IDENTIFICATION AND CHARACTERIZATION OF CONSERVED REGULATORY ELEMENTS BY COMPARATIVE GENOMICS

KRISH JON MATHAVAN

(B.Sc. (Hons.) University of New South Wales)

A THESIS SUBMITTED FOR THE

DEGREE OF DOCTOR OF PHILISOPHY

INSTITUTE OF MOLECULAR AND CELL BIOLOGY

NATIONAL UNIVERSITY OF SINGAPORE

2008

Acknowledgements

I would like to thank firstly my supervisor Byrappa Venkatesh, especially for the patience and support shown to me during the writing of this thesis.

I would also like to thank the past and present members of the SB and FUGE lab for the friendship and help with techniques and reagents, especially Tay Boon Hui, Sumanti Tohari, Elizabeth Yeoh and Diane Tan.

I would also like to thank Jian Liang from Walter’s lab; and Guo Ke, Li Jie and Bin Qi from the histology lab who taught me histology and provided much expertise in helping me fine-tune the various techniques involved, and who went out of their way to help whenever possible. I would also like to thank Arun from BRC who helped to make the transgenic work run more smoothly for me.

I would like to thank members of my supervisory committee: Walter Hunziker and Wang Yue for the feedback given during the development of this project.

Finally I would like to thank my loved one and friends both here and in Australia, who have been supporting me during the whole doctorate, and who have kept me strong when I was disheartened and who encouraged me through the thesis.

TABLE OF CONTENTS

Acknowledgements……………………………………………………………..ii

Table of Contents……………………………………………………………..…iii

Summary…………………………………………………………………vii

List of Tables…………………………………………………………….ix

List of Figures……………………………………………………………..x

List of Abbreviations……………………………………………………xii

Chapter 1 Introduction…………………………………………………….1

1.1 Functional sequences in the human genome…………………………………..2

1.2 *Cis*-regulatory elements……………………………………………………….3

1.3 *Cis*-regulatory elements and genetic diseases…………………………………5

1.4 Identification of *cis*-regulatory elements……………………………………...7

1.4.1 Traditional methods…………………………………………...……..8

1.4.2 High throughput methods……………………………………….10

1.5 Using comparative genomics to identify *cis*-regulatory elements…………...12

1.5.1 Comparison of closely related species……………..…………….13

1.5.2 Extreme conservation within mammals………………………….16

1.5.3 Comparison of distantly related vertebrates……………………..18

1.5.4 Alignment and visualization tools for comparative genomics…..24

1.6 Objectives of the present study……………………………………………27

Chapter 2 Materials and methods……………………………………………......32

2.1 Genomic sequence alignment and prediction of conserved noncoding sequences……………………………………………………………………...33

2.2 Generation of DNA constructs for microinjection…………………………..35

2.3 Isolation and sequencing of fugu cosmid to map the orexin locus………….36

2.4 Generation of transgenic mice……………………………………………….37

2.5 Preparation of DNA for microinjection……………………………………...38

2.6 Genotyping…………………………………………………………………...39

2.7 In situ hybridization………………………………………………………….41

2.7.1 Preparation of embryos and tissues for whole-mount or section in situ hybridization……………………………………………………...41

2.7.2 Synthesis of RNA probes for in situ hybridization………………..43

2.7.3 Pretreatment of embryos and sections……………………………..44

2.7.4 Hybridization, washing and antibody addition…………………….46

2.7.5 Visualization………………………………………………………47

2.7.6 Double in situ hybridization………………………………………..49

Chapter 3 Results: Identification of CNEs in forebrain genes………………………50

3.1 Introduction…………………………………………………………………..51

3.2 Identification of human, mouse and fugu forebrain genes…………………..52

3.3 Prediction of CNEs…………………………………………………………..52

3.4 Summary……………………………………………………………………..58

Chapter 4 Results: Regulation of *Six3*……………………………………………...60

4.1 Introduction…………………………………………………………………..61

4.2 *Six3* loci in human, mouse and fugu; and identification of CNEs…………...62

4.3 Expression pattern of mouse *Six3*……………………………………………67

4.4 Functional assay of *Six3* CNEs………………………………………………70

4.4.1 Basal promoter region (includes CNE13) of mouse *Six3* is sufficient to recapitulate most aspects of expression in the forebrain and eye during early and late stages of development…….………………………………70

4.4.2 Expression patterns directed by CNE1, CNE2/3/4 and CNE5/6/7...74

4.4.3 Expression patterns directed by CNE8/9 and CNE12…….………76

4.4.4 CNE10/11 silences the mouse *Six3* promoter at all developmental stages….…………..…………………………………………………...81

4.4.5 Expression pattern directed by CNE14……...……………………..81

4.4.6 Summary of the regulatory potential of mouse *Six3* CNEs………..82

4.5 Discussion………………………………………………………………83

4.5.1 Comparison of results from *Six3* regulation in medaka……………86

Chapter 5 Results: Regulation of *Foxb1*………………………………………………90

5.1 Introduction……………………………………………………………..91

5.2 Comparison of *Foxb1* loci in human, mouse and fugu………………………92

5.3 Expression pattern of mouse *Foxb1*…………………………………………….96

5.4 Functional assay of *Foxb1* CNEs…………………………………………….99

5.4.1 Basal promoter region (includes CNE3) of mouse *Foxb1* is sufficient to recapitulate most aspects of endogenous expression during early and late stages of development……………………………………………….99

5.4.2 Expression patterns directed by CNEs 1, 2, 4 and 5……………...102

5.4.3 Summary of the regulatory potential of mouse *Foxb1* CNEs…….107

5.4.4 Conservation of regulation of *Foxb1* between fugu and mouse….108

5.5 Discussion…………………………………………………………..111

Chapter 6 Results: Regulation of Orexin …………………………………...…….118

6.1 Introduction…………………………………………………………119

6.2 Comparison of *ORX* loci in human, mouse and fugu………………………121

6.3 Expression of fugu *ORX* in mouse…………………………………………123

6.4 Comparative analyses and validation of *ORX* regulatory elements common in human, mouse and fugu……………………………………………………...127

6.5 Discussion…………………………………………………………..133

Chapter 7 General Discussion...138

7.1 Summary……………………………………………………………139

7.2 High-success rate in identifying functional *cis*-regulatory elements……….140

7.3 Cooperativity and redundancy in *cis*-regulatory elements………………….142

7.4 Conserved function of *cis*-regulatory elements in mammals and fish without apparent sequence conservation……………………………………………...143

References……………………………………………………………………...146

Annex I………………………………………………………………………….159

Summary

Comparative genomics is a powerful approach for identifying *cis*-regulatory elements in the human genome. Noncoding sequences that exhibit high level of conservation between genomes are likely to be under purifying selection and represent functional elements such as *cis*-regulatory elements. The pufferfish (fugu) is a particularly attractive model for discovering *cis*-regulatory elements in the human genome because of its compact intronic and intergenic regions, and its maximal evolutionary distance (~420 million years) from human. The aim of this study is to use fugu to predict conserved noncoding elements (CNEs) in genes expressing in the human forebrain, and to characterize selected CNEs in transgenic mice to identify *cis*-regulatory elements that direct tissue-specific expression in developing embryos. To this end, genomic sequences for 50 human genes that express in the forebrain were aligned with their orthologous sequences in mouse and fugu using a global algorithm program (MLAGAN) and CNEs were predicted using the criteria of at least 60% identity over 50 bp. Altogether 206 CNEs (total length ~30 kb) associated with 29 genes were identified. CNEs associated with two transcription factor genes, *Six3* and *Foxb1*, were assayed in transgenic mice using a *lacZ* reporter gene. All the CNEs assayed were found to function as *cis*-regulatory elements by either enhancing or suppressing expression of the reporter gene in a tissue- and developmental-stage specific manner. Interestingly, the highly conserved basal promoter regions of *Six3* and *Foxb1* genes were found to contain regulatory elements required for expression in almost all the domains in early and late stages of development, while the CNEs dispersed in the intergenic regions were found to 'fine-tune' the expression driven by the basal promoter by enhancing or silencing expression in particular domains. Many CNEs were found to have overlapping expression patterns reflecting the redundancy built into the regulatory code for ensuring the correct spatial and temporal expression patterns of genes. These results demonstrate that comparative genomics using fugu is a useful approach for identifying evolutionarily conserved *cis*-regulatory elements in the human genome.

I also analyzed the regulatory region of orexin (*ORX*) gene which did not contain CNEs, in order to understand the molecular basis of cell-specific expression of such genes. Despite the absence of CNEs, the fugu *ORX* regulatory region was able to direct neuron-specific expression in the hypothalamus of transgenic mice. Close inspection of sequences revealed *cis*-regulatory elements with sequence identities below the threshold level of CNEs. These vertebrate genes appear to be associated with two types of enhancers: one that is highly constrained in structure and organization and detected by a high level of sequence conservation in distant vertebrates; and another one that is weakly constrained and flexible in its organization and requires comparison with closely and distantly related species and identification by conservation at the level of transcription factor-binding sites. Thus, alternative strategies are required for the identification of all the *cis*-regulatory elements in the human genome.

List of Tables

1: List of 50 forebrain genes with the number and total length of CNEs associated with each gene…………………………………………………………………………………55

2: Number of CNEs identified and the functional categories of genes…………………..58

3: *Six3* CNEs tested in transgenic mice……………………………………………….65

4: Enhancer function of mouse *Six3* CNEs across different developmental stages and in different tissues………………………………………………………………………..83

5: *Foxb1* CNEs tested in transgenic mice……………………………………………..96

6: Enhancer function of mouse *Foxb1* CNEs across different developmental stages and in different tissues………………………………………………………………………108

List of Figures

1: Schematic diagram of the developing forebrain……………………………………29

2: Identification of CNEs in *Otp* locus in human, mouse and fugu……………………...54

3: *Six3* loci of human, mouse and fugu………………………………………………..63

4: Conserved noncoding elements in the *Six3* locus………………………………….65

5: Expression patterns of *Six3* in the developing mouse embryo………………………..68

6: A 860-bp promoter region of mouse *Six3* directs expression of *lacZ* mRNA to the forebrain and eye during embryonic development……………………………………71

7: Expression patterns directed by CNE1, CNE2/3/4 and CNE5/6/7……………………75

8: Expression patterns directed by CNE8/9 and CNE12………………………………...78

9: Expression pattern directed by CNE14 at E9.5-E11.5………………………………...82

10: Summary of the regulatory code that controls the expression of *Six3* in mouse…….89

11: *Foxb1* loci of human, mouse and fugu……………………………………………93

12: Conserved noncoding elements in the *Foxb1* locus…………………………………95

13: CNEs selected for testing in transgenic mice……………………………………..95

14: Expression patterns of *Foxb1* in the developing mouse embryo……………………98

15: A 400-bp basal promoter region of mouse *Foxb1* directs expression of *lacZ* mRNA to the diencephalon, midbrain and hindbrain during embryonic development……………100

16: Whole mount in situ hybridization showing expression patterns directed by *Foxb1* CNE1, CNE2, CNE4 and CNE5……………………………………………………..104

17: A fugu construct containing CNEs 1, 2, 4 and 5 upstream of the basal promoter containing CNE3 reproduces mouse endogenous *Foxb1* expression in the diencephalon, midbrain and hindbrain………………………………………………………………110

18: Summary of the regulatory code that controls the expression of *Foxb1* in mouse…116

19: *ORX* locus in fugu, mouse and human…………………………………………..122

20: Expression of fugu *ORX* gene in transgenic mice compared with the expression of the endogenous mouse *ORX* expression…………………………………………………125

21: Poorly conserved mouse and human regulatory elements in the fugu *ORX* locus…129

22: Analysis of the regulatory region of fugu *ORX* in transgenic mice………………...131

23: Expression of fugu *ORX* in transgenic mice compared with the expression of the endogenous mouse *ORX*…………………………………………………………..132

List of Abbreviations

bp	base pair
CTP	cytosine triphosphate
DNase	deoxyribonuclease
DEPC	diethyl pyrocarbonate
EDTA	ethylenediamine-N,N,N’,N’-tetraacetic acid
HCl	hydrochloric acid
kb	kilobase
LHA	lateral hypothalamus
MYA	million years ago
NaCl	sodium chloride
NaOAc	sodium acetate
NaOH	sodium hydroxide
PBS	phosphate buffered saline
PCR	polymerase chain reaction
RT	room temperature
SDS	sodium dodecyl sulfate
tRNA	transfer RNA
TE	tris and EDTA
TFBS	transcription factor binding sites
UTR	untranslated region (of an mRNA)

Chapter 1

Introduction

1.1 Functional sequences in the human genome

The Human Genome Project is the largest project ever attempted in biological sciences. Its main objectives are to determine the complete sequence of the human genome, and to identify and characterize all functional elements which would lead to a more complete understanding of the structure, function and evolutionary history of the human genome. The first objective was largely accomplished in 2001 when two “draft” sequences were generated (Lander et al., 2001; Venter et al., 2001). Most of the gaps in these draft sequences have since been filled-up and now the human genome sequence is essentially complete (International Human Genome Sequencing Consortium, 2004). However, although about 21,000 protein coding genes comprising about 1.5% of the human genome have been predicted, we are still far from identifying all functional elements. Since we have a good understanding of the genetic code and structure of protein coding genes, on hindsight, predicting protein-coding sequences was the easiest part of the annotation. Identifying and characterizing the “other” functional elements in the human genome which do not have a well-defined structure like the protein-coding genes, has become a major challenge in this post-human genome sequencing era.

How much of the 3000-Mb human genome sequence is functional? This is a highly debated issue with estimates ranging from 3% to 70% depending on the method used for identifying functional elements (Pheasant and Mattick, 2007; Waterston et al., 2002). A typical method for identifying functional sequences is by comparing the human genome sequence with related genomes and estimating the portion of the genome that is evolving more slowly than the neutrally evolving sequences. The slowly evolving sequences that are under selective constraint are likely to be functional elements in these genomes. A systematic comparison of the whole genome sequences of the human and mouse genomes has indicated that about 5% of these genomes are under selective constraint since they diverged from a common ancestor. This implies that at least 5% of the human and mouse genomes comprise functional sequences (Waterston et al., 2002). Since the protein-coding sequences account for 1.5% of these genomes, this analysis indicates that noncoding functional elements are three-fold higher than protein-coding sequences, and underscores the challenge in identifying and characterizing these functional elements. The non-coding functional sequences in the human genome include RNA genes such as transfer RNA (tRNA), ribosomal RNA (rRNA), and small RNAs like small interfering RNA (siRNA) and micro RNA (miRNA); transcriptional regulatory elements; splicing regulatory elements; sequences conferring structural chromatin features; and sequences playing a role in chromosomal replication and recombination. The main objective of my work is to identify and characterize transcriptional regulatory elements (referred to as "*cis*-regulatory elements" or "enhancers" in this thesis) in the human genome.

1.2 *Cis*-regulatory elements

Cis-regulatory elements are DNA sequences that mediate spatial and temporal expression patterns of genes. Transcription factors bind to *cis*-regulatory elements and activate or repress transcription of target genes associated with the *cis*-regulatory element in a cell-type or tissue-specific manner at specific developmental stages. *Cis*-regulatory elements comprise binding sites for transcription factors that are often organized into clusters called *cis*-regulatory modules (CRMs) or enhancers. The CRMs typically span a few hundred nucleotides, and can contain dozens of binding sites for ~3-10 transcription factors that activate or repress gene transcription (Chen and Rajewsky, 2007). Complex gene expression patterns frequently evolve from an orchestrated activity of several different *cis*-regulatory modules with distinct spatiotemporal activity patterns. For instance, in the Drosophila embryos the *even-skipped* (*eve*) gene, a pair-rule gene, is transcribed in alternate embryonic parasegments to generate a zebra pattern of seven stripes. The transcriptional state of this gene - either ON or OFF depending on which parasegment - is under the control of a series of CRMs, with about one module responsible for expression in each stripe (Sackerson et al., 1999).

Cis-regulatory elements also confer regulatory control in the timing of gene expression. For example, there is emerging evidence that the precise temporal expression of *Hox* genes is crucial for the establishment of regional identities. Deletion of the *Hoxd11* enhancer in mice delays expression of both *Hoxd10* and *Hoxd11* during somitogenesis, but at a later stage, normal expression of both genes is restored (Zakany et al., 1997). However this regulatory deletion could not prevent the occurrence of defects in patterning and specification of the vertebral column, although these were of less severity than the complete *Hoxd11* gene knockout (Zakany et al., 1997). Another similar study showed that the deletion of an early enhancer of *Hoxc8* resulted in a significant delay in the temporal expression but did not eliminate the expression of the Hoxc8 protein. It delayed the attainment of control levels of expression and anterior and posterior boundaries of expression on the anterior-posterior axis and this temporal delay in *Hoxc8* expression was sufficient to produce phenocopies of many of the axial skeletal defects associated with the complete absence of the *Hoxc8* gene product (Juan and Ruddle, 2003).

Cis-regulatory elements can reside close to the basal promoter, in introns, or in the 5' and 3'-flanking sequences of their target genes. In some vertebrate genes, *cis*-regulatory elements termed 'long-range enhancers', are located at several hundred kilobases away from the target gene (Bagheri-Fam et al., 2006; de la Calle-Mustienes et al., 2005; Kimura-Yoshida et al., 2004; Nobrega et al., 2003). In some instances, the long-range enhancers are embedded in the introns of the neighbouring genes. For example the limb enhancer of *Sonic Hedgehog* (*SHH)* gene has been found in the 5th intron of the neighbouring *limb region 1 homolog* (*LMBR1*) gene that is 1Mb upstream (Lettice et al., 2003); and the retina enhancer of the paired box gene 6 (*Pax6*) gene was found located in the intron of the neighbouring *elongation protein 4 homolog* (*ELP4*) gene that is located 200 kb downstream (Kleinjan et al., 2001). Thus, *cis*-regulatory elements can be potentially located within the introns or anywhere in the flanking regions of their target genes.

1.3 *Cis*-regulatory elements and genetic diseases

Cis-regulatory elements have emerged as primary candidates that are likely to harbour mutations contributing to human disease phenotypes. Although disease-associated genetic changes commonly affect gene coding regions, some may exert their effect through abnormal gene expression that results from mutations in *cis*-regulatory elements that affect their interaction with the promoter and/or disrupt the chromatin structure of the locus (Kleinjan and van Heyningen, 2005). The most obvious cases of transcriptional misregulation as the cause of genetic disease are associated with visible chromosomal rearrangements. For example, aniridia (absence of iris) and related eye anomalies are caused mainly by haploinsufficiency of the paired box / homeodomain gene *Pax6* at human chromosome 11p13 (van Heyningen and Williamson, 2002). A number of aniridia human subjects have been described with no identifiable mutation in the transcription unit. Instead chromosomal rearrangements that disrupt the region downstream of the *Pax6* transcription unit have been implicated. Detailed mapping of the breakpoints placed them about 125 kb beyond the final exon. Analysis of the region beyond this breakpoint revealed the presence of a downstream regulatory region (DRR) located about 200 kb away and within the intron of the adjacent ubiquitously expressed *ELP4* gene (Kleinjan et al., 2001). Deletion of this DRR showed that it is absolutely essential for expression of *Pax6* in the retina and iris, even in the presence of more proximal known retinal enhancers, and explains why the aniridia phenotype in ‘position effect’ patients is indistinguishable from aniridia in patients carrying coding region mutations in *Pax6* (Kleinjan et al., 2006).

On the other hand, the phenotype caused by a regulatory mutation can be very different from that caused by a coding region mutation, because such mutations may only be affecting a subset of expressing tissues. The involvement of *SHH* in preaxial polydactyl (the formation of additional anterior digits in the vertebrate limb) fits such a scenario, because while *SHH* functions in brain and neural development, it also plays a key role in defining the limb anterior-posterior axis (Kleinjan and van Heyningen, 2005). Normally

SHH is transiently expressed in the posterior part of the mouse limb and sets up a morphogen gradient from this zone of polarizing activity to instruct cells with respect to their antero-posterior fates and to specify digit identities. The limb-specific long-distance enhancer of *SHH* is located at the extreme distance of 1 Mb from the gene it regulates, residing in the intron of a neighbouring gene *Lmbr1*, and genetic lesions affecting this element is responsible for the ectopic expression of *SHH* in the limb bud, resulting in preaxial polydactyl in humans (Lettice et al., 2002). These instances of genetic diseases highlight the need for a comprehensive cataloging and characterization of *cis*-regulatory elements in the human genome, which should facilitate the identification and validation of functionally significant variants and pathological mutations in the regulatory regions of the genome.

1.4 Identification of *cis*-regulatory elements

Given that *cis*-regulatory elements comprise clusters of transcription factor binding sites and such sites are typically short (6 to 10 bp long) and allow degeneracy in their sequences, identifying functional *cis*-regulatory elements in the vast non-coding regions of the human genome is a non-trivial exercise. Although individual transcription factor binding sites can be predicted in *silico* based on similarity to experimentally validated binding sites, such predictions are likely to contain a large number of false positives. Following are some of the techniques used for identifying and validating *cis*-regulatory elements.

1.4.1 Traditional methods

Traditional methods of identifying *cis*-regulatory elements can be categorized into biochemical and genetic methods. Biochemical methods typically make use of the way DNA is packaged in the cell. Histone proteins act like molecular spools that coil the strands of DNA into bead-like units called nucleosomes, which help to organize the higher levels of chromatin structure. Genes in these tightly condensed regions are not as accessible for gene expression as compared to genes that have been unwound from their nucleosome structure. As such, DNA that is ‘unpacked’ would often be hypersensitive to endonucleases such as DNase I, and DNase I hypersensitive sites are good indicators of the presence of enhancers. To identify DNase I hypersensitive sites, nuclei are prepared from cells or a tissue and incubated with various concentrations of DNase I, and the DNA is then extracted and digested with a restriction enzyme to make a defined end from which the hypersensitive sites can be located. Early observations suggest that hypersensitivity is associated with the removal of nucleosomes but more recent analyses detect the presence of histones in modified form such as acetylation of histone H3 on lysines 9 and 14 that reduce the affinity of the DNA for the nucleosome (Bernstein et al., 2005). This in turn would facilitate the interaction of DNA with trans-acting factors, and this property is made use of in DNase I footprinting where bound transcription factors will tend to protect the ‘unpacked’ enhancer DNA from DNase I and produce a characteristic ‘footprint’ when fractionated on a gel. However this method requires prior knowledge of the transcription factors that bind the enhancer. Gel shift assays, known as electrophoretic mobility shift assays (EMSA) can also be used to show that a known transcription factor binds to a site in the *cis*-regulatory element. The labeled DNA in the form of an oligo is incubated with nuclear extract containing the transcription factor, and the mix is fractionated on an acrylamide gel. The transcription factor will retard the DNA to which it is bound as compared to the unbound DNA, and the ‘shifted’ band can be recognized easily on the gel. This method also requires prior knowledge of the transcription factor, nuclear extract from the cell types in which the gene is expressed (could be a problem if genes express in a small population of cells) and may involve a large number of oligos if the candidate *cis*-regulatory regions span a large distance.

Candidate *cis*-regulatory elements can be validated for their transcription activating potential using genetic assays that provide the appropriate array of transcription factors and conditions in which they can bind. The best assay system is an *in vivo* whole organism but tissue explants may be used when more rapid alternatives are needed. Assays in cell lines offer an attractive rapid system, if appropriate cell lines that show specific expression of genes of interest are available. Whole animal *in vivo* assay, however, provides the best means of assessing functional elements in a biologically relevant and tissue-specific context, and is the method of choice if the gene of interest is developmentally regulated. The region of the candidate *cis*-regulatory element is cloned upstream of a reporter gene and introduced into the system, and the expression of the reporter mRNA or protein is measured in specific cells or tissues and in response to regulatory signals. To locate the exact position of the *cis*-regulatory element, progressive deletions are carried out until the minimal region required for activity is identified. These experiments, however, are tedious, time consuming and expensive particularly if the candidate *cis*-regulatory regions are large as in the case of human genes.

1.4.2 High-throughput methods

The human genome sequencing era heralded the development of high-throughput methods to discover functional elements on a genome-wide scale. These methods can be classified into biochemical methods and computational methods. One recently developed biochemical method involves the use of DNase I hypersensitivity to measure the appearance and disappearance of functional sites on a genome-wide scale by comparing between cells of different tissues or comparing within the same type of cell in response to changes in the cellular environment. This method has taken form in two recently developed techniques known as quantitative chromatin profiling (Dorschner et al., 2004) and massively parallel signature sequencing (Crawford et al., 2006). At present these high throughput methods are limited in scope by the number of cell lines or tissue types available, and can produce many false positives caused by non-specific digestion of DNase at non-hypersensitive sites (Crawford et al., 2004).

Another increasingly popular method is the chromatin immunoprecipitation (ChIP) assay, which is a modification of the 'pull down' assays in which target proteins are precipitated from solution using an antibody coupled to a retrievable tag. ChIP assays capture in *vivo* protein-DNA interactions by cross-linking proteins to their DNA recognition sites using formaldehyde, fragmenting the protein-bound DNA, probing this DNA with a transcription factor-specific antibody and then reversing the cross-linking to release the bound DNA for subsequent detection by PCR amplification. Caveats to using the ChIP assay include recovering indirect interactions caused by protein-protein contact rather than protein-DNA interactions and the inability to detect precise contacts of binding within the 500 bp region (average fragment size after shearing the chromatin) of the DNA probe (Elnitski et al., 2006). High-throughput variations of the ChIP technique use ligation-mediated PCR to amplify the pool of DNA sequences as uniformly as possible, generating many copies of all genomic binding sites for a given protein. The assortment of enhancers containing these binding sites recovered in a ChIP assay can then be visualized by hybridization to a microarray of genomic sequences (Elnitski et al., 2006). This approach called ChIP-chip has been used recently to interrogate protein-DNA interactions in intact cells and in a genome-wide fashion (Kim et al., 2005). Unfortunately ChIP-chip results are dependent on the availability of suitable microarrays with high coverage and resolution, and on the affinity and specificity of the antibodies used to recognize and bind the native protein of interest (Hudson and Snyder, 2006). In addition, there is the possibility of background DNA being 'pulled down' by nonspecific interactions of protein and DNA, leading to false positives. Optimization of ChIP-chip has helped somewhat to decrease the false positive rate by paying attention to several key basics like immunoprecipitation handling, optimization of array binding conditions and the use of appropriate controls (Wu et al., 2006). Arrays used should contain a representation of the entire genome whenever possible so as to facilitate comparison between different loci represented on the array and to identify the 'best' candidate enhancers (Hanlon and Lieb, 2004).

Computational methods of identifying enhancers generally rely on their modular nature that comprises multiple transcription factor binding sites often in close proximity to each other. This clustering of sites for relevant transcription factors is considered a reliable indicator of regulatory function and has been used for the computational prediction of enhancers in coregulated genes that would share the same cluster of binding sites. Most of this kind of work has been carried out in Drosophila (Berman et al., 2004; Markstein et al., 2004). However these computational methods rely on previous knowledge of the transcription factor binding sites and composition of several experimentally characterized *cis*-regulatory elements in order to construct the predictive models, but the number of such datasets are very limited in vertebrates, which poses an obstacle in the training and testing of these methods. Recently a landmark study was carried out that identified more than 118,000 *cis*-regulatory modules in the human genome using existing transcription factor binding site information, but with no prior knowledge about coregulated genes or combinations of factors that are likely to co-occur in a module (Blanchette et al., 2006). Although a subset of these modules was shown to be bound in *vivo* by transcription factors using ChIP-chip, the predictions nevertheless contained a significant number of false positives (Blanchette et al., 2006). On the other hand, computational approaches have been more successful in identifying *cis*-regulatory elements when used in sequence comparisons between related vertebrate species, and this is elaborated in the next section.

1.5. Using comparative genomics to identify *cis*-regulatory elements

Soon after the completion of the human genome sequence, genomes of several vertebrates were sequenced starting with the genome of the pufferfish, *Fugu rubripes,* in August 2002 (Aparicio et al., 2002) and mouse in December 2002 (Waterston et al., 2002). Since then the genomes of several vertebrates have been completed (Miller et al., 2007). The availability of whole genome sequences of these vertebrates has provided an unprecedented opportunity to identify functional elements in the human genome using a comparative genomics approach. This approach relies on the principle that functionally relevant sequences are under purifying selection whereas non-functional regions are subject to neutral evolution and become divergent between species whereby functional sequences tend to stand out as more conserved than non-functional sequences. This approach is also known as “phylogenetic footprinting” because the constrained sequences leave behind a ‘footprint’ in the alignment of DNA sequences from multiple species. Phylogenetic footprinting, particularly in the non-coding region, reduces the sequence search space in a biologically meaningful way. The comparison of genomes for identifying functional noncoding elements in the human genome can be based on vertebrate genomes that are phylogenetically closely related to human (e.g., other mammals) or distantly related to human (e.g., teleost fishes). The comparisons at the extreme ends of the vertebrate phylogenetic tree have their own advantages and disadvantages.

1.5.1 Comparison of closely-related species

A pioneering study that used close-species comparison for identifying functional noncoding sequences in the human genome is that by Loots et al. (2000). In this study, about 1 Mb of human 5q31 region spanning the *interleukin-4* (*IL-4*), *interleukin-13* (*IL-13*), and *interleukin-5* (*IL-5*) gene clusters was compared with its orthologous region in the mouse and 90 noncoding sequences that exhibited equal to or greater than 70% identity over 100 bp or longer were identified. Functional characterization of the largest of these noncoding sequences (401 bp long residing in the intergenic region between *IL-4* and *IL-13*) in transgenic mice revealed that it functions as a coordinate regulator of three IL genes (*IL-4*, *IL-13* and *IL-5*) spread across 120 kb (Loots et al., 2000). Because the functional assay demonstrated that the noncoding element is a functional element (transcriptional enhancer), the threshold values used in this study for defining conserved noncoding sequences (≥70% identify across ≥100 bp) has since been routinely used for identifying conserved putative functional noncoding sequences, such as in a comprehensive comparative analysis of human chromosome 21 with syntenic regions of the mouse genome (Dermitzakis et al., 2002). Although highly conserved noncoding sequences have proven to be good indicators of regulatory elements, not all human-mouse alignments identified using a single conservation criterion necessarily indicate functional sequences, owing to the substantial variation in the rate of evolution from region to region in human and mouse genomes (Hardison et al., 2003; Waterston et al., 2002). Furthermore, because of the relatively short divergence period (70 million years) between human and mouse lineages, the high level of similarities in some regions could be due to a lack of adequate time for divergence of non-essential DNA rather than due to purifying selection. Thus, although human-mouse (or other phylogenetically closely related species) comparison is effective in identifying a large number of conserved noncoding sequences, such comparisons suffer from low specificity and contain many false positive predictions.

One effective way to increase the specificity in close species comparisons is to increase the number of species compared. The rationale of multiple genome alignment is to maximize combined branch length of the phylogenetic tree to ensure enough evolutionary time has elapsed so that non-functional regions have sufficiently diverged, resulting in higher specificity in detecting functional conserved sequences (Margulies et al., 2003). Therefore increasing the number of species used in genome comparisons makes it progressively less probable that sequences are conserved by chance, and helps in the identification of truly functional conserved sequences to be prioritized for experimental analysis. The recent examples that utilized multiple alignment of mammalian genome sequences include the discovery of regulatory motifs in human promoters and 3' UTRs by comparing the human genome with the mouse, rat and dog genomes (Xie et al., 2005), and the comprehensive identification of conserved non-coding sequences that were missed in human-mouse comparisons alone by aligning up to 12 mammalian genomes, in the analysis of a 1.8 Mb interval on human chromosome 7 (Margulies et al., 2005; Thomas et al., 2003).

Comparisons of much more closely related species such as human and non-human primates are generally dismissed as uninformative owing to their inherent sequence similarity caused by the relatively short period since they diverged from their last common ancestor in the primate branch, which is for example about 25 million years for humans and old world monkeys (Boffelli et al., 2003). On the other hand, human-primate comparisons have been used more widely to detect sequence differences that reflect positive selection in protein-coding (Enard et al., 2002) and noncoding regions (Pollard et al., 2006) that would give rise to rapid evolution in the human lineage. The closely related species do indeed contain biological insights that are not available from comparisons between species that are more evolutionarily divergent, for example primate-specific functional elements that arose in the primate lineage, which are responsible for phenotypes unique to primates. To overcome the lack of sequence variation observed between human and their primate relatives, a different approach called “Phylogenetic shadowing” involving comparisons of numerous closely related primate species has been developed (Boffelli et al., 2003). This approach takes into account the phylogenetic relationship of the set of species analyzed and identifies regions that accumulate variation at a slower rate in all the species (Boffelli et al., 2003). This method is uniquely suited to identifying primate-specific functional elements and has only been used in the context of particular loci of interest since there are currently not enough completed primate genomes to facilitate genome-wide comparisons. More recently, phylogenetic shadowing has been used to uncover conserved regulatory elements in a comparison of as few as 6 non-human primates and notably, the mouse orthologs of these elements retained regulatory activity despite the lack of significance sequence conservation (Wang et al., 2007). Therefore, comparisons between primate genomes can be used to detect both primate-specific and ancestral mammalian regulatory elements.

1.5.2 Extreme conservation within mammals

In an attempt to identify a core set of highly conserved functional elements in the human genome, extreme conservation has been used as an indicator of function. The extremely conserved elements, known as “ultraconserved elements” (UCEs), are defined as sequences that are 200 bp or longer and completely conserved (100% identity without insertions or deletions) in the human, mouse and rat genomes (Bejerano et al., 2004).

Using these criteria Bejerano et al. (2004) identified 481 UCEs. Of these, 256 are nonexonic UCEs located in the noncoding regions of the genome. Unlike the exonic UCEs which tend to be associated with RNA genes, nonexonic UCEs tend to cluster around transcription factor-encoding genes and genes involved in development. It was therefore proposed that the nonexonic UCEs function as transcriptional enhancers directing the precise spatial and temporal expression patterns of the developmental regulatory genes (Bejerano et al., 2004). Consistent with this hypothesis, experimentally validated enhancers of some transcription factor genes (e.g., *DACH1*, *Iroquois*) overlap nonexonic UCEs (de la Calle-Mustienes et al., 2005; Nobrega et al., 2003; Poulin et al., 2005). Furthermore, functional assay of 84 nonexonic UCEs in transgenic mice have confirmed that 51 of them are positive enhancers that directed tissue-specific reporter gene expression at embryonic day 11.5 (e11.5) (Pennacchio et al., 2006). This revealed a high propensity (~60%) of ultraconserved human noncoding sequences to behave as *cis*-regulatory elements in *vivo*. Interestingly, knockout of four nonexonic UCEs that had shown transcriptional enhancer activity in *vivo*, had no measurable phenotypic consequences on the knockout mice, implying that these UCEs are functionally redundant in spite of their remarkable conservation (Ahituv et al., 2007). Moreover, a large-scale transgenic mouse assay comparing the enhancer activity of almost all 256 nonexonic UCEs, with a similar number of extremely constrained CNEs lacking ultraconservation but having high human-rodent *P*-values (Prabhakar et al., 2006) showed that developmental enhancers were equally prevalent (about 50%) in both types of conserved elements (Visel et al., 2008). These results indicate that UCEs are only a subset of extremely constrained human-rodent noncoding elements that posses enhancer function.

As such, although non-exonic UCEs provide a high likelihood of identifying enhancers, they represent a relatively small subset of functionally conserved sequences that are under similar constraint in the human genome, and that many functional elements will still be missed if ultraconservation is used as the sole criteria for screening noncoding regions (Visel et al., 2007).

1.5.3 Comparison of distantly-related vertebrates

Comparison of human genome with phylogenetically distant vertebrates such as teleost fishes that diverged from the mammalian lineage about 420 million years ago is an effective method for identifying conserved functional noncoding sequences because all the neutrally evolving sequences would have diverged beyond recognition during this long evolutionary period and those that have not diverged are likely to be under purifying selection. Such deep comparison essentially offers low sensitivity but high specificity whereby most of the conserved sequences identified are likely to be functional elements. The proof of principle for this approach was first demonstrated by Aparicio et al. (1995) who used mouse and fugu comparison to identify developmental enhancers in the *Hoxb-4* locus. Of the three blocks of conserved noncoding sequences (designated CR1, CR2 and CR3) identified at this locus, one element (CR1) was found to be responsible for directing expression in the mesoderm and ectoderm while CR3 was capable of directing expression to neural tube in 10.5 day old mouse embryos (Aparicio et al., 1995). Subsequently this approach was used to identify and validate *cis*-regulatory elements in several loci in the human genome, such as *Sox9* (Bagheri-Fam et al., 2001); *Pax6* (Griffin et al., 2002); Pax9 and *Nkx2-9* (Santagati et al., 2003); and the *Dlx* bigene clusters

(Ghanem et al., 2003). Human-fugu comparison has been found to be particularly useful in prioritizing conserved noncoding sequences identified in the large intergenic regions of gene deserts. For example, the human gene *DACH* is expressed in numerous tissues and involved in the development of brain, limb and sensory organs, and is located in one of the gene deserts. It is flanked by 870 kb 5’ intergenic and 1.3 Mb 3’ intergenic regions. Comparison of the human and mouse *DACH* loci identified more than 1000 conserved noncoding elements (each longer than 100 bp long and >70% identical), but this number was reduced to 32 by comparison with several distant vertebrates including fugu. *In vivo* mouse transgenic assay of nine of these elements showed that seven of them functioned as transcriptional enhancers recapitulating several aspects of the complex endogenous *DACH* expression in 12.5 and 13.5 days post coitum mouse embryos (Nobrega et al., 2003). This demonstrates that distant vertebrates such as fugu help in prioritizing conserved elements for functional assays.

Besides fugu, other teleost fishes such as zebrafish have also been used in mammalian-fish comparisons of homologous gene loci to identify *cis*-regulatory elements. For example, two blocks of conserved noncoding sequences were identified between the zebrafish *Dlx5/Dlx6* genes and their mammalian homologs with over 80% identity across >600 bp of sequence, and their functionality was demonstrated in transgenic mice (Zerucha et al., 2000). A sequence comparison of the human and zebrafish *SHH* loci detected short stretches of conservation in the intronic regions and the upstream promoter (Muller et al., 1999). When the conserved intronic fragment was introduced into transgenic animals, the zebrafish homolog directed floor plate and notochord expression in both developing mouse and zebrafish embryos while the mouse homolog was exclusively floor-plate-specific, suggesting that some of the *cis*-regulatory mechanisms involved in regulating *SHH* expression are conserved between zebrafish and mice (Jeong and Epstein, 2003; Muller et al., 1999). However, unlike the fugu genome with its tendency toward a compact genome, the zebrafish genome has retained a higher number of duplicated genes that were generated as a result of a 'fish-specific' whole genome duplication event (Christoffels et al., 2004; Taylor et al., 2003), and it is necessary in most instances to compare the mammalian gene locus with its two zebrafish orthologs. This can be complicated if one of the duplicate genes has diverged considerably and acquired novel expression domains.

Whole genome comparisons of human and teleost fishes have also been effective in identifying a large number of putative *cis*-regulatory elements in the two genomes. Alignment of human and fugu genomes using the local alignment algorithm MegaBLAST identified 1,373 highly conserved noncoding elements (>100 bp long and >70% identical). These elements are distributed in a non-random manner in the genome, with a large number of them found in clusters predominantly in the vicinity of genes involved in transcription and development (Woolfe et al., 2005). Functional assay of 25 of these conserved elements in transgenic zebrafish indicated that 23 of them exhibit enhancer activity in one or more tissues (Woolfe et al., 2005). Taken together, these data indicate that a majority of the elements conserved in the human and fugu genomes function as *cis*-regulatory elements of transcription factor-encoding and developmental genes. A similar genome-wide comparison of human and fugu using a different approach based on quantifying the rate of decline of noncoding sequence conservation with increasing evolutionary distance by employing probability scores instead of a conservation window (Prabhakar et al., 2006), identified about 5,700 human-fugu conserved noncoding sequences. Functional assay of 137 of these elements in transgenic mouse showed that 57 of them direct tissue-specific expression in 11.5 day old embryos (Pennacchio et al., 2006). Genome-wide comparisons of human and zebrafish using the ECR browser (Ovcharenko et al., 2004) that utilized the local alignment BLASTZ were also able to identify a large number of putative regulatory elements. Using a conservation criteria of more than 70% identity and over 80 bp in length a total of about 4,800 conserved noncoding sequences were identified (Shin et al., 2005). 16 of these conserved elements were randomly chosen for experimental validation, and 11 were found to be positive for transcriptional upregulation using a dual luciferase system in transgenic zebrafish. A dual reporter system was used to allow for normalization of reporter activity due to the mosaic expression known to occur in zebrafish transient transgenesis. These elements were also found to be enriched for genes involved in development and transcription factor activity, consistent with the findings of human-fugu whole-genome comparisons (Shin et al., 2005). Yet a recent study also showed that conserved regulatory modules might be found in genes other than transcription factor and developmental regulators more frequently, if one included the possibility that regulatory modules were rearranged or shuffled within the loci (Sanges et al., 2006). This study first identified conserved noncoding sequences in at least three mammalian genomes (human, mouse and dog or rat) of at least 100 bp in length having a percentage identity of at least 70%. These conserved elements were then used to screen the fugu, zebrafish and *Tetraodon* genomes to identify shorter conserved fragments of at least 40 bp in length and 60% identity to the mouse element using a method called CHAOS (Brudno et al., 2003a) that allowed for the identification of short 10 bp regions that are reversed or moved in the fish locus with respect to the corresponding mammalian locus. Approximately 21,500 conserved elements were found, with 72% of the elements shuffled. Of the total of 27 of these elements selected for functional assay, 22 were able to direct tissue-specific expression of a reporter gene in transgenic zebrafish embryos (Sanges et al., 2006). While this unique approach has been more sensitive in identifying conserved noncoding sequences in fish, the use of short word sizes in the algorithm to aid in fish-mammalian alignments is likely to make it more difficult to distinguish between biological features preserved through evolution and neutrally evolving short fragments in the genome.

Cartilaginous fishes are a more ancient group of vertebrates than teleost fishes. They diverged from the common ancestor of human and teleost fish lineages about 450 million years ago. Therefore, comparisons of the human and cartilaginous fish genomes offer the highest stringency to identify highly conserved noncoding elements. Indeed, a comparison of the human genome with a 1.4× assembly of the elephant shark genome (comprising 134,109 scaffolds of average length 2.6 kb and covering ~75% of the genome) using the local alignment algorithm discontiguous MegaBLAST was able to identify about 5,000 highly conserved noncoding elements (≥100 bp long and ≥70% identical) (Venkatesh et al., 2006). Like the highly conserved human-fugu (Woolfe et al., 2005) and human-zebrafish (Shin et al., 2005) elements, these human-elephant shark elements were found to be predominantly associated with transcription factor genes in the human genome suggesting that they may function as *cis*-regulatory elements of transcription factor genes. However, an unexpected finding of this study was that the number of human-elephant shark elements was almost twice that identified in human-fugu and human-zebrafish (Venkatesh et al., 2006). This implies that the regulatory regions of elephant shark are evolving slower than the regulatory regions of teleost fish and as such, elephant shark is a useful distantly related genome for identifying putative *cis*-regulatory elements in the human genome. However, the currently available highly fragmented assembly of the elephant shark precludes a comprehensive comparison of human and elephant shark genomes.

In summary, whole-genome comparisons of human and distantly-related vertebrates have been effective in identifying a large number of highly conserved noncoding elements, and many of the conserved elements experimentally validated in *vivo* have been shown to function as *cis*-regulatory elements. However, whole-genome comparisons, particularly between distantly related genomes such as human and teleost fish, can fail to identify and align all the correct orthologous sequences. This is because the local alignment algorithms used are designed to be highly sensitive but less specific, and according to the scoring scheme and seeding strategy used, they will find all possible sequence similarities, not just the contiguous ones (Ureta-Vidal et al., 2003). Some of these methods were developed when the bulk of available sequences to be aligned were coding sequences, and it has been shown that such algorithms are not as efficient in aligning noncoding sequences (Bergman and Kreitman, 2001). Indeed a recent study measuring the accuracy of whole-genome local alignments at human Chromosome 1 showed that misalignments tend to occur often in noncoding regions and become more prominent with increasing phylogenetic distance from humans, with the ambiguous alignments ranging from 3% in human-mouse alignments to almost 30% in human-zebrafish alignments (Prakash and Tompa, 2007). Therefore, a comprehensive and accurate alignment requires aligning the exact orthologous regions locus-by-locus using suitable global alignment algorithms.

1.5.4 Alignment and visualization tools for comparative genomics

A number of computational tools and web-based resources have been developed for comparing genomic sequences, locus-by-locus as well as whole-genomes, for discovering and visualizing putative *cis*-regulatory elements in the human genome. Identification of conserved elements by comparative genomics is generally a two-step process. First, orthologous regions of two or more different genomes are aligned at the nucleotide level so that for each nucleotide position in the reference genome, a best fit with the nucleotide at the respective position in the other genome(s) is determined. Second, based on this alignment, the different genomes are compared at the nucleotide level and statistical methods identify regions that are more constrained than would be expected for neutrally evolving DNA.

Alignment algorithms generally fall into two categories: local and global alignment approaches. The commonly used local alignment programs include MegaBLAST (Zhang et al., 2000), discontiguous MegaBLAST (Ma et al., 2002), BLASTZ (Schwartz et al., 2003), and MULAN (Ovcharenko et al., 2005). While the MegaBLAST, discontiguous

MegaBLAST and BLASTZ are pairwise alignment algorithms, MULAN is a multiple alignment program. Local alignment programs compute similarity scores between subregions of input sequences and are used when the input sequences vary in ways that prevent an accurate end-to-end alignment, for example when rearrangements, insertions or deletions are present in one or more sequences (Frazer et al., 2003). However because they do not take into account the region surrounding these matches, they can result in a false hit, for example detecting a paralogous sequence instead of the true ortholog (Visel et al., 2007). Pipmaker (http://bio.cse.psu.edu) is a worldwide web server that combines the use of the BLASTZ algorithm with a visualization of the aligned segments in comparing two long genomics sequences (Schwartz et al., 2000). A companion server at the same site called MultiPipmaker will align three or more genomic DNA sequences. Visualization of this alignment takes the form of a percent identity plot ("Pip") displaying the position, length and percent identity (50-100%) of each gap-free segment in the pairwise BLASTZ alignments of the reference sequence with DNA from each of the other species.

In contrast to local alignment algorithms, global alignment programs compute a similarity score over the entire length of input sequences, and are used for comparing sequences that are expected to share similarity over their entire length such as regions with conserved gene order and orientation, and are likely to be more sensitive in detecting highly divergent but orthologous regions in two contiguous sequences (Frazer et al., 2003). They are less prone to return false-positive matches but fail to recognize homologous regions that have been locally rearranged by translocations or inversions

(Visel et al., 2007). Examples of global aligners are AVID (Bray et al., 2003), LAGAN (Brudno et al., 2003b), and MLAGAN (Brudno et al., 2003b). AVID looks for exact matches, limiting the comparison to closely related organisms, whereas LAGAN was designed to align both distantly and closely related organisms by using short inexact words, with level of degeneracy modified by the user. MLAGAN permits the multiple alignments of large genomic sequences. It involves a progressive alignment phase based on LAGAN, which first aligns the genomes of the most closely-related organisms, then incorporates the others in order of phylogenetic distance (Brudno et al., 2003b). MLAGAN has been found to perform better in multiple genome alignments containing distantly related genomes (Prakash and Tompa, 2007), and therefore is a useful tool in aligning and comparing mammalian and fish genomic sequences. Shuffle-LAGAN is a local-cum-global alignment program that has been specifically developed to find rearrangements during alignments and is useful to identify rearranged conserved noncoding sequences in related genomes (Brudno et al., 2003a). The VISTA server (http://www.gsd.lbl.gov/vista) is used to predict and display conserved noncoding sequences in the alignments generated first using the BLAT local alignment program and then globally aligned using AVID or LAGAN or MLAGAN (Frazer et al., 2003). VISTA plot visualizes pairwise global alignments between the reference sequence and DNA of other species by sliding a specified window (e.g., 100 bp) along each pairwise sequence alignment and calculating the percent identity at each base pair position.

1.6. Objectives of the present study

The main aim of my project is to use a comparative genomics approach for identifying evolutionarily constrained noncoding elements associated with human genes known to express in the forebrain, and to systematically validate the function of elements associated with selected genes in transgenic mice using a β-galactosidase reporter construct. I chose the forebrain genes since the forebrain is one of the most complex organs in vertebrates. It comprises many structural and functional components, with a wide range of tissue types making up each component. Furthermore, the structure and development of forebrain is highly conserved across vertebrates making it an attractive system for using a comparative genomics approach. The forebrain arises from anterior neuroectoderm during gastrulation, and by the end of somitogenesis it comprises the dorsally positioned telencephalon and the more caudally located diencephalon (see Figure 1). The dorsal telencephalon, or pallium, develops into the cerebral cortex, and the ventral telencephalon, or subpallium, becomes the basal ganglia, also known as the striatum. The diencephalon is primarily composed of the thalamus and the hypothalamus that is ventrally positioned (Figure 1). As such, forebrain morphogenesis is more complex than morphogenesis of other regions of the central nervous system. There are at least three major steps in the formation of the prospective forebrain. The ectodermal cells must acquire neural identity, the rostrally positioned neural tissue must adopt anterior character, and the regional patterning must take place within the rostral neural plate (Wilson and Houart, 2004). These steps result in a segment-like genetic organization of the forebrain, called the prosomeric model that attributes morphological meaning to known gene expression patterns and other data in the forebrain (Puelles and Rubenstein, 2003). In recent years it has become evident that several of the genetic mechanisms for establishing and patterning the vertebrate nervous system are conserved in insects (Kammermeier and Reichert, 2001) and annelids (Tessmar-Raible et al., 2007). However, despite the underlying homologies between vertebrate and invertebrate forebrains, the vertebrate forebrain is massively more complex. The vertebrate forebrain has been greatly expanded and shows evidence of compartmentalization not seen in other chordates, with the telencephalon known to be unique to vertebrates (Holland and Holland, 1999). Remarkably the general organization of the forebrain is conserved in all vertebrates including fish, reptiles, birds and mammals. What makes the brain of each species unique is not the initial presence or absence of different subdomains of the forebrain, but the way these domains are elaborated as they form the various structures that comprise the mature brain. Comparative studies in mammals, reptiles and fishes have shown conserved patterns of gene expression in the forebrain, suggesting homologies between regions in distant species (Broglio et al., 2005; Medina et al., 2005; Metin et al., 2007). Studying the *cis*-regulatory elements associated with vertebrate forebrain genes should help to better understand the expression and developmental regulation of these genes, and shed light on the regulatory complexities of forebrain development.

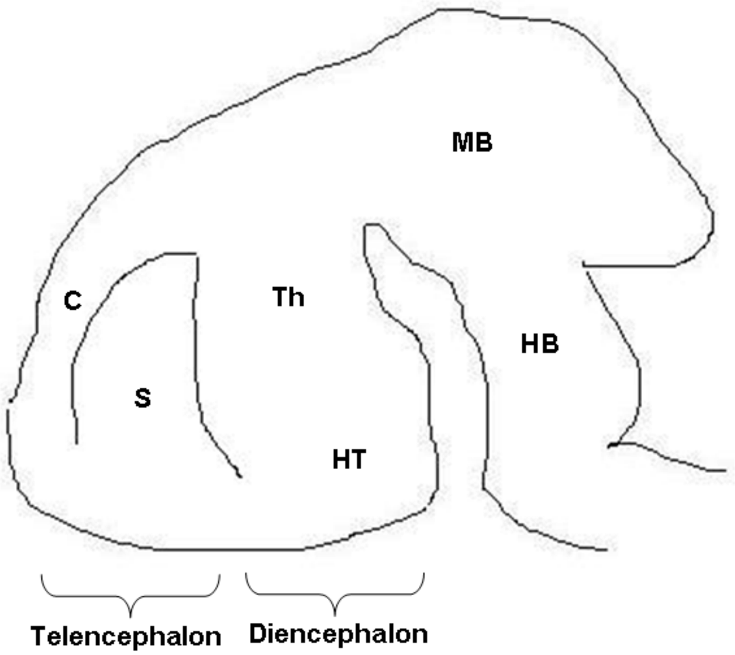


Figure 1: **Schematic diagram of the developing forebrain**. The forebrain consists of the telencephalon and diencephalon. The telencephalon comprises the cerebral cortex (C) and striatum (S), while the diencephalon comprises the thalamus (Th) and hypothalamus (HT). MB: midbrain; HB: hindbrain. Diagram was modified from Mathis and Nicolas (2006)..

Since my objective was to identify evolutionarily constrained noncoding elements, I chose to use fugu as a model for comparative genomics. The common ancestors of human and fugu diverged about 420 million years ago, and the noncoding sequences conserved in the two genomes over 840 million years of divergent evolution are likely to be under purifying selection. At 400 Mb, fugu genome is among the smallest vertebrate genomes (Brenner et al., 1993). It is about one-eighth the size of the human genome. However, fugu and human genomes contain a similar number of genes. The compaction has occurred mainly in the intergenic and intronic-regions which are typically short in fugu due to a paucity of repetitive elements (<10%). The short noncoding regions of fugu reduces the noise to signal ratio in the prediction of conserved noncoding sequences and are useful for assaying the function of multiple putative *cis*-regulatory regions at the same time. Additionally, fugu genome was the second vertebrate genome to be completely sequenced (Aparicio et al., 2002), the first being that of human (Lander et al., 2001; Venter et al., 2001). The availability of the whole genome sequence of the compact fugu genome has made it an attractive fish model genome for comparative studies for identifying *cis*-regulatory elements. Although whole genome comparisons of human and fugu have been carried out, such comparisons can fail to identify and align all the exact orthologous sequences, particularly between distantly related genomes like human and fish. Furthermore genome-wide comparisons are predicted to contain up to 25% misalignments between human and fugu (Prakash and Tompa, 2007). On the other hand, locus-by-locus comparison of orthologous sequences would be more effective in identifying all the associated CNEs, and the use of global alignment algorithms here would have more power in detecting weakly conserved regions than local alignments (Frazer et al., 2003).

To make a comprehensive search for *cis*-regulatory elements, I carried out a locus-by-locus alignment of human, mouse and fugu genes using MLAGAN. Multiple alignments of human-mouse-fugu were found to be better than pairwise fugu-human alignments in anchoring the alignment seeds, and thereby detecting conserved regions with higher specificity (Alison Lee and B.Venkatesh, unpublished data). Among the global alignment programs, MALGAN has been shown to be adept in identifying CNEs with relatively high specificity (Prakash and Tompa, 2005). For this project, I selected at random 50 human genes that are known to express in the forebrain and whose regulation has not been elucidated. From among the genes containing conserved noncoding elements, three genes representing different levels in the hierarchy of the gene regulatory network were selected and the function of the CNEs associated with them were systematically assayed in a transgenic mouse enhancer assay.

Chapter 2

Materials and Methods

2.1 Genomic sequence alignment and prediction of conserved noncoding sequences

Human genes for this study were selected by searching the Pubmed database (http://www.ncbi.nlm.nih.gov/sites/entrez) using key words such as “forebrain”, “transcription factor” and “development” to look for genes known to express in the developing forebrain, and whose regulation had not yet been well understood. The protein and nucleotide sequences for the genes were retrieved from Ensembl (http://www.ensembl.org/index.html). The mouse and fugu orthologs for these genes were identified from Ensembl BioMart and their sequences were also retrieved from Ensembl. BioMart typically identifies a single ortholog in mouse and fugu. However, fugu contains duplicate genes for many human genes due to a ‘fish-specific’ whole genome duplication (Christoffels et al., 2004). In order to identify