, 8 / 8, 8, 7</td>
<td></td>
</tr>
<tr>
<td>Huntingtin</td>
<td>Human, chimp, gorilla, mouse</td>
<td>6, 8</td>
<td></td>
</tr>
<tr>
<td>interleukin-3</td>
<td>Human, chimp, macaque, callithrix, mouse</td>
<td>7, 9, 29, 22, 28</td>
<td></td>
</tr>
<tr>
<td>interleukin-5</td>
<td>Human, mouse</td>
<td>11, 5, 6, 16</td>
<td></td>
</tr>
<tr>
<td>L-plastin</td>
<td>Human, mouse</td>
<td>8, 15, 9, 6 / 6</td>
<td></td>
</tr>
<tr>
<td>monoamine oxidase</td>
<td>Human, chimp, gorilla, orangutan</td>
<td>7, 7, 7</td>
<td></td>
</tr>
<tr>
<td>msh2</td>
<td>Human, chimp, gorilla, orangutan, cercopithecus, callithrix</td>
<td>7, 5, 5, 6</td>
<td></td>
</tr>
<tr>
<td>mucin 1</td>
<td>Human, gibbon, mouse</td>
<td>7, 28, 14, 12, 6, 17, 8, 32, 6, 13, 14</td>
<td></td>
</tr>
<tr>
<td>mucin 2</td>
<td>Human, mouse</td>
<td>8, 10 / 10</td>
<td></td>
</tr>
<tr>
<td>MyoD</td>
<td>Human, mouse</td>
<td>7, 17, 6, 15, 22</td>
<td></td>
</tr>
<tr>
<td>Myoglobin</td>
<td>Human, mouse</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>neurofilament M</td>
<td>Human, chimp, gorilla, orangutan, mouse</td>
<td>7, 7, 12</td>
<td></td>
</tr>
<tr>
<td>olfactory marker protein</td>
<td>Human, mouse</td>
<td>23, 11, 11</td>
<td></td>
</tr>
<tr>
<td>Oxytocin</td>
<td>Human, mouse</td>
<td>6, 6, 6, 6</td>
<td></td>
</tr>
<tr>
<td>plasminogen activator</td>
<td>Human, rat</td>
<td>8, 10</td>
<td></td>
</tr>
<tr>
<td>platelet glycoprot. IB_</td>
<td>Human, mouse</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Preproinsulin</td>
<td>Human, chimp, cercopithecus, aotus, mouse</td>
<td>11, 30, 8, 5, 6, 7, 30, 8</td>
<td></td>
</tr>
<tr>
<td>Prohagcagon</td>
<td>Human, rat</td>
<td>37 / 91</td>
<td></td>
</tr>
<tr>
<td>p-selectin</td>
<td>Human, mouse</td>
<td>/ 23, 9, 7</td>
<td></td>
</tr>
<tr>
<td>rh50</td>
<td>Human, chimp, gorilla, orangutan, gibbon, macaque, mouse</td>
<td>6, 6, 13, 8, 7, 7</td>
<td></td>
</tr>
<tr>
<td>s- cardiac troponin C</td>
<td>Human, mouse</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Sry</td>
<td>Human, chimp, gorilla, cercopithecus, mouse</td>
<td>7, 8</td>
<td></td>
</tr>
<tr>
<td>Surfl-2</td>
<td>Human, mouse</td>
<td>12, 11, 14, 6, 6, 11</td>
<td></td>
</tr>
<tr>
<td>Thyroglobulin</td>
<td>Human, rat</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>TNF</td>
<td>Human, chimp, gorilla, orangutan, gibbon, macaque, baboon, aotus, mouse</td>
<td>10, 10, 10, 10, 8, 8, 10, 6, 6, 8 /</td>
<td></td>
</tr>
</tbody>
</table>

* Size of binding sites mapped in the human sequence and used in the analysis. See also Figure 1 for binding factors

b Rodent-specific binding sites
Figure 3.1a: Regulatory regions with available functional data only for human. Graphical representation of the promoter data used for human (top), and rodents (bottom). Shapes indicate binding sites and symbols inside the human sequence indicate the binding factor. Some sites do not have a definitive binding factor but they are verified to be functional (indicated with U), and some others have multiple binding factors with overlapping sequences (indicated with M). In the case where no sequence data were available, the space is left blank for the respective species. Dashed lines and dashed shapes indicate that the regulatory sequence was available, but no significant alignment was found for the respective region. Del indicates deletion of the binding site due to a larger deletion of the sequence in the respective position. Numbers inside the shapes for the sequences of rodent or human (for the rodent-specific sites) indicate the number of nucleotide that are different this species sequence from the human reference functional sequence (including gaps). For the sizes of the binding sites refer to Table 1;
Figure 3.1b: Regulatory regions with available functional data for both human and rodent. Arrows indicate the species in which the binding site is functional. Graphical representation of the promoter data used for human (top), and rodents (bottom). Shapes indicate binding sites and symbols inside the human sequence indicate the binding factor. Some sites do not have a definitive binding factor but they are verified to be functional (indicated with U), and some others have multiple binding factors with overlapping sequences (indicated with M). In the case where no sequence data were available, the space is left blank for the respective species. Dashed lines and dashed shapes indicate that the regulatory sequence was available, but no significant alignment was found for the respective region. Del indicates deletion of the binding site due to a larger deletion of the sequence in the respective position. Numbers inside the shapes for the sequences of rodent or human (for the rodent-specific sites) indicate the number of nucleotide that are different this species sequence from the human reference functional sequence (including gaps). For the sizes of the binding sites refer to Table 1;
Figure 3.2a: Distribution of divergence within binding sites for human-macaque.
**Figure 3.2b:** Distribution of variance from 1000 simulations of a random Poisson process of substitution within binding site sequence for the human-macaque divergence level; the observed value is indicated with an arrow.
Figure 3.3a: Distribution of divergence within binding sites for all the data between human-rodents.
Figure 3.3b: Distribution of divergence for the binding sites with shared function between human-rodents.
Figure 3.3c: Distribution of divergence for the binding sites with species-specific function in human and rodents.
CHAPTER 4: Differential selection after duplication in mammalian developmental genes
4.1 Abstract.

Gene duplication provides the opportunity for subsequent refinement of distinct functions of the duplicated copies. Either through changes in coding sequence or changes in regulatory regions, duplicate copies appear to obtain new or tissue-specific functions. If this divergence were driven by natural selection, then we would expect duplicated copies to have differentiated patterns of substitutions. We tested this hypothesis using genes that duplicated before the human/mouse split and whose orthologous relations are clear. The null hypothesis is that the number of amino acid changes between human and mouse is distributed similarly across different paralogs. We used a method modified from Tang and Lewontin (1999, Genetics 153: 485) to detect heterogeneity in the amino acid substitution pattern between those different paralogs. Our results show that many of the paralogous gene pairs appear to be under differential selection in the human/mouse comparison. The properties that led to diversification appear to have arisen before the split of the human and mouse lineages. Further study of the diverged genes reveals insights regarding the patterns of amino acid substitution that resulted in differences in function and/or expression of these genes. This approach has utility in the study of newly identified members of gene families in genome-wide data mining and for contrasting the merits of alternative hypotheses for the evolutionary divergence of function of duplicated genes.
4.2 Introduction.

Duplication is a nearly ubiquitous property of genes in higher eukaryotes. It is believed that gene duplication provides the material for the generation of new genes and subsequently new or more specialized functions. The details of how members of duplicated gene pairs acquire new function are largely unknown. The classical view is that selection is relaxed, because if one gene gains a mutation that alters function, the remaining copy serves as a backup to retain the original function (Ohno 1970; Ohta 1987). The divergent copy is then free to acquire random substitutions, and by chance, these substitutions may result in new tissue specificity or other function.

Recently several alternative models have been proposed that consider the dynamics of such processes and determine the relative likelihoods of gain of new function versus loss of function under the pressure of random mutations. Ota and Nei (1994) and Nei et al. (1997) have proposed a birth-and-death process for the evolution of immunity genes, suggesting that gain and loss of genes plays a dominant role in the evolution of multigene families. Clark (1994) showed that the classical model cannot account for the levels of duplicated gene preservation observed and that populations in mutation-selection balance are likely to experience weak positive selection for the initial duplication. Walsh (1995) and Hughes (1994) suggested that the evolution of new functions must be rather frequent. Zhang et al. (1998) demonstrated that positive selection is probably acting on ribonuclease genes leading to new adaptive functions. Recently, Force et al. (1999) formalized and modeled the idea of subfunctionalization of regulatory regions, whereby the original genes had multiple expression domains, and are then free to lose those
subfunctions in a complementary gene-specific manner. Subfunctionalization appears to be consistent with the nature of differences in regulatory regions, and this model can explain the relatively high preservation of duplicate copies compared to the neutral expectation.

In another study, Lynch and Force (2000) modeled the evolution and preservation of duplicated genes under their DDC model (Duplication — Degeneration — Complementation) and concluded that the data provide good support for the idea of subfunctionalization. Furthermore, many proteins have subunits with distinct functions, and the subfunctionalization model can be applied to these subfunctions within the coding region (Henikoff et al. 1997). If a gene has two domains, each interacting within a specific tissue, subsequent duplication could eventually lead to subfunctionalization of the protein sequence such that each copy can interact with components of one tissue in a complementary manner. This divergence could occur independent of the subfunctionalization of the regulatory sequences, but some interactions are possible. Interest in detection of regions that contribute to functional divergence is increasing but very few studies have proposed rigorous methods for detecting them (e.g. Gu 1999). Here we propose a novel method for identifying heterogeneity in patterns of substitutions along duplicated genes and present results from the analysis of 12 gene families present in the human and mouse genomes.
4.3 Materials and Methods.

**Sequences used:** The gene families that we used were taken from the literature, consisting exclusively of genes for which phylogenetic information was available to guarantee orthologous relationships. Orthology was independently investigated by searching OMIM (Online Mendelian Inheritance of Man, www.ncbi.nlm.nih.gov/omim) and MGI (Mouse Genome Informatics, www.informatics.jax.org/mgihome) for presence of those genes in syntenic positions (e.g. Nadeau 1989) and for similar or identical tissue-specific expression patterns. Also most of the information about the expression patterns of those genes, presented below, was obtained mainly from these two databases and references mentioned therein. In all cases, full-length amino acid sequence was used in the analysis. The gene families and respective gene members that were used include:

(Genbank accession number in parentheses for human and mouse respectively): **Mef2:** Mef2A (Q02078, Q60929), Mef2B (Q02080, JC4221), Mef2C (Q06413, A47561);

**MyoD:** MyoD1 (P15172, P10085), Myf5 (P13349, P24699), Myf6 (P23409, P15375), MyoG (P15173, P12979);

**Bmp:** Bmp2 (P12643, P21274), Bmp6 (P22004, P20722), Bmp7 (P18075, P23359);

**Notch:** Notch1 (P46531, Q01705), Notch4 (AAC32288, P31695);

**Elav:** HuA (AAB41913, AAB17967), HuB (AAA69698, AAC52644), HuC (AAA58677, AAC52999);

**Hox1:** Hoxa1 (U10421, NM_010449), Hoxb1 (NM_002144, NM_008266);

**Hox3:** Hoxa3 (NM_043365, P02831), Hoxb3 (NM_002146, P09026), Hoxd3 (P31249, P09027);

**Hox4:** Hoxa4 (AAB97947, P06798), Hoxb4 (P17483, P10284), Hoxc4 (Q08624, P09017), Hoxd4 (P09016, P10628);

**Hox5:** Hoxa5 (P20719, P09021), Hoxb5 (P09079, P09067);

**Hox7:** Hoxa7 (NP_008827, P02830), Hoxb7
(P09629, P09024); **Hox9:** Hoxa9 (P31269, P09631), Hoxd9 (P28356, P28357); **Hox13:**
Hoxa13 (P31271, Q62424), Hoxd13 (P35453, P70217).

**Description of the paralog heterogeneity test:** The test is designed to identify
differences between pairs of duplicated genes in the pattern of interspecific sequence
divergence. Human and mouse sequences were used because of the rapidly expanding
sequence data available for both species. We extracted Genbank amino acid sequence
entries that correspond to gene families from the two species. We aligned all members of
each family for both species using ClustalW and deleted all the gapped sites (complete
deletion). Deleting the gapped sites may eliminate some of the regions that functionally
differentiate the proteins of the gene family. However, we wanted to have a complete
correspondence of homologous sites for comparison between different paralogs.
Phylogenetic analysis has resolved the relationships of most of the genes used in this
study (Zhang and Nei 1996; Hughes 1999). However, for each family we constructed a
phylogenetic tree to confirm that our alignments produce the correct orthologous
relationships. We then calculated a statistic $Q_i$ (from Tang and Lewontin 1999) for each
variable site in each orthologous comparison as:

$$Q_i = \frac{j}{v} - \frac{i}{l}$$

where $j = 1, 2, 3$, $v$ is the index of the current, or $j$th, variable site, $v$ is the total number
of variable sites, $i$ is the nucleotide or amino acid position of $j$th variable site in the
sequence and $l$ is the total length of the sequence in amino acids or nucleotides (Figure
1). $Q_i$ is positive for runs of sites that are closer together than a uniform distribution, and
$Q_i$ is negative in regions of sparse variation. We subsequently calculated $Q_i$ for non-
variable sites by linear interpolation between the two flanking variable sites. So for a
non-variable site at position $i$ between variable sites at positions $x$ and $k$ on the sequence the value of $Q_i$ is:

$$Q_i = Q_x \frac{k-i}{k-x} + Q_k \frac{i-x}{k-x}$$  \hfill (2)

For sites before the first variable site and after the last variable site we consider $Q_i$ to be zero at the two ends of the sequence.

The main idea behind the test is that even when the distribution of substitutions (variable sites) is uneven, the pattern of this heterogeneity will be similar between different paralogs if the functional domains are the same. To test homogeneity of $Q_i$ we calculate the difference between $Q_i$ s at the same position for two different human-mouse comparisons (two different paralogs):

$$D = Q_{1,i} - Q_{2,i}$$  \hfill (3)

where $Q_{1,i}$ and $Q_{2,i}$ are the $Q_i$ values for $i$th position in human-mouse comparison in paralogous gene copy 1 and paralogous gene copy 2 respectively. The expectation is that, if the substitution pattern is similar in the two copies, the $D$ — values will be low given the number of variable sites. Finally, we calculate the longest stretch of monotonic increase or decrease of the $D$ values for these two copies and we report the absolute differential $R$ of $D$ values in this run. Significance levels were determined by randomly assigning the same number of observed substitutions in each of the sequences, determining the largest monotonic run of $D$, and calculating the $R$-value for this run. By doing this 1000 times, we obtain a null distribution of the $R$-values. Analysis was also performed with the alternative hypothesis that there are two or three regions of differential substitution pattern. In this case we calculated the sum of differentials ($R$ - values) for the top two or three regions. Significance levels again were determined by random permutations.
identical to the ones for one region, but this time the sum of the largest two or three $R$-values was used to generate the null distribution. The need for this additional analysis is illustrated with simulations in Tang and Lewontin (1999). The rationale is that there may be two or more regions that each exhibits rather small heterogeneity effect, in which case the effect is distributed across several different regions. The fact that we use random permutation to determine significance levels eliminates any bias in the analysis, because we calculate the same statistic for the permuted sets. The test is directional in the sense that decline of $D$ means lower variation, within the region examined, in the first paralog than the second paralog in the region examined, while increase of $D$ means higher variation in the region examined in the first paralog than in the second paralog.

We assume under the null hypothesis that if the different orthologous groups have the same relative substitution properties across the sequence, there should be no heterogeneity in the distribution of substitutions across the sequence, and the test should fail to reject the null hypothesis. On the other hand, if different regions were evolving differently among paralogs, there would be detectable heterogeneity. This method cannot detect substitution rate acceleration if the effect is proportional across the gene for each copy, but it will detect accelerated rates of substitutions if they occur in different regions across different paralogs (Figure 2).
**Power calculations:** In order to calculate the power of the test we simulated the accumulation of substitutions in a protein of a length of 500 residues consisting of two equal-sized domains of 250 residues. Numbers of substitutions in the two domains for the two orthologous pairs were sampled from a Poisson distribution such that the average overall rate of their sequence divergence was preserved. The two domains in each orthologous pair were assumed to diverge at different rates, and we report the ratio of those rates (rate in domain 1 / rate in domain 2). In all simulations the domains accumulated substitutions differently across orthologous pairs such that the domain that had the high rate in the one orthologous pair had low rate in the other pair and vice-versa for the other domain. The ratio of the divergence rates of the two domains of one orthologous pair was precisely the reciprocal of the ratio of rates in the other pair. We simulated 1000 sets of two pairs of genes (two orthologous groups in two species) with both orthologous groups evolving with three different scenarios:

**Scenario 1:** both pairs of orthologs were diverging with high and equal rates (divergence of orthologs with rate 0.45, labeled *high* in Figure 3).

**Scenario 2:** both pairs of orthologs were diverging with low and equal rates (divergence of orthologs with rate 0.13, labeled *low* in Figure 3).

**Scenario 3:** one of the two pairs of orthologs was diverging with low rate (0.13) and the other with high rate (0.45) (labeled *dif* in Figure 3). The ratios of rates between domains were still precisely reciprocal across the two comparisons.

We did not simulate the divergence of paralogs before speciation because the divergent sites will be shared within the orthologs so they do not contribute to the pattern. The rates
of divergence that were simulated are within the range observed in the analyzed data of the twelve families studied.

Subsequently, we calculated $R$ for each of the 1000 sets of the two orthologous pairs. The significance was determined by comparing the values of $R$ with a null distribution of $R$-values derived from simulations assuming equal rates for the two domains and identical to the simulated sets above for the whole amino acid sequence. The test in this case becomes more conservative because we did not generate a null distribution by permutation for the actual number of substitutions for each simulated set, so the variance of the null distribution is higher given that different counts of substitutions are included in this distribution.

**Software and alignments:** Computer programs were written in Perl and are available upon request. All the alignments used in this analysis are also available upon request.

### 4.4 Results — Discussion.

We examined 12 gene families that have at least two orthologous pairs in mouse (*Mus musculus*) and human. The pairwise paralog heterogeneity tests revealed 7 out of 12 families with significant heterogeneity (Table 1). This heterogeneity appears to have occurred before the human-mouse split, because when we tested for heterogeneity in the amino acid substitutions between paralogs in the two species, this time obtaining the pattern of amino acid substitutions between paralogs within the same species and comparing it to the pattern of amino acid substitution in the same paralogous comparison in the other species, the null hypothesis was never rejected. In other words, the
heterogeneity appears as differences among paralogous gene copies rather than a species-specific effect in the amino acid substitution pattern.

Out of the seven Hox gene families tested only three showed significant heterogeneity, while the proportion of non-Hox genes that showed significant heterogeneity was much higher. These results are consistent with the widely accepted hypothesis that Hox genes evolve mainly through their regulatory regions, rather than their coding region (Holland et al. 1994). This illustrates once more that gene regulation is an essential component of gene function and evolutionary potential.

**Power of the test:** The results of the calculations for the power of the test are summarized in Figure 3. We simulated a rather simple model that just assumes two differentially evolving regions. In reality the patterns of substitutions can be much more complicated. However we illustrate that for this case the power of the test is very good if the ratio of rates between the two regions is more than 0.5, which is realistic if the diversification occurred long enough before present. This ratio of rates is not uncommon since we are considering the scenario where one of the two domains has lost function in one paralog and the other domain has lost function in the other paralog.

**Function of genes that have diverged:** One of the key predictions of this model is that the regions that will appear to be divergent in substitution pattern are the ones that confer subfunctions. For example, in the case of a transcription factor, the domain that would probably be conserved is the DNA binding domain, while tissue-specific domains (including transactivation domains) will diverge because of subfunctionalization. In order to find specific examples of such patterns, we examined more closely two of the
gene families, MyoD and Mef2, consisting of transcription factors involved in myogenic fate determination in mammals.

The central region of the MyoD genes encodes a helix-loop-helix domain responsible for DNA binding. From the plots of $D$ values across the different pairwise comparisons, it is clear that there is little difference in the amino acid substitution pattern in this central region. On the other hand, the trans-activation domain, which is at the C-terminus of the proteins, appears quite diverged (Figure 4a). In particular, MyoG has accumulated a large number of amino acid differences in both mouse and human. This is consistent with the idea that MyoG is a member of the MyoD family with very distinct functions and expression patterns. However, it appears that MyoD and Myf5 are also different in the pattern of amino acid substitutions as indicated in Table 1, with the greatest heterogeneity residing in the transactivation domain (C-terminus). The general pattern of divergence in the transactivation domain is consistent with evolution driven by the interactions of the molecule with the cellular environment.

A similar pattern is observed for the Mef2 (MADS box) gene family. Differences are pronounced in the C-termini, the transactivation domain, while the DNA binding domain (MADS and Mef2 domains) is generally conserved (see Figure 4b). The three members appear to have different roles in muscle differentiation and all three have different patterns of substitution across the amino acid sequence. Mef2A is expressed in skeletal and cardiac muscle, while Mef2C is expressed in skeletal muscle and the brain. Mef2A is thought to be involved in the induction of the differentiation process in skeletal muscle, while Mef2C maintains this differentiated state. Mef2B on the other hand appears to be more redundant in terms of function and expression, expressed in skeletal
and cardiac muscle and the brain. Consistent with the prediction, Mef2A and Mef2C appear to have significantly different patterns of amino acid substitution, while Mef2B, which overlaps in function and expression with both, is not significantly different from either of the two others. In addition, we are able to separate candidate regions for the functional divergence in the Mef2A — Mef2C comparison. The maximum differential (R) is observed between sites 86 and 125 (in the ungapped alignment - R = 0.609; increase of D) followed by 126 to 183 (R = 0.3769; decline of D) and 255 to 301 (R = 0.206; decline of D). These become likely candidates for regions that are responsible for interactions with the cardiac tissue (Mef2A specific) or the brain (Mef2C specific). In particular, region 86 — 125 is highly variable in Mef2A but conserved in Mef2C. Therefore it is a candidate region for interactions with components of the brain. On the other hand, regions 126 — 183 and 255 — 301 are conserved in Mef2A and are highly variable in Mef2C, which makes them good candidates for interactions with components of the cardiac muscle tissue.

One possibility is that the functional divergence observed in these two gene families may also be due to the fact that Mef2 genes interact genetically and biochemically with the genes of the MyoD family. Intriguingly, the regions that appear to be different between different homologous groups are the C-termini, which are the transactivation domains for both families. It has been demonstrated that there is direct biochemical interaction between the Mef2 proteins and the MyoD proteins (Molkentin et al. 1995). This interaction in some cases makes the C-terminus (or parts of it) of either the Mef2 or the MyoD gene dispensable. Assuming that the pairwise interactions are specific in terms of which Mef2 protein interacts with which MyoD protein, either due to
co-expression in the same tissue or because of biochemical affinity, there would be co-evolution of the protein pairs that exhibit the interaction. In addition, the transactivation domains, which interact in the nucleus after DNA binding, may have evolved according to the specific components they bind to. This would expose the transactivation domains to different selective pressures.

On the other hand, the expectation is that duplicated genes evolving through their regulatory regions will not show heterogeneity in the rate of amino acid substitutions. Interestingly, a recent study (Greer et al. 2000) has shown that it is not the amino acid sequence responsible for the functional differences between Hoxa3 and Hoxd3, but rather quantitative modulation in gene expression. The Hox3 family did not show heterogeneity in our analysis ($P > 0.2$, Figure 4C), consistent with no subfunctionalization in the coding sequence, as predicted by the experimental data.

**Conclusion:** Our results demonstrate that coding subfunctionalization is an important component in the evolution of duplicated genes. They also support previous theoretical work (Clark 1994, Lynch and Force 2000) that claims that the classical model of evolution of duplicated genes does not adequately explain the level of duplicate gene preservation. Duplicated genes undergo many steps of functional divergence some of which leave their footprint on the amino acid sequences. Finally, we propose a simple and flexible method for the analysis of homologous genes that have been a product of gene duplication. This method not only can detect heterogeneity in different duplicate copies, but it can also indicate candidate regions for functional divergence. This is of particular interest in mammalian genomics, because of the ubiquity of gene duplication.
As we gain access to complete genome sequences of human and mouse, such approaches could be of great value in generating decisive experimentally testable hypotheses.

4.5 References.


Table 4.1: Data for all the pairwise comparisons of the gene family members.
Comparisons with statistically significant heterogeneity (determined by the paralog heterogeneity test) are in bold ($P < 0.05$). Results are presented for the alternative hypotheses that there are 1, 2 or 3 divergent regions.

<table>
<thead>
<tr>
<th>Gene family</th>
<th>Comparison</th>
<th>1 region</th>
<th>2 regions</th>
<th>3 regions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p-value</td>
<td>p-value</td>
<td>p-value</td>
<td></td>
</tr>
<tr>
<td>Mef2</td>
<td>Mef2A — Mef2B</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td>Mef2A — Mef2C</td>
<td>0.027</td>
<td>0.026</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>Mef2B — Mef2C</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>MyoD</td>
<td>Myf6 — MyoG</td>
<td>0.130</td>
<td>0.081</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>Myf6 — MyoD</td>
<td>0.120</td>
<td>0.075</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>Myf6 — Myf5</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td>MyoG — MyoD</td>
<td>0.120</td>
<td>0.035</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>MyoG — Myf5</td>
<td>0.021</td>
<td>0.008</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>MyoD — Myf5</td>
<td>0.03</td>
<td>0.014</td>
<td>0.035</td>
</tr>
<tr>
<td>Bmp</td>
<td>Bmp6 — Bmp7</td>
<td>0.045</td>
<td>0.075</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>Bmp6 — Bmp2</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td>Bmp7 — Bmp2</td>
<td>0.037</td>
<td>0.017</td>
<td>0.008</td>
</tr>
<tr>
<td>Notch</td>
<td>Notch1 — Notch4</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Elav</td>
<td>HuA — HuB</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td>HuA — HuC</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td>HuB — HuC</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Hox1</td>
<td>Hoxa1 — Hoxb1</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Hox3</td>
<td>Hoxa3 — Hoxb3</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td>Hoxa3 — Hoxd3</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td>Hoxb3 — Hoxd3</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Hox4</td>
<td>Hoxa4 — Hoxb4</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td>Hoxa4 — Hoxc4</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td>Hoxa4 — Hoxd4</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td>Hoxb4 — Hoxc4</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td>Hoxb4 — Hoxd4</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td>Hoxc4 — Hoxd4</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Hox5</td>
<td>Hoxa5 — Hoxb5</td>
<td>0.038</td>
<td>0.074</td>
<td>0.052</td>
</tr>
<tr>
<td>Hox7</td>
<td>Hoxa7 — Hoxb7</td>
<td>0.120</td>
<td>0.073</td>
<td>0.079</td>
</tr>
<tr>
<td>Hox9</td>
<td>Hoxa9 — Hoxd9</td>
<td>0.037</td>
<td>0.015</td>
<td>0.025</td>
</tr>
<tr>
<td>Hox13</td>
<td>Hoxa13 — Hoxd13</td>
<td>0.041</td>
<td>0.073</td>
<td>0.052</td>
</tr>
</tbody>
</table>

ns: indicates $P > 0.2$
Figure 4.1: A graphical description of the test in a case of nucleotide sequence.

Equations below the sequences demonstrate the method for the calculation of $Q_j$ for each position $i$ in the sequence. $j$ is the index of a variable site, _ is the number of mismatches between the mouse and the human sequence and $l$ is the total length of the sequence.
**Figure 4.2:** Hypothetical tree structure for each of two domains when there is: (a) accelerated substitution rate in one orthologous pair across the whole sequence in both species and (b) accelerated substitution rate differing across domains in the two different orthologous groups. The paralog heterogeneity test will be able to detect heterogeneity in (b) but not in (a).
**Figure 4.3:** The plots indicate the proportion of tests rejected for the various cases described in the methods; *high* is the case where both divergence rates between orthologous genes were 0.45, *low* where both divergence rates where 0.13 and *dif.* is the case where one orthologous group had a rate of 0.13 and the other a rate of 0.45. Not that the Rate ratio indicated in this figure is the ratio of low to high rate in all cases.
Figure 4.4: Plots of the $D$-values for comparisons in three gene families. Increase or decline indicates differential rate of substitution, while plateaus indicate very similar substitution rate across paralogs. Note that the amino acid position refers to the alignment after complete deletion (ungapped). (a) Plot of $D$-values for the pairwise comparisons in MyoD family; note the divergent domains at the C-termini of the proteins; the gray filled bar indicates one of the two candidate regions of functional divergence between MyoG and Myf5; the black filled bar indicates the candidate region for functional divergence in all four significant comparisons; (b) Plot of the $D$-values for the pairwise comparisons Mef2 family; note the divergent domain at the C-termini of the proteins; the black filled bars indicate the candidate regions for functional divergence between Mef2A and Mef2C; (c) Plot of the $D$-values for the pairwise comparisons in Hox3 gene family; no comparison is significantly heterogeneous. Comparisons with statistically significant heterogeneity are starred (*).
Figure 4.4: (continued)
Figure 4.4: (continued)
CHAPTER 5: Evidence for interaction of nucleotide positions within transcription factor binding sites: applications for binding site prediction.
5.1 Abstract.

A lot of attention has been given recently on the annotation of gene regulatory sequences. This process requires good knowledge of the features, the content and structure of these sequences. One of the approaches used for this annotation is the prediction of the location of binding sites of transcription factors suspected to be regulating the gene under study. This is usually achieved with the use of position weight matrices which describe the probability of each of the four nucleotides in the experimentally verified binding sites of a given transcription factor. In this report we add one more parameter to these models and tease apart interactions between nucleotide positions within binding sites. We statistically demonstrate that positions within binding sites are associated with each other, leading to non-random occurrence of nucleotide pairs in experimentally verified binding sites. We attempt to quantify these interactions and their effect in real and simulated binding site data. We finally incorporate these interaction features in a binding site prediction model that takes into account the non-random association of adjacent positions within a binding site sequence. Our results strongly suggest that by incorporating these features in prediction models we achieve higher binding site prediction efficiency and reduction in the number of false positives, a frequent problem in binding site prediction.
5.2 Introduction.

Interpretation of the content and structure of regulatory regions has been one of the major focuses in genomic research (Clark 2001; Hardison 2000; Bergman and Kreitman 2001). Regulatory regions consist of a variety of transcription factor binding sites that are organized in such a manner that facilitate protein-protein and protein-DNA interactions and result in the efficient regulation of the expression of the gene. One of the main characteristics of these binding sites is that they have some degree of degeneracy (Claverie 1994; Stormo 2000). In other words, a given transcription factor binds a variety of short sequences (usually 6-14 bps), which are similar but not identical to each other. This sequence degeneracy may also be associated with varying degrees of binding affinity of the protein with the binding site sequence (Berg and von Hippel 1987). In addition, the degree of degeneracy is not the same across all the positions of the binding site, and therefore some nucleotide positions have stricter requirements than others, in order for the sequence to be bound by the transcription factor.

These unique features of the transcription factor binding sites combined with their very short length increase their occurrence by chance, making their identification within regulatory regions problematic (Stormo 1998; Hardison et al 1997). Binding site prediction methods frequently make use of position weight matrices, which describe the frequency of each of the four nucleotides along the binding site sequence and by using approaches such as Hidden Markov Models (Eddy 1998; Durbin et al. 1998; Yada et al. 1998) attempt to assign a likelihood of functional role of a given sequence, in some cases weighted by the nucleotide composition bias in the genome. Some approaches have gone
even further trying to uncover and use the relative positioning of the binding sites within
the regulatory region in order to reduce false positives and obtain a more meaningful
signal (GuhaThakurta and Strom 2001; Liu et al. 2001) or quantify the protein-DNA
interactions (Benos et al. 2001). Other methods combine binding site prediction with
conservation of sequence (Zhu et al. 1998; Wasserman et al. 2000; Dubchak et al. 2000;
Sumiyama et al. 2001) between species to obtain independent sorts of evidence for the
potential functional role of a short sequence.

A characteristic that has been consistently excluded in all these studies is the
interdependence of nucleotides within the binding site. All the above models assume that
the frequencies of the four nucleotides in each position of the binding site are
independent of other positions. However, given the molecular interactions occurring
when a protein binds DNA it is very likely that two or more nucleotide positions within
the binding site interact in order for the protein-DNA interaction to be efficient and only
certain combinations of nucleotides are allowed. In such case, a simple position weight
matrix does not reflect the complexity of the binding site sequence and will result to
many false positives and low resolution in binding site prediction.

In this report we attempt to tease apart some of the properties of the interaction of
nucleotide positions within binding sites of well-characterized transcription factors in
Drosophila and mammals. We first describe the degree of the interaction between
positions as this is inferred by statistical tests and then we develop and test models of
binding site prediction that take into account these interactions. We discuss the
implications of our results for the annotation of regulatory regions and for the
understanding of the molecular interactions between transcription factors and their DNA targets.

5.3 Methods.

**Binding site collection:** Binding sites were collected from the SELEX database ([http://wwwmgs.bionet.nsc.ru/mgs/systems/selex/](http://wwwmgs.bionet.nsc.ru/mgs/systems/selex/)), from the TRANSFAC database (Wingender et al. 2000) for NF-kB, YY1 and MEF-2 in humans and from data kindly provided by Casey Bergman for BCD and HB in Drosophila (see Chapter 2). We also generated 50 artificial sets of 100 binding sites each, with a fixed binding site length of 10 bps. The simulation was performed by constructing $4 \times 4 \times 9$ random matrices describing the association of adjacent sites within a binding site, and sampling from the frequencies of those matrices to obtain real sequence for 100 binding sites that corresponded to the properties of the matrix they derived from.

**Regulatory regions:** For a real test of our binding site prediction method we used regulatory regions that contained experimentally verified binding sites for the 5 well-characterized transcription factors. Experimental information in these regions was compared with the results from the binding site prediction method described below. These regions with the Genbank accession numbers are: human $TNF_\alpha$ promoter (L11698) and Drosophila $eve$ stripe-2 enhancer (AF042709).

**Binding site prediction models:** Two models of prediction were used. One of the models assumed independence of nucleotide positions within the binding site sequence (see Fig 1
for probability weight matrices) hereafter referred to as 1-site model and the other assumed association of adjacent sites within the binding site sequence and is referred to as 2-site model.

2-site model: For the 2-site model, the probability of a nucleotide in each position was conditioned on the nucleotide of one of the two adjacent nucleotides. This approach requires a starting position within the binding site to initiate the calculation of the likelihood under the adjacent-position-dependence model. We assigned this position to be the one with the most biased nucleotide frequencies in the 1-site probability weight matrix. This was quantified by subtracting the expected frequency of each nucleotide under no bias (0.25) from the frequency of each of the 4 nucleotides and summing up the absolute values of these 4 differences. This metric $b$ has the maximum value ($b_{\text{max}} = 1.5$) if a position has a frequency of 1 for one nucleotide and 0 for the other three and the minimum value ($b_{\text{min}} = 0$) when all four nucleotides have the same frequency (0.25). We then calculate the likelihood ($L$-value) of each short sequence of the same length as the binding site of the factor tested going outwards from the most biased position, assigned previously, to the two ends of the binding site sequence according to the equation below:

$$L = \log ps \leftrightarrow \log p_i \leftrightarrow \log p_j$$

where $s$ is the starting position, $i$ is the position downstream of the starting position, $j$ is the position upstream of the starting position, $ps$ is the frequency of the observed nucleotide in the starting position, and $p_i$ and $p_j$ is the frequency of the observed nucleotide in the 2-site weight matrix, which is the probability of this nucleotide conditioned on the nucleotide of the previous position going outwards of the starting position (see Figure 2 for graphical representation).
The calculation is similar to the one described in Chapter 2 with the only difference being that instead of using the frequency of the observed nucleotide in the 1-site weight matrix, we use the conditional probability of the observed nucleotide given the nucleotide in the previous position, as we go outwards from the most biased site (starting position of the model). We believe that this is the most appropriate starting position of the model since a highly biased site (e.g. one that has a very high frequency for one of the nucleotides) will not be significantly associated with any other position and thus it is a good independent starting point. For the cases where more than one position had the same \( b \) value, we arbitrarily started from the one towards the 5' of the binding site. Test results revealed no difference in the pattern described when started from alternative equally biased positions.

1-site model: Calculation of the \( L \)-values for the 1-site model was done according to Chapter 2.

For the regulatory region binding site prediction thresholds for both the 1-site and the 2-site models we used the distribution of \( L \)-values from a bootstrap approach (as described in Chapter 2) to determine the 95% confidence interval of \( L \)-value of the binding sites of the transcription factor we scanned for.

Degree of association: For the simulated binding site sets we correlated the degree of adjacent site dependence with the difference in the \( L \)-values produced by the 1-site model vs. the 2-site model. The degree of association was determined by the sum of chi-square statistics of the 4x4 matrices for each pair of adjacent sites. The simulated binding sites had a length of 10 nucleotides and therefore there were 9 chi-square values for each of the 50 simulated sets. For the simulated sets of binding sites the expectation was that the
higher the sum of chi-square values the higher the $L$-values from the 2-site prediction method.

### 5.4 Results.

**Degree of association between adjacent positions within binding sites:** We tested the degree of association of adjacent positions within sets of experimentally verified binding sites of 5 well-characterized mammalian (MEF-2, NF-kB, YY1) and *Drosophila* (BCD, HB) transcription factors. We observed strong significant associations between nucleotide positions in HB, BCD, NF-kB, Mef-2 and weaker associations in YY1 (Fig. 1). Our results are not biased by the structure of the data since the associations did not result from the sampling properties of the binding sites. The HB binding site haplotypes were present in sites from more than one regions sampled, the NF-kB sites were collected from SELEX data and experimental regulatory regions but there was no bias in the presence of the haplotypes in one or the other set and the MEF-2 and YY1 sites were entirely from SELEX data. The above associations suggest that adjacent positions have dependent structure to accommodate the requirements for the transcription factor binding to the DNA target.

The strong association observed suggests that the degeneracy of the binding site sequences for a given transcription factor has some underlined structure. This structure could be used to construct binding site prediction models, which will be able to incorporate more information than just the frequency of each of the four nucleotides in each position of the binding site. We attempt to construct these models and explore the
efficiency of binding site prediction when the association of positions is incorporated and how this efficiency correlates with the degree of adjacent position association.

**Properties and association in simulated binding site sets:** Fifty sets of 100 binding sites each were generated from simulated matrices (see methods). We then searched for the most biased position of the binding site in each of the sets and constructed 1-site and 2-site probability matrices. We then used these matrices to obtain the $L$-value for each of the 100 binding sites for each set from the 1-site and 2-site prediction models. For each binding site within each set we obtained the difference between the 2-site $L$ and the 1-site $L$.

The sum of the chi-square values for the association of adjacent positions was calculated for each of the 50 binding site sets. We then run a regression between the average difference in $L$-values between 2-ste and 1-site prediction models in each set of binding sites with the sum of chi-square values for each set. The regression was highly significant with a positive slope ($R^2 = 50\%, P = 0.0001$) indicating that the stronger the interposition association (higher sum of chi-square values) the higher the $L$-value from the 2-site model was from the 1-site model $L$-value (Fig. 3). Note that if there is no association between adjacent sites the $L$-values should be identical to the ones obtained from the 1-site model since the transition probability of the nucleotide in position $i$ will be the same regardless of the nucleotide in the adjacent position ($i+1$ or $i-1$ depending on the starting position) and the same as the frequency of that nucleotide in position $i$ in the 1-site probability weight matrix.
Association of positions within binding sites of known transcription factors: We collected binding sites from 5 well-characterized transcription factors, BCD, HB, NF-kB, YY1 and MEF-2. We restricted the analysis to only these factors because we needed a large number of binding sites to detect the associations of adjacent nucleotides. We constructed matrices for 1-site and 2-site prediction models.

We then compared the distributions of $L$-values from the two models and two of the factors, HB and MEF-2, showed significantly lower 1-site $L$-values when compared to the distribution of $L$-values produced by the 2-site model (Mann-Whitney: HB: $W = 10705, P = 0.0014$; MEF-2, $W = 8985, P = 0.0163$). This suggests that the information incorporated in the 2-site model describes more efficiently the sequence properties of the collected binding sites of these factors.

One of the caveats of this comparison is that by comparing the $L$-values from the 2-site and the 1-site model we may be comparing values that are not directly comparable since they are generated based on different models. In addition, sampling properties may make the 2-site model over-trained and it may only represent the properties of the given set of binding sites and not the properties of any given set of binding sites of the transcription factor. In order to bypass such a problem we generated a binding site set, with the same number of binding sites as the assembled one, solely sampling from the frequencies of the 1-site weight matrix. Therefore these binding site sets have the same frequencies with the real sets for the 1-site model (assuming independence of positions) but they are random in terms of the 2-site association. We constructed 2-site weight matrices for these simulated sets and obtained the $L$-values for the 2-site model. These values are directly comparable to the 2-site $L$-values from the real datasets. In 3 of the 5
cases (HB, MEF-2, YY1) the \( L \)-values for the real binding sites were significantly higher than the simulated sites (Mann-Whitney: HB, \( W = 8691, P = 0.049 \); MEF-2, \( W = 6956, P = 0.0007 \); YY1, \( W = 35381, P = 0.0111 \)). This result clearly demonstrates that there is real association between positions of binding sites and it can be efficiently incorporated in models for binding site prediction.

Pattern of binding site prediction in regulatory regions: The ultimate goal of this analysis is to generate improved models for binding site prediction that could be used to identify binding sites in real regulatory regions. The next step is to search regulatory regions that contain binding sites for transcription factors for which we have devised 2-site prediction models. In order to obtain a measurement of significance for the binding site prediction we followed the method described in Chapter 2. For each transcription factor we obtained a threshold \( L \)-value from a distribution of bootstrapped \( L \)-values (see methods).

In Fig. 4 we show the pattern of predicted NF-kB and HB binding sites (\( L \)-values above the threshold) for 2-site vs. 1-site model in two regulatory regions, \( TNF-\alpha \) in humans and \( eve \) stripe-2 enhancer in \( Drosophila \), respectively. Careful observation of the pattern of prediction in these two regions leads to two conclusions: a) most of the experimentally verified binding sites are correctly predicted with both models; b) fewer not yet experimentally verified binding sites were predicted with the 2-site model than the 1-site model and the ones that are predicted by both models have higher \( L \)-values with the 2-site model. These observations suggest that by incorporating the association of nucleotide positions (2-site model) within binding sites we have probably removed some
spurious or weak matches to the 1-site weight matrix and thus reduced the degree of false positives.

5.5 Discussion.

Transcription factor binding site prediction is one of the most useful tools for the annotation of regulatory sequences, especially when combined with sequence conservation measurements and expression profile comparisons (Ren et al. 2000; Bussemaker et al. 2001). Construction of models that incorporate many of the properties of binding site sequences will facilitate the identification of functional elements within non-coding regions. However, it has been shown that the degeneracy of sequence requirements for transcription factor binding does not allow very accurate construction of prediction models, thus resulting in many spurious matches. Identification of more complex properties of binding sites and their incorporation into models for prediction reduces the degree of false positives (Stormo 1998).

We have successfully identified and used an additional characteristic of transcription factor binding sites. We observed that there seem to be interactions between nucleotide positions of the binding site sequences in a manner that generates statistical association between the occurrences of nucleotides in two adjacent positions of the binding site. We demonstrated this association with chi-square tests and after incorporating this property in a binding site prediction model we showed that the efficiency of binding site prediction is improved.
This result has implications for the future of annotation of regulatory regions. By incorporating the association between sites we were able to more accurately describe the properties of binding sites and possibly reduce the degree of false positives. In fact, the higher the degree of association and the higher the degree of degeneracy as measured by the 1-site weight matrix the more efficient the prediction method will be for the proper identification of binding sites.

Moreover there are some additional aspects that one can test regarding the evolutionary dynamics of transcription factor binding sites. A model of association of nucleotide positions may reveal pairs of substitutions within the binding site sequence that when present together are not detrimental to the binding affinity of the sequence but each one by itself could be. Some of these changes may be predicted as too extreme when tested with a 1-site model but they may be very well compatible with a 2-site model if the two positions substituted strongly interact and as a pair provide the same binding efficiency (see Kirby et al. 1995 and Leigh et al. 1995 for compensatory evolution).

As we begin to understand more the structure and molecular interactions of transcription factors with their DNA targets we will be able to construct even more sophisticated models for the prediction of transcription factor binding sites and the annotation of regulatory sequences. Our results provide one more parameter that can improve binding site prediction that has to be considered in the multifactorial models required for the identification of transcription factor binding sites.
5.6 References.


<table>
<thead>
<tr>
<th>Transcription factor</th>
<th># sites collected</th>
<th>Binding site length (bps)</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCD</td>
<td>51</td>
<td>8</td>
<td>C.M. Bergman</td>
</tr>
<tr>
<td>HB</td>
<td>97</td>
<td>14</td>
<td>C.M. Bergman</td>
</tr>
<tr>
<td>MEF-2</td>
<td>90</td>
<td>7</td>
<td>SELEX database</td>
</tr>
<tr>
<td>NF-kB</td>
<td>42</td>
<td>10</td>
<td>SELEX database, TRANSFAC</td>
</tr>
<tr>
<td>YY1</td>
<td>181</td>
<td>6</td>
<td>SELEX database</td>
</tr>
</tbody>
</table>
**Figure 5.1**: Position weight matrices for the binding sites of the five transcription factors analyzed in this study assuming independence of positions within the binding site (1-site model). Stars between positions indicate statistically significant association of the two positions.
$L_{\text{allsite}} = L_1 + L_{t1} + L_{t2} + L_{t3} + L_{t4} + L_{t5} + L_{t6}$

**Figure 5.2**: Graphical description of the calculation of probability assuming adjacent position dependence.
Figure 5.3: Regression of the sum of chi-square statistic for the 50 simulated binding site sets with the mean difference in the $L$-value produced by the 2-site model and the $L$-value produced by the 1-site model. $R^2 = 50\%$, $P = 0.000$. 

$P = 0.0001$
Figure 5.4: Pattern of predicted NF-kB sites in the *TNF-α* promoter and HB sites in the *eve* stripe-2 enhancer using the 2-site and the 1-site models.
Figure 5.4: (continued)
APPENDIX A: Supplementary information for Chapter 1.

bcd proximal promoter

bcdp3-1
------TGCA CATTTTGTC AATTCAGGCT GGACTGGAAT GGAGCATATT AATATTATAA
bcdp3-2
-------- ........ ........ ........ ........ ........ ........ ........
bcdp3-14
-------- ........ ........ ........ ........ ........ ........ ........
bcdp3-31
-------- G------- ........ ........ ........ ........ ........ ........
bcdp3-32
-------- ........ ........ ........ ........ ........ ........ ........
bcdp3-46
-------- ------ ....... ........ ........ ........ ........ ........
bcdp3-47
-------- ........ ........ ........ ........ ........ ........ ........
bcdp3-54
-------- ---ATG... ........ .-........ ........ ........ ........ ........
bcdpH-12
-------- ........ ........ ........ ........ ........ ........ ........
bcd3-25
-------- ------- ........ ........ ........ ........ ........ ........
bcd3-51
ACAAAG.... ........ ........ ........ ........ ........ ........ ........
bcdh-16
-------- ........ ........ ........ ........ ........ ........ ........
bcdp-sim
-------- ------ ........ ........ ........ ........ ........ ........

bcdp3-1
TATTAACAAA AATCCTAAAT AAACATTCGA CACTTGTCTA ATTGATTCCT AAATTTGGGG
bcdp3-2
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-14
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-31
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-32
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-46
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-47
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-54
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-80
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-81
......... ........ ........ ........ ........ ........ ........ ........
bcdpH-12
......... ........ ........ ........ ........ ........ ........ ........
bcd3-25
......... A... ........ ........ ........ ........ ........ ........
bcd3-51
......... ........ ........ ........ ........ ........ ........ ........
bcdh-16
......... ........ ........ ........ ........ ........ ........ ........
bcdp-sim
......... TT.. ........ ........ ........ ........ ........ ........

bcdp3-1
TGCCT----G TTTGTTAATT AAATGTTAAT ATTATGAAGT TCCAAACAGA GCAAAGAGTT
bcdp3-2
......----. ....C..... ........ ........ ........ ........ ........ ........
bcdp3-14
......----. .......... .......... .......... .......... ........ ........
bcdp3-31
......----. .......... .......... .......... .......... ........ ........
bcdp3-32
......----. .......... .......... .......... .......... ........ ........
bcdp3-46
......----. .......... .......... .......... .......... ........ ........
bcdp3-47
......----. .......... .......... .......... .......... ........ ........
bcdp3-54
......----. .......... .......... .......... .......... ........ ........
bcdp3-80
......----. .......... .......... .......... .......... ........ ........
bcdp3-81
......----. .......... .......... .......... .......... ........ ........
bcdpH-12
......----. .......... .......... .......... .......... ........ ........
bcd3-25
......----. .......... .......... .......... .......... ........ ........
bcd3-51
.....----. .......... .......... .......... .......... ........ ........
bcdh-16
.....----. TTT. TGGG. ........ ........ ........ ........ ........ ........
bcdp-sim
.....----A ...T..T... G-----.... .......... ...G.G.... C.....G---

bcdp3-1
TAAGTTTAAT TGGTTCTACT TATTTGTTAC AATATTCAAG CTTTTTTTTA TTATTATTCT
bcdp3-2
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-14
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-31
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-32
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-46
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-47
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-54
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-80
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-81
......... ........ ........ ........ ........ ........ ........ ........
bcdpH-12
......... ........ ........ ........ ........ ........ ........ ........
bcd3-25
......... ........ ........ ........ ........ ........ ........ ........
bcd3-51
......... ........ ........ ........ ........ ........ ........ ........
bcdh-16
......... -A.... ..T.T.. ...G---- ........ ........ ........ ........ ........
bcdp-sim

bcdp3-1
CAATGCCAAA TCTCTACAAA TAAATACCC TCCGACGT TTAGAACAT TCACC-TTTT
bcdp3-2
......... ........ ........ ........ ........ ........ ........ ........
bcdp3-1   TGGCGCAACC GCCGCCAGAT CAAAACTTTT ACCATCATCC -GCTGCCCCA CACGCACACA

bcdp3-1   CATCCGCATC CGCACTCCCA TCCGCATCCG CACTCGCATC CGCACCCACA TCACCAA-CA
bcdp3-2   .......... .......... .......... .......... .......... .......... .......A-
bcdp3-14  .......... .......... .......... .......... .......... .......... .......A-
bcdp3-31  .......... .......... .......... .......... .......... .......... .......A-
bcdp3-54  .......... .......... .......... .......... .......... .......... .......A-
bcdp3-80  .......... .......... .......... .......... .......... .......... .......A-
bcdp3-81  .......... .......... .......... .......... .......... .......... .......A-
bcd3-51   .......... .......... .......... .......... .......... .......... .......A-

bcdp3-1   TCCGCAGCTT CAGTTGCCGC C-----
bcdp3-2   .......... .......... AACATA
bcdp3-14  .......... .......... AACATA
bcdp3-31  .......... .......... .A-----
bcdp3-32  .......... .......... .A-----
bcdp3-46  .......... .......... .A-----
bcdp3-47  .......... .......... .A-----
bcdp3-54  .......... .......... .A-----
bcdp3-80  .......... .......... .A-----
bcdp3-81  .......... .......... .A-----
bcdH-12   .......... .......... .A-----
bcd3-25   .......... .......... .A-----
bcd3-51   .......... .......... .A-----
bcuh-16   .......... .......... .A-----
bcdp-sim  .......... .......... .A-----

142
<table>
<thead>
<tr>
<th>ftzp-3-1</th>
<th>CAGCA-CCGT</th>
<th>CTCAAGTCGC</th>
<th>CGCAGTAGGA</th>
<th>GAAGCGCGCG</th>
<th>GGCAGATAAA</th>
<th>TCAG-TGTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ftzp-3-14</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-2</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-31</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-32</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-46</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-47</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-51</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-54</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-80</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-81</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-h16</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-w111</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-sim2</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ftzp-3-1</th>
<th>TAATGGGCGC</th>
<th>GATGGGTAGG</th>
<th>TAATAAGCCG</th>
<th>CGCAGCAGGT</th>
<th>AGGCACC-GT</th>
<th>ACGGATAAAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ftzp-3-14</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-2</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-31</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-32</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-46</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-47</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-51</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-54</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-80</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-81</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-h16</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-w111</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-sim2</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ftzp-3-1</th>
<th>TTGCCAGGAC</th>
<th>CTCCGGATAAC</th>
<th>TTCCCCTCTC</th>
<th>CGTGCCTGCA</th>
<th>AGGACATTTC</th>
<th>-GCCGGAGGG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ftzp-3-14</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-2</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-31</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-32</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-46</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-47</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-51</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-54</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-80</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-81</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-h16</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-w111</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-sim2</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ftzp-3-1</th>
<th>GTGGCTGCGA</th>
<th>ACAGCACCCG</th>
<th>TCCCCCTCTC</th>
<th>CGTGCCTGCA</th>
<th>AGGACATTTTC</th>
<th>-GCCGAGGG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ftzp-3-14</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-2</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-31</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-32</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-46</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-47</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-51</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-54</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-80</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-81</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-h16</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-w111</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-sim2</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ftzp-3-1</th>
<th>GCGAGCGTGT</th>
<th>GCCGA-----</th>
<th>----------</th>
<th>----------</th>
<th>----------</th>
<th>----------</th>
</tr>
</thead>
<tbody>
<tr>
<td>ftzp-3-14</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-2</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-31</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-32</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-46</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-47</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-51</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-54</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-80</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-81</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-h16</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-w111</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-sim2</td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
<td><del>-</del></td>
</tr>
<tr>
<td>ftzp-3-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>ftzp-3-31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ftzp-3-32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ftzp-3-46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ftzp-3-47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ftzp-3-51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ftzp-3-54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ftzp-3-80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ftzp-3-81 C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ftzp-h16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ftzp-w111</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ftzp-sim2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GGGCT CTCTGATT TT GTCTATATCA TGCTAGTGA C TCTG
hb distal promoter

```
hbp1-3-1 ACTATAATA ATGAATAATA ATAATAAGTTT TTTTTT-AG TCCAAAATTG GTCAATTAAC

hbp1-3-1 GATGCCAAA AACGCAAAG AGCCAAA-AA GGCAGCTGCA TTCGGCCAAA ATGCAGCGCC

hbp1-3-1 AGAAAATGCA AAAGGATAAA ATGAGCGAGT CAGAGCGAGA GAGTGGGTGA GTGAGTGAAA

hbp1-3-1 GA-GCGAACG CCACGAAGGG GATGCTGTTT AGTTATTGCT TT-TGGGGAT GGGGAAAGTC

hbp1-3-1 ACTCAGATTT ACAGCTAGCA TCCGTATCCG TTTTGAGTGA GTTTCGTCGT TACGTTGTTG

152
<table>
<thead>
<tr>
<th>153</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>hbp1-3-1</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-14</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-2</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-46</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-51</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-7</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-80</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-81</strong></td>
</tr>
<tr>
<td><strong>hbp1-sim</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-1</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-14</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-2</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-46</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-51</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-7</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-80</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-81</strong></td>
</tr>
<tr>
<td><strong>hbp1-sim</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-1</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-14</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-2</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-46</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-51</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-7</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-80</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-81</strong></td>
</tr>
<tr>
<td><strong>hbp1-sim</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-1</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-14</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-2</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-46</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-51</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-7</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-80</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-81</strong></td>
</tr>
<tr>
<td><strong>hbp1-sim</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-1</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-14</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-2</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-46</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-51</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-7</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-80</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-81</strong></td>
</tr>
<tr>
<td><strong>hbp1-sim</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-1</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-14</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-2</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-46</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-51</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-7</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-80</strong></td>
</tr>
<tr>
<td><strong>hbp1-3-81</strong></td>
</tr>
<tr>
<td><strong>hbp1-sim</strong></td>
</tr>
<tr>
<td>nos proximal promoter (partial alignment)</td>
</tr>
<tr>
<td>------------------------------------------</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nosp3-1</th>
<th>TGTGC-GTAA</th>
<th>TCC-GATGTT</th>
<th>G--CCGGCC</th>
<th>GATATCATG</th>
<th>GATGTGTGTA</th>
<th>CAGTGGGCCGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>nosp3-2</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-31</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-32</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-46</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-47</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-51</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-H12</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-35</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-43</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-54</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-80</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-81</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-h16</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-sim</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nosp3-1</th>
<th>-AATTCGAAT</th>
<th>GGTGGTGTGC</th>
<th>AGTGATTGTG</th>
<th>TTGTGACGGC</th>
<th>GAGTGGCGCG</th>
<th>TGTGGGTGCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>nosp3-2</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-31</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-32</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-46</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-47</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-51</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-H12</td>
<td>G--G--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-35</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-43</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-54</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-80</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-81</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-h16</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-sim</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nosp3-1</th>
<th>G-TAGTAC--</th>
<th>TATTTTC-CA</th>
<th>TTGCGCGGTG</th>
<th>TCCAGCTT-T</th>
<th>TAATTAGTGG</th>
<th>CACATATT-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>nosp3-2</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-31</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-32</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-46</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-47</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp3-51</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-H12</td>
<td>G--G--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-35</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-43</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-54</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-80</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-3-81</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-h16</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
<tr>
<td>nosp-sim</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nosp3-1</th>
<th>TT-AGCAAGT</th>
<th>AAAATAAATT</th>
<th>TTGCATACTA</th>
<th>TTA-AATTC</th>
<th>TTATAATTA</th>
<th>TTTCTAAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>nosp3-2</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
<td>..--.--</td>
</tr>
</tbody>
</table>

154
nosp3-31 ..-........-..-..-..-..............................
nosp3-32 ..-........-..-..-..-..-..............................
nosp3-46 ..-........-..-..-..-..-..............................
nosp3-47 ..-........-..-..-..-..-..............................
nosp3-51 ..-........-..-..-..-..-..............................
nosp-H12 ..-........-..-..-..-..-A................G........
nosp-3-35 ..-........-..-..-..-..-..............................
nosp-3-43 ..-........-..-..-..-..-..............................
nosp-3-54 ..-........-..-..-..-..-T..............................
nosp-3-80 ..-........-..-..-..-..-..............................
nosp-3-81 ..-........-..-..-..-..-..............................
nosp-h16 ..-........-..-..-..-..-..............................
nosp-sim ..-G........-..-..-..-GATGCCAC. C----------'

APPENDIX B: Supplementary information for Chapter 2

en interspecific alignment

enp-mel  ATATGTTCGG AATCCACAT CTATACACTA TTTGCAAGCAG ATGTTTCACA TTTGTAGATC
enp-sim2  -------------------- -------------------- -------------------- -------------------- -------------------- ----
enp-sech  -------------------- -------------------- -------------------- -------------------- -------------------- --------------------
enp-yak  -------------------- -------------------- -------------------- -------------------- -------------------- --------------------
enp-orena  -------------------- -------------------- -------------------- -------------------- -------------------- --------------------

enp-mel  GTGGTTGAT TAAAGGCCCA CAAAAGGCCA TTATTGTGTA TGTCAATGCA CAAATGGCGT
enp-sim2  ............... ............... ............... ............... ............... ............... ...........
enp-sech  ............... ............... ............... ............... ............... ............... ...........
enp-yak  ............... ............... ............... ............... ............... ............... ...........
enp-orena  ............... ............... ............... ............... ............... ............... ...........

enp-mel  TCGTTTTTAA CACCTTCTAA GGTTTATTTT GACAGTTAGC GATGGTTTTT CATTCGCTCG
enp-sim2  .......... .......C.. A....G.... .......... .........C ...........
enp-sech  .......... .......... A....G.... .......... .........C ...........
enp-yak  ............... ............... ............... ............... ............... ............... ...........
enp-orena  ............... ............... ............... ............... ............... ............... ...........

enp-mel  TCTGGCAAAT CGATATTCGA TGGGCAAATA AATAAATTAA ATGTCAATTA AATATCAATC
enp-yak  ............... ............... ............... ............... ............... ............... ...........

enp-mel  AATTTCGTCA GCTGTTTTTC AAGGCACATT TAACTGGTAA ATTGAAGGCC TCA---AAAA
enp-sim2  ....G..... .......... .......... .......... .......... ..C---....
enp-sech  ....G..... .......... .......... .......... .......... ..C---....

enp-mel  TAAATTGTCA TCGTGTATT TAGCGTATTT TTGTGAAAAA TCGCAAATGA TACTGGTGGT
enp-sim2  .......... .T-C...... .......... .......... .......... ..T..C....
enp-yak  .....C.... ..AC...... .......... .......... .........G .T.....A.C
enp-orena  A......... .T-CG..... .......... .......... .....C.... .T....C...

enp-mel  GGGAG----T GGGGGTGATA AAAAGTTTCC CT-TCATTTA AATTAAAAC- AATTAGCTTT
enp-sim2  .....----. .......... .......... ..-C...... .........- ..........
enp-sech  ..A..----. .......... .......... ..-CT..... .........- ..........
enp-yak  ...G.CAGA. .......... .......... ..TC...... T........T ..........
enp-orena  ...GA----. .......... .......... ..-C...... .........- ..........
	enp-mel  GCGGTTTT-- -TTTCCTCTC TTT--CGATT CCCCTCCTAC AGATCGCTTA AAGGTGTGAT
enp-sech  .A........ .......... ........T. ........T. .GTT...... ..G..T....
enp-yak  .......... .......... ........T. ........T. .GTT...... ..G..T....
enp-orena  .......... .......... ........T. ........T. .GTT...... ..G..T....
	enp-mel  TCAACAAAAA CTAAATGCTT ATGAAAAGTG TCTG-TGGAA AGAGATGGCA TGTGGCTCTC
	enp-mel  CCCCTCATGG AAAGGCAGCC ATTTTCCT-G GCCTACTCGC A-GAGGGAGT GAACA-GTGC
enp-sim2  .......... .......... ........-. .......... .-.T...... ..G..-....
enp-sech  .A........ .......... ........-. .......... .-.....T.. ..G..-....
enp-yak  .-........ .......... ........T. ........T. .GTT...... ..G..T....
enp-orena  .......... .......... ........-. .......... .-.T...... ..G..-....

156
ftz zebra element interspecific alignment

ftzp-mel  CGCAAACAGC CGACCTCGGG CCAAGGGTATAT TCCTGATCCCT AGGGATCGGA CGTAATGTTA
ftzp-sim2  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-sech  GGGAAAAGGC CACCTCGGG CCAAGGGTATAT TCCTGATCCCT AGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  TCTTTTGCCC GCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-sim2  GGGAAAAGGC CACCTCGGG CCAAGGGTATAT TCCTGATCCCT AGGGATCGGA CGTAATGTTA
ftzp-sech  GGGAAAAGGC CACCTCGGG CCAAGGGTATAT TCCTGATCCCT AGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  TCCCTTTGGCC GCCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-sim2  GGGAAAAGGC CACCTCGGG CCAAGGGTATAT TCCTGATCCCT AGGGATCGGA CGTAATGTTA
ftzp-sech  GGGAAAAGGC CACCTCGGG CCAAGGGTATAT TCCTGATCCCT AGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  ACAACTGGCA GCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-sim2  AACACTGGCA GCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-sech  AACACTGGCA GCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  GGGAACCATA AACGGGCCGG GGAAAA-AGC CTCTGCCCCG AAGGAACGTT TTCAGCAACA
ftzp-sim2  AACACTGGCA GCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-sech  AACACTGGCA GCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  GTTTACAGTT T-----TTA- ---------- -----TGTCT TTATGATTAT TGCAATTAGA
ftzp-sim2  AACACTGGCA GCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-sech  AACACTGGCA GCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  GGGAGA---- ------TCGG CTGAGAGTCG CGCCCTCTCG CTCTGCGCAC CTCATAGGTA
ftzp-sim2  TCCCTTGCCC GCCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-sech  TCCCTTGCCC GCCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  CGCGGGCGGA TAAATCGCGA TGATAATGGG CGCGATGGG- TAGGTAATAA GCC-GCGCAG
ftzp-sim2  TCCCTTGCCC GCCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-sech  TCCCTTGCCC GCCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  GGCACCTCAT GGCCGTAATT ACTGCAGCAC C-GTCTCAAG GTCGCCGAGT AGGAGAAGCG
ftzp-sim2  TCCCTTGCCC GCCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-sech  TCCCTTGCCC GCCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  CGCGGGCGGA TAAATCGCGA TGATAATGGG CGCGATGGG- TAGGTAATAA GCC-GCGCAG
ftzp-sim2  TCCCTTGCCC GCCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-sech  TCCCTTGCCC GCCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  GGCACCTCAT GGCCGTAATT ACTGCAGCAC C-GTCTCAAG GTCGCCGAGT AGGAGAAGCG
ftzp-sim2  TCCCTTGCCC GCCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-sech  TCCCTTGCCC GCCCAGTGCCC AGGAAATAAA TCGGAGGGGA AGGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  TCTCAAGGC CCA-GTCAGCA CCCTGCCCAC GCCTGCCGCT GCTGCCGAGT AGGGGATCGGA CGTAATGTTA
ftzp-sim2  CCA-GTCAGCA CCCTGCCCAC GCCTGCCGCT GCTGCCGAGT AGGGGATCGGA CGTAATGTTA
ftzp-sech  CCA-GTCAGCA CCCTGCCCAC GCCTGCCGCT GCTGCCGAGT AGGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  CAGGTAGGCA CC-GTACGGA TAAAGTTGCC AGGACCTCGG ATAACTTCCC CTCTCCGTGC
ftzp-sim2  CCA-GTCAGCA CCCTGCCCAC GCCTGCCGCT GCTGCCGAGT AGGGGATCGGA CGTAATGTTA
ftzp-sech  CCA-GTCAGCA CCCTGCCCAC GCCTGCCGCT GCTGCCGAGT AGGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  CAGGTAGGCA CC-GTACGGA TAAAGTTGCC AGGACCTCGG ATAACTTCCC CTCTCCGTGC
ftzp-sim2  CCA-GTCAGCA CCCTGCCCAC GCCTGCCGCT GCTGCCGAGT AGGGGATCGGA CGTAATGTTA
ftzp-sech  CCA-GTCAGCA CCCTGCCCAC GCCTGCCGCT GCTGCCGAGT AGGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  CTGCAAGGAC ATTTCGCCGG AGGGGTGGCT GCGAACAGCA GGCGGCAAAG TGTCATGCGC
ftzp-sim2  CCA-GTCAGCA CCCTGCCCAC GCCTGCCGCT GCTGCCGAGT AGGGGATCGGA CGTAATGTTA
ftzp-sech  CCA-GTCAGCA CCCTGCCCAC GCCTGCCGCT GCTGCCGAGT AGGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  CTGCAAGGAC ATTTCGCCGG AGGGGTGGCT GCGAACAGCA GGCGGCAAAG TGTCATGCGC
ftzp-sim2  CCA-GTCAGCA CCCTGCCCAC GCCTGCCGCT GCTGCCGAGT AGGGGATCGGA CGTAATGTTA
ftzp-sech  CCA-GTCAGCA CCCTGCCCAC GCCTGCCGCT GCTGCCGAGT AGGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  CAGGTAGGCA CC-GTACGGA TAAAGTTGCC AGGACCTCGG ATAACTTCCC CTCTCCGTGC
ftzp-sim2  CCA-GTCAGCA CCCTGCCCAC GCCTGCCGCT GCTGCCGAGT AGGGGATCGGA CGTAATGTTA
ftzp-sech  CCA-GTCAGCA CCCTGCCCAC GCCTGCCGCT GCTGCCGAGT AGGGGATCGGA CGTAATGTTA
ftzp-yak  --------------- --------------- --------------- --------------- --------------- ---------------
ftzp-orena  --------------- --------------- --------------- --------------- --------------- ---------------

ftzp-mel  CAGGTAGGCA CC-GTACGGA TAAAGTTGCC AGGACCTCGG ATAACTTCCC CTCTCCGTGC
ftzp-yak  .-....-.-. ...T...: .-...:. ..-...-. ---.-. ---.-. --.-.-. -
ftzp-orena.-....-.-. ...T...: .-..-. ..-.-. .-. ..-. G ..-.-.-.

ftzp-mel  CGCAGAGTTA GAGAAGAAAT
ftzp-sim2  ---.-.
ftzp-sech  ....-.
ftzp-yak  ---.-.
ftzp-orena  ---.-.
<table>
<thead>
<tr>
<th></th>
<th>mel-hbp</th>
<th>sech-hbp</th>
<th>ore-hbp</th>
<th>yak-hbp</th>
</tr>
</thead>
<tbody>
<tr>
<td>hb</td>
<td>TTTTATAT</td>
<td>TAACGTGCTT</td>
<td>ACTCCCTGCTG</td>
<td>TGCA---CTC</td>
</tr>
<tr>
<td></td>
<td>..........</td>
<td>..........</td>
<td>..........</td>
<td>.......</td>
</tr>
<tr>
<td>mel-hbp</td>
<td>CTGACCAACG</td>
<td>TAATCCCAT</td>
<td>AGAAAACCGG</td>
<td>TGAAAATTC</td>
</tr>
<tr>
<td>sech-hbp</td>
<td>.C........</td>
<td>........G.</td>
<td>........A.</td>
<td>GCAGCTGCCT</td>
</tr>
<tr>
<td>ore-hbp</td>
<td>.A.C.....C</td>
<td>..........</td>
<td>..........</td>
<td>GCAAGCTGGG</td>
</tr>
<tr>
<td>yak-hbp</td>
<td>.A.C.....T</td>
<td>..........</td>
<td>..........</td>
<td>GCAAGCTGGG</td>
</tr>
<tr>
<td>mel-hbp</td>
<td>CCATCCGCTA</td>
<td>AGCTCCCGGA</td>
<td>TCATCCAAAT</td>
<td>CTAATTTTTT</td>
</tr>
<tr>
<td>sech-hbp</td>
<td>..........</td>
<td>..........</td>
<td>..........</td>
<td>GTTCTGCTC</td>
</tr>
<tr>
<td>ore-hbp</td>
<td>..........</td>
<td>..........</td>
<td>..........</td>
<td>GCAAGCTGGG</td>
</tr>
<tr>
<td>yak-hbp</td>
<td>..........</td>
<td>..........</td>
<td>..........</td>
<td>GCAAGCTGGG</td>
</tr>
<tr>
<td>mel-hbp</td>
<td>TAATCCAGAA</td>
<td>TGGATCAAGA</td>
<td>GCACATGCTT</td>
<td>AACTAATGCC</td>
</tr>
<tr>
<td>sech-hbp</td>
<td>..........</td>
<td>..........</td>
<td>..........</td>
<td>TGTTGCAATT</td>
</tr>
<tr>
<td>ore-hbp</td>
<td>..........</td>
<td>..........</td>
<td>..........</td>
<td>GTTCTGCTC</td>
</tr>
<tr>
<td>yak-hbp</td>
<td>..........</td>
<td>..........</td>
<td>..........</td>
<td>GCAAGCTGGG</td>
</tr>
<tr>
<td>mel-hbp</td>
<td>A---CCGATC</td>
<td>CGCGACCTGT</td>
<td>ACCTGACTTC</td>
<td>TGCCCATCTA</td>
</tr>
<tr>
<td>sech-hbp</td>
<td>G--------</td>
<td>........G.</td>
<td>..........</td>
<td>ATCCCTGACT</td>
</tr>
<tr>
<td>ore-hbp</td>
<td>G--------</td>
<td>........G.</td>
<td>..........</td>
<td>ATCCCTGACT</td>
</tr>
<tr>
<td>yak-hbp</td>
<td>.TCC.....</td>
<td>..........</td>
<td>..........</td>
<td>ATCCCTGACT</td>
</tr>
<tr>
<td>mel-hbp</td>
<td>GCGTGCAATTC</td>
<td>GTCTACCTGA</td>
<td>GCGATATAA</td>
<td>ACTAATGCC</td>
</tr>
<tr>
<td>sech-hbp</td>
<td>.G.........</td>
<td>..........</td>
<td>..........</td>
<td>TGTTGCAATT</td>
</tr>
<tr>
<td>yak-hbp</td>
<td>.G.C......</td>
<td>..........</td>
<td>..........</td>
<td>TGTTGCAATT</td>
</tr>
</tbody>
</table>
knirps enhancer interspecific alignment

kn-mel   AAGACATAG CC-GAGATGC AGCCGTTTTA TCCCTTTTCC GCTTAAAAAA TTGCGGAT-G
kn-sim2  .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........
kn-sech  .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........
kn-yak   .-....GA.A.G.... .-....G..T.... .-....A.
kn-orena ..T-....AT G.C-....G- .-........ .-........ .-........ .-........

kn-mel   GGTGGATGGA TGGGTGAAT TCGGGTCTG CTTGTTTGGC TCGGGTTAT- AAACCAGG
kn-sim2  .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........
kn-sech  .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........
kn-yak   .-....GA.A.G.... .-....G..T.... .-....A.
kn-orena .-....T-.... ..A-.... .-........ .-........ .-........ .-........

kn-mel   CTGCTCGACA ATTTT-ATGG TTACGACCGA ACG----GCC TGACTTTATG A-CTCGCACG
kn-sim2  .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........
kn-sech  .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........
kn-yak   .-....GA.A.G.... .-....G..T.... .-....A.
kn-orena .-....T-.... ..A-.... .-........ .-........ .-........ .-........

kn-mel   GCCC-ACAAC GTGC-AAACG TGGCCCGAA- CGCAGGCTGA GTTTTTTAGG CCAATT-CTT
kn-sim2  .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........
kn-sech  .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........
kn-yak   .-....GA.A.G.... .-....G..T.... .-....A.
kn-orena .-....T-.... ..A-.... .-........ .-........ .-........ .-........

kn-mel   GTGGGTACCT AAGCCAGCGA TTT-CGTTAC CTAAT-CGCG GGATCAGCTT ACCTAAGCTG
kn-sim2  .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........
kn-sech  .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........
kn-yak   .-....TA..... .-....T........ .-....T........ .-....T........ .-....T........
kn-orena .-....TA-.... .-....A-.... .-....T........ .-....T........ .-....T........

kn-mel   CAGATTATCC T-AGGATCAT AATTATGA-C G
kn-sim2  .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........
kn-sech  .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........
kn-yak   .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........ .-........
kn-orena .-....T........ .-........ .-........ .-........ .-........ .-........ .-........ .-........

161
eve stripe 2 enhancer interspecific alignment

eve-2-mel AATATAACCC AATAATTGTA AGTAACTGCC AGG--AGCGA GG------- -------TA

eve-2-sim .C. -------------- C. -------------- C. -------------- C. --------------

eve-2-yak .C. .G. .C. -------- .CATCCTTG CATCCTTGCC.


eve-2-mel TCCTTCTGGA TTAC------ CCGGTACTGC ATAAAACATGG AACCAGAACC GTAACGTGGGA

eve-2-sim ----------------- ----------------- ----------------- -----------------

eve-2-yak ------.G.G .TGTT.TG.. TGTT.C. ------.C. -------------- GGCATTTGCAA

eve-2-ere -------.GTG --------.-------.-------.-------.-------.-------.-------

eve-2-mel CAGATCGAAA AGTGGGCCTG ---GTTTCTC GCTGTGTGTG CCGTGTTAAT CCGTTTGCAA

eve-2-sim ----------------- ----------------- ----------------- -----------------

eve-2-yak ------..G.G .TGGT.TG.. TGT.C. ------.C. -------------- GGCATTTGCAA

eve-2-ere -------.CTC --------.-------.-------.-------.-------.-------

eve-2-mel TCAGCGAGAT TATTAGTCAA TTGCAGTTGC AGC------- ---------- -GTTTCGCTT

eve-2-sim ---.C.... .......... ..-------.----------.----------.----------

eve-2-yak ---.C.... .......... ..-------.----------.----------.----------

eve-2-ere ---.CTC --------.-------.-------.-------.-------.-------

eve-2-mel TCGCAGTTTG GTAACACGCT GTGCCAT--- ---ACTTTC- ---------- ----------

eve-2-sim .....C.--- ---......- ---------- ---------- ---------- ----------

eve-2-yak .....C.--- ---......- ---------- ---------- ---------- ----------

eve-2-ere .....C.... .......... ..-------.----------.----------.----------

eve-2-mel ----ATTTAG ACAGAATCGA GGGACCTTG ---ACTATA ATCGACAAC CAGA------

eve-2-sim ------.G.------------.C.-------------A--------.-------.-------.-------

eve-2-yak ------.G.------------.C.-------------A--------.-------.-------.-------

eve-2-ere ------.GGG--------- ..-------.----------.----------.----------

eve-2-mel ------ACAAAACTT AA--------- .A.--------- .A.TACGCC. .ATCCA.C.

eve-2-sim ------.A.--------- .A.--------- .A.TACGCC. .ATCCA.C.

eve-2-yak ------.A.--------- .A.--------- .A.TACGCC. .ATCCA.C.

eve-2-ere ------.A.--------- .A.--------- .A.TACGCC. .ATCCA.C.

eve-2-mel CATCGCCATC TTCTGGCGCC GGTGTTGCTG GGTAGAGCCA AGGGCTTGAC

eve-2-sim -------.C.------- ..-------.----------.----------.----------

eve-2-yak -------.C.------- ..-------.----------.----------.----------

eve-2-ere -------.C.------- ..-------.----------.----------.----------

eve-2-mel ----ATTTAG ACAGAATCGA GGGACCTTG ---ACTATA ATCGACAAC CAGA------

eve-2-sim ------.G.------------.C.-------------A--------.-------.-------.-------

eve-2-yak ------.G.------------.C.-------------A--------.-------.-------.-------

eve-2-ere ------.GGG--------- ..-------.----------.----------.----------

eve-2-mel ------ACAAAACTT AA--------- .A.--------- .A.TACGCC. .ATCCA.C.

eve-2-sim ------.A.--------- .A.--------- .A.TACGCC. .ATCCA.C.

eve-2-yak ------.A.--------- .A.--------- .A.TACGCC. .ATCCA.C.

eve-2-ere ------.A.--------- .A.--------- .A.TACGCC. .ATCCA.C.

eve-2-mel CATCGCCATC TTCTGGCGCC GGTGTTGCTG GGTAGAGCCA AGGGCTTGAC

eve-2-sim -------.C.------- ..-------.----------.----------.----------

eve-2-yak -------.C.------- ..-------.----------.----------.----------

eve-2-ere -------.C.------- ..-------.----------.----------.----------

eve-2-mel ----ATTTAG ACAGAATCGA GGGACCTTG ---ACTATA ATCGACAAC CAGA------

eve-2-sim ------.G.------------.C.-------------A--------.-------.-------.-------

eve-2-yak ------.G.------------.C.-------------A--------.-------.-------.-------
BCD weight matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>0.20</th>
<th>0.22</th>
<th>0.94</th>
<th>0.98</th>
<th>0.00</th>
<th>0.02</th>
<th>0.08</th>
<th>0.08</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.35</td>
<td>0.08</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
<td>0.04</td>
<td>0.86</td>
<td>0.49</td>
<td>0.31</td>
</tr>
<tr>
<td>G</td>
<td>0.10</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
<td>0.24</td>
<td>0.04</td>
<td>0.08</td>
<td>0.08</td>
<td>0.41</td>
</tr>
<tr>
<td>T</td>
<td>0.35</td>
<td>0.69</td>
<td>0.02</td>
<td>0.00</td>
<td>0.73</td>
<td>0.08</td>
<td>0.33</td>
<td>0.18</td>
<td>0.33</td>
</tr>
</tbody>
</table>

HB weight matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>0.14</th>
<th>0.33</th>
<th>0.55</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.40</td>
<td>0.25</td>
<td>0.25</td>
<td>0.09</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.04</td>
<td>0.20</td>
<td>0.16</td>
</tr>
<tr>
<td>G</td>
<td>0.19</td>
<td>0.16</td>
<td>0.31</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.32</td>
<td>0.13</td>
<td>0.48</td>
</tr>
<tr>
<td>T</td>
<td>0.27</td>
<td>0.26</td>
<td>0.12</td>
<td>0.69</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.08</td>
<td>0.22</td>
<td>0.31</td>
<td>0.31</td>
</tr>
</tbody>
</table>

BCD binding sites

Bolded sequence indicates the experimentally verified binding sites used to construct the weight matrix

>AE003831; 129914 - 129962; BCD_EVE S2E 5+; REF. 1
TTTCTGCGCTGTCGCCGCTGTTAATCCGTTTGCCATCAGCGAGATTAT

>AE003831; 129933 - 129936; BCD_EVE S2E 4+; REF. 1
CGCTGCACTGCAATTGACTAATACTCCTCGCTGAGGCAACCGGAT

>AE003831; 130083 - 130036; BCD_EVE S2E 3+; REF. 1
GGAATCGAGGGACCCTGGACTATAATCGCACAACGAGACCGGGTTAG

>AE003831; 130227 - 130180; BCD_EVE S2E 2+; REF. 1
GATTTCAACTGCAAGCCCCTGGCTATATCCCAAGAACAAACAAACAC

>AE003831; 130348 - 130300; BCD_EVE S2E 1+; REF. 1
GTTGCCCTGACCTGCGGCCCTTACATCCTCGACATCATATTATG

>AE003680; 34016 - 33968; BCD_HB A1+; REF. 2
TGCTGTCGACCTCCTGGACCAACGTAAAATCCTCCATAGAAAACCGGTTGAAA

>AE003680; 33906 - 33858; BCD_HB A2+; REF. 2
GCATATTATTTTCTTGTCTCATAATCCAGAATGGATCAAGAGCGCAAT

>AE003680; 33790 - 33753; BCD_HB A3+; REF. 2
CTTCCGTTACCTCTGCCCATCTATACCTCTCGCGTCACTCCGCT

>AE003680; 36067 - 36019; BCD_HB B1+; REF. 2
ATTTTACTTTTATTAATCTATGCTAATCTGACTGATAACCAATTTCG

>AE003680; 36017 - 35969; BCD_HB B2+; REF. 2
TATTCTATGTCATAATCACCCTTTAAATCCCAAGTACTCAACTTTCTTCT

>AE003680; 33900 - 33938; BCD_HB X1+; REF. 2
AACCGGTTGAAAAATCCGCAGCTCTCGCTGTAAGCTCCGGACTCCG

>AE003680; 33949 - 33922; BCD_HB X2+; REF. 2
GCTGGCCATCCGCTAAGCTTCCGGATCA

>AE003680; 33932 - 33892; BCD_HB X3+; REF. 2
CTTCCGTTACATCTCAAATCAGGCCAATTTTTTGT

>AE003680; 33880 - 33819; BCD_HB X+; REF. 3
CCAGAATGCAGACGCGTCTACATCTCTGCGATCGTCGATCGTCCAGCGAT

>AE003466; 247497 - 247450; BCD_KR 730 1+; REF. 4
GGATTATAATGCTGAAAAATATCCGTGGTTCTAGGATTAGCCT

>AE003466; 247497 - 247450; BCD_KR 730 1+; REF. 4
GGATTATTTGCTGGAAAAATATCCGTGGTTCTAGGATTAGCCT
GGATTAAATTTTTTCAGACAAATAATCCAGCCTTAAGCATGGTG

>AE003466; 247539 - 247492; BCD_KR 730 3-; REF. 4

CCCTTGGTAGGGATCAAGCTTAATCACCATGCTTAAGGCTGGATTA

>AE003466; 247541 - 247574; BCD_KR 730 4+; REF. 4

GTAATTGAGGATTTTCCTTTTTTCTAATATCGTCTGT

>AE003466; 247570 - 247601; BCD_KR 730 5+; REF. 4

CCGTCTGTTAATCTCCGGCTTAGAGCGCGACGCGT

>AE003466; 247872 - 247932; BCD_KR 730 6+; REF. 4

ACAATGCAAAAAATATGTGTAACTGAACTAAATCCGGCTTAGGATTCTTGCGTCATAAACG

>AE003592; 68569 - 68539; BCD_KNI 64 1+; 5

GTGGTACCTAAGCAGGATTTTCCTTAAATCCGTCTGT

>AE003592; 68534 - 68566; BCD_KNI 64 2-; REF. 5

GCGATTAGGTAACGAAATCGCTGGCTTAGGTAC

>AE003592; 68534 - 68566; BCD_KNI 64 2-; REF. 5

GTCCTCAAAAGGCAACGCCTAATCTGGCTCAGCGGTAAG

>AE003775; 180502 - 180560; BCD_TLL 7+; REF. 6

AAATACTAAAAACTGCTGAAAATCTTTTAATCTTGATAGAATCAAAAAAAAAAAAAGAT

>AE003554; 48929 - 48979; BCD_H S7E 5+; REF. 7

TATGTTTTTAAGTCGAAAAGCCAAAAAGTCAAATGCGATTTTTTATGGGAACAAGAC

>AE003554; 49897 - 49897; BCD_H S7E 6+; REF. 7

GGGTTTTACGACCTCCGTCCGTTTTTTAAGCCTTTCTGCTCTGCCATCCTCTGACCC

>AE003554; 50671 - 50612; BCD_H S6E 2-; REF. 8

TTAACACCAAGGTTGCGGGCTGTTTTTTAAGCCTTTTGCCTAATCAAAAAATGTGGAAAT

>AE003554; 50713 - 50765; BCD_H S6E 3-; REF. 8

TATGTTTTTAAGTCGAAAAGCCAAAAAGTCAAATGCGATTTTTTATGGGAACAAGAC

>AE003554; 50765 - 50765; BCD_H S6E 3-; REF. 8
>AE003632; 142053 - 142104; BCD_SPALT 1-; REF. 9
CTATTTTTTCGGTCATTGCAAGCCGTTTTTCAGGGGCGTTACCGAAGTCGA

>AE003632; 141836 - 141785; BCD_SPALT 2+; REF. 9
AAGGAGCGAGGATCGATCTTCGATAAGCCGGAGGAAAATGCAACGCCCACGG

>AE003632; 141806 - 141754; BCD_SPALT 3+; REF. 9
GAGGAAAATGCAACGCCACGGACAAATCCTTTTGCGCAACAAAAAGGGTAATGG

>AE003632; 141772 - 141724; BCD_SPALT 4+; REF. 9
GCAACAAAGGTAATGGCTGCAAATCGGCATATGCAATCC

>AE003632; 141772 - 141724; BCD_SPALT 5+; REF. 9
GCAACAAAGGTAATGGCTGCAAATCGGCATATGCAATCC

>AE003632; 141668 - 141720; BCD_SPALT 6-; REF. 9
AGGCAAATATTTTGCGTACTAATTAAGCATGGCTCAAGACCCGGAATCGT

>AE003632; 141626 - 141590; BCD_SPALT 7+; REF. 9
GGCTGCTGCCCTGGCGTAATTATAATCCCTTCGATCG

Hunchback binding sites

Bolded sequence indicates the experimentally verified binding sites used to construct the weight matrix

>AE003825; 59502 - 59552; HB_EN 1-; REF. 1
GATATTTAATTCACATTTATTTATTTAGGCTCATCAGATCTGTTGG
>AE003825; 59502 - 59454; HB_EN 1+; REF. 1
CAATCAATTTCGTCAGCTGTTTTTCAAGGCAACATTTAATGTTGG
>AE003715; 11564 - 11610; HB_ABD-B IAB-2 1-; REF. 2
TTCTAGTTTCCGTCATTTTTATGAGACCTGACAGTGGCGCAF
>AE003715; 11507 - 11450; HB_ABD-B IAB-2 2+; REF. 2
ATAAATACGTAAAATGCCCCTTTTTATTTTGTCTGTGTAACATTACGT
>AE003715; 11399 - 11449; HB_ABD-B IAB-2 3-; REF. 2
AAGATTTCATGTCACGCAATTTATTTATGTTTGAAGTGTGCACATTTC
>AE003715; 11173 - 11226; HB_ABD-B IAB-2 4-; REF. 2
CATAATATCCTATGCTACCTGTGCCGAATTTTCGCCGAGCGGTGCAGAGCTCC
>AE003715; 11142 - 11192; HB_ABD-B IAB-2 5-; REF. 2
TCCTGTCTAGTTTTTTTCTTTTCTTTATCTAATTTCCATCTGCTGCTGA
>AE003714; 133459 - 133417; HB_UBX BRE A; 3
TTAGGAAACGTTTTTATGTCTGTCGACACACATGAG
>AE003714; 133376 - 133438; HB_UBX BRE A; 3
TTAATGATCACCATCGGCTCTGGAACGATTTTTAATGTTTCTCATGTGCTGCTAGCGCAC
>AE003714; 133260 - 133289; HB_UBX BRE C1; 3
TAACATTGTACTTTTATGACCTCGFAAA
AATAGTATCATTCGTGTTAGACCAATTTTTTTCCCAAGCGGATGTTCGCTTTT
gtacat>AE003466; 247394 - 247421; HB_KR 730 2+; REF. 7
AAACCAATTTGCTACAATTTATTTTTTTGCTTTTCCTTTTAAGCATG
>AE003466; 247852 - 247822; HB_KR 730 8-; REF. 7
GTTGTTTTTAGAGAAATTTGACTTCTTAAT
gatagtagcatgttttttttttttttggcgttttcttgggcggagaaattctgt
gatagttcccgtattgggtgggtggtggtgtt
>AE003554; 48264 - 48324; HB_H S3/4E 9+; REF. 8
AAAAATCTGCGAGCAGGTCGCTCTTTTTTTTGTTTGTTCATC
gcttcgttcacagtcgactatgtgttccttcgcccttgagaaaat
gtacat>AE003554; 48001 - 48048; HB_H S3/4E 4+; REF. 8
ATTTCTTACTGTAAGGACCCAGACATGCATGTCGATTAAGTT
gggttaagtaagctgtaagattgagctgttttctgttggtgttttgagcag
>AE003554; 48040 - 48083; HB_H S3/4E 5+; REF. 8
GGAGAGCTCATATGTTTTTTTTTTTTTTCCTCGGACTCCGATTTTAGATTTTG
>AE003554; 47926 - 47985; HB_H S3/4E 3-; REF. 8
ATGGATCATCCAGGACATGTAACATTTTTTTTCTTCCTTCGCCACTCATTAGGATTTCAG
>AE003554; 47917 - 47957; HB_H S3/4E 1-; REF. 8
GGGAAAAAGTCCGATATGCGACGAGATTTTTGCGTAATTTCTCAAGGCCAGACAGGGACA
gggttaagtaagctgtaagattgagctgttttctgttggtgttttgagcag
>AE003554; 48001 - 48048; HB_H S3/4E 4+; REF. 8
ATTTCTTACTGTAAGGACCCAGACATGCATGTCGATTAAGTT
gggttaagtaagctgtaagattgagctgttttctgttggtgttttgagcag
>AE003554; 48040 - 48083; HB_H S3/4E 5+; REF. 8
GGAGAGCTCATATGTTTTTTTTTTTTTTCCTCGGACTCCGATTTTAGATTTTG
>AE003554; 47926 - 47985; HB_H S3/4E 3-; REF. 8
ATGGATCATCCAGGACATGTAACATTTTTTTTCTTCCTTCGCCACTCATTAGGATTTCAG
>AE003554; 47917 - 47957; HB_H S3/4E 1-; REF. 8
GGGAAAAAGTCCGATATGCGACGAGATTTTTT

TATTAATCTTTTTTTTTTTTTTTGATTCTATCAAGATTAAAAGATTTTC
ATTAGTCTTGCTTTGCTTTGAGGCAACTTTTTGGGAATTTGTTTTGAGCAACGCAT
GGAAGAACTGACTCAGAGTTTTTGTGTTTTTTCTGCTTTGTTAAAAATTAGTCTTGGCTTT
GGAAGTAATGCTTGCTTTGCTTTGAGGCAACTTTTTGGGAATTTGTTTTGAGCAACGCAT
ACTTTTTTATTGCGCGCATGGCCGATGCTCTACCTTTTTATTGCGCGCATGGTGCG
CAATGATGTCCTGCGATGCGCTCTACTTTTTTATTGGCCGGCATTGTGCG
TCGACCCGGATGCCAGTTTTTTACGATCTCATCTTCGTTTTTTACTTTTTTTTACGACCT
CTCAACCGGTTTTTACGACCTCCGTCCTTTTTAAAAAGCTTCTGCTGCTGCAATCCTCTCTGA
GGGTGAAATCCCGCTGAAACAAATGAGTTTTTGGACCATCGGGAAGTGCTGAGTTTTTTAGGCCAATTCTTGTGGGTACCTAAGCCAGCGA
TCCACCCATCCCAACCATCCACCCATCCGCTACCTCTGGCTTTTTAAAAAGCGGCAAACGCATTGCTGAC
CCTGACGTTTGGCGTTTTTGAAGCCTTTTTTTTTTTTATGTCGTTTTTTTGTTTTTTTACG
TGTTTTTTTATTATTTTTTTGTTTTTTTATGTCGTTTTTTTACG
GGGTGAAATCCCGCTGAAACAAATGAGTTTTTGGACCATCGGGAAGTGCTGAGTTTTTTAGGCCAATTCTTGTGGGTACCTAAGCCAGCGA
CCTGACGTTTGGCGTTTTTGAAGCCTTTTTTTTTTTTATGTCGTTTTTTTACG
TGTTTTTTTATTATTTTTTTTGGATTCTATCAAGATTAAAAGATTTTC
ATTAGTCTTGCTTTGCTTTGAGGCAACTTTTTGGGAATTTGTTTTGAGCAACGCAT
GGAAGAACTGACTCAGAGTTTTTGTGTTTTTTCTGCTTTGTTAAAAATTAGTCTTGGCTTT
GGAAGTAATGCTTGCTTTGCTTTGAGGCAACTTTTTGGGAATTTGTTTTGAGCAACGCAT
ACTTTTTTATTGCGCGCATGGCCGATGCTCTACCTTTTTATTGCGCGCATGGTGCG
CAATGATGTCCTGCGATGCGCTCTACTTTTTTATTGGCCGGCATTGTGCG
APPENDIX C: Supplementary information for Chapter 3.

Accession numbers:

Accession numbers for functional data in humans:

The Genbank accession numbers are (following the order of species in Table 1): acute reg. protein: U29098, AY007224; adenosine deaminase: M13792, M34242; adh1: M32656, L30115, AF052976, X76984; aldolase A: X12447, J05517; apoA-I: J04066, X64263; apoA-II: X02905, M32360; apoCIII: M60674, X77900, L04149; apoE: AF261279, AF261280, D00466; app: D87675, AF067971; atrial natriuretic factor: K02043, K02781; C/EBP: S75625, M62362; ccr5: AF246924, AF252556, AF252560, AF252561, AF252567, AF252594, U68565; cftr: M58478, X95928, X95929, X95930, L04873; c-reactive protein: M11880, X13588; erythropoietin: X02158, M12930; factor IX: X55008, X65472, X54634, X54633; _interferon: M32746, X64659; growth hormone: J03071, U02293, AJ297563, Z46663; haptoglobin: AB025321, U01948, U04852; huntingtin: Y07981, Y07989, Y07988, L34008; il-3: L10616, X74875, X51890, X74877, AF147878; il-5: J03478, X06271; monoamine oxidase: alignment silver database (http://sayer.lab.nig.ac.jp/~silver/); msh2: U23824, AJ002051, AJ002050, AJ002052, AJ002049, AJ002053; mucin 1: X69118, L41624, U16175; myoglobin: X00371, X04405; neurofilament M: primate alignment from silver database (http://sayer.lab.nig.ac.jp/~silver/), mouse: X05640; oxytocin: M11186, M88355; platelet glycoprotein IB: M22403, U91967; preproinsulin: M10039, X61089, X61092, J02989, X04725; rh50: AF237382, AF177627, AF177628, AF177629, AF177630, AF177632,
AF177632, AF057527; *slow cardiac troponin C*: M37984, J04971; *sry*: L08063, AJ222687, AJ003068, AF008916, AF070933; *TNF*: from ref. 4, mouse: U68414.

Accession numbers for functional data in both human and rodents:

*acid labile subunit*: af192554, af220294; *c/ebpa*: s75265, m62362; *cd68*: af060540, ab009287; *c-myb*: z18893, z18894; *colla1*: af017178, x54876; *cyp1a1*: d10855, af210905; *dio2*: af188709, af195884; *gnrh*: u56735, u29674; *grp78*: x59969, j03377; *haptoglobin*: m34230; *l-plastin*: s54531, u82611; *mucin2*: u68061, af221746; *myoD*: u21227, s82831; *omp*: u01212, u01213; *plat*: k03021, m26065, m31184; *glucagon*: x03991, z35161; *p-selectin*: l01574, af031662; *surf 1/2*: ac002107, m14689

*thyroglobulin*: ab052952, x06162

References used for binding site characterization and mapping:

Genes characterized in humans only:

Acute reg. protein:

Sugawara T, Saito M, Fujimoto S.
Sp1 and SF-1 interact and cooperate in the regulation of human steroidogenic acute regulatory protein gene expression.

Sugawara T, Kiriakidou M, McAllister JM, Kallen CB, Strauss JF 3rd.
Multiple steroidogenic factor 1 binding elements in the human steroidogenic acute regulatory protein gene 5’-flanking region are required for maximal promoter activity and cyclic AMP responsiveness.

Adh1: GOOO184 (TRANSFAC)

ApoCIII:

Hadzopoulou-Cladaras M, Lavrentiadou SN, Zannis VI, Kardassis D.

Talianidis I, Tambakaki A, Toursounova J, Zannis VI. Complex interactions between SP1 bound to multiple distal regulatory sites and HNF-4 bound to the proximal promoter lead to transcriptional activation of liver-specific human APOCIII gene. Biochemistry. (1995) 34: 10298-309

ApoE: G000207 (TRANSFAC)


CCR5:


Cftr and Mucin1:


Factor IX:


Interferon-gamma: GOO1174 (TRANSFAC)

Growth hormone: GOOO282(TR)

Haptoglobin: G000285

Huntigtin:


Il-3: G000314

Monoamine oxidase:


Msh2:


Neurofilament M:


Preproinsulin: G000320

Rh50:


Sry:

TNFa:


aldolase A:


adenosine deaminase:


platelet glycoprotein Ib alpha:


oxytocin:


interleukin-5:

Of the GATA-binding proteins, only GATA-4 selectively regulates the human interleukin-5 gene promoter in interleukin-5-producing cells which express multiple GATA-binding proteins.

atrial natriuretic factor:

Wu J, LaPointe MC, West BL, Gardner DG.
Tissue-specific determinants of human atrial natriuretic factor gene expression in cardiac tissue.

cardiac troponin C: G001178
c-reactive protein: G000233
myoglobin: G000348
apolipoprotein A-I: G000203
apolipoprotein A-II: GOOO204
erythropoietin:

Tsuchiya T, Ueda M, Ochiai H, Imajoh-Ohmi S, Kanegasaki S.
Erythropoietin 5’-flanking sequence-binding protein induced during hypoxia and cobalt exposure.

Genes characterized in both humans and rodents:

olfactory marker protein:

Buiakova OI, Krishna NS, Getchell TV, Margolis FL. Human and rodent OMP genes: conservation of structural and regulatory motifs and cellular localization.
Genomics 1994 Apr;20(3):452-62

Tumor necrosis factor aplha:

Kuprash DV, Udalova IA, Turetskaya RL, Kwiatkowski D, Rice NR, Nedospasov SA.
Similarities and differences between human and murine TNF promoters in their response to lipopolysaccharide.

coll1a1:

Peterkofsky B, Gosiewska A, Singh K, Pearlman S, Mahmoodian F.
Species differences in cis-elements of the proalpha1(I) procollagen promoter and their binding proteins.

l-plastin:

Lin CS, Chang CH, Huynh T.
The murine L-plastin gene promoter: identification and comparison with the human L-plastin gene promoter.

glucagon:

Nian M, Drucker DJ, Irwin D.
Divergent regulation of human and rat proglucagon gene promoters in vivo.

mucin 2:

Aslam F, Palumbo L, Augenlicht LH, Velcich A.
The Sp family of transcription factors in the regulation of the human and mouse MUC2 gene promoters.

grp78:

Resendez E Jr, Wooden SK, Lee AS.
Identification of highly conserved regulatory domains and protein-binding sites in the promoters of the rat and human genes encoding the stress-inducible 78-kilodalton glucose-regulated protein.

gnrh:

Structure of the distal human gonadotropin releasing hormone (hGnrh) gene promoter and functional analysis in Gt1-7 neuronal cells.
dio2:

Gereben B, Salvatore D, Harney JW, Tu HM, Larsen PR.
The human, but not rat, dio2 gene is stimulated by thyroid transcription factor-1 (TTF-1).

cyp1a1:

Boucher PD, Hines RN.
In vitro binding and functional studies comparing the human CYP1A1 negative regulatory element with the orthologous sequences from rodent genes.

c-myb:

Seib T, Welter C, Engel M, Theisinger B, Dooley S.
Presence of regulatory sequences within intron 4 of human and murine c-myb genes.

cd68:

Greaves DR, Quinn CM, Seldin MF, Gordon S.
Functional comparison of the murine macrosialin and human CD68 promoters in macrophage and nonmacrophage cell lines.

surf 1/2:

Lennard A, Gaston K, Fried M.
The Surf-1 and Surf-2 genes and their essential bidirectional promoter elements are conserved between mouse and human.

myoD:

Embryonic activation of the myoD gene is regulated by a highly conserved distal control element.

plat:

Holmberg M, Leonardsson G, Ny T.
The species-specific differences in the cAMP regulation of the tissue-type plasminogen activator gene between rat, mouse and human is caused by a one-nucleotide substitution in the cAMP-responsive element of the promoters.
thyroglobulin:


haptoglobin:


acid-labile subunit:


p-selectin:

CURRICULUM VITAE

Emmanouil T. Dermitzakis

EDUCATION
1994: University of Crete, Greece: B.Sc. in Biology (Grade: Very Good)
1998: University of Crete: M.Sc. in Biology (Grade: Excellent, Advisor: E. Zouros)
1994: Centro de Biologia Molecular, Madrid, Spain: Erasmus undergraduate fellowship (6 months)

PUBLICATIONS
Manuscripts Published or In press:


Manuscripts submitted:


Dermitzakis, E. T., C. M. Bergman and A. G. Clark. High evolutionary potential in regulatory sequences of early development genes in Drosophila. Submitted to Genome Research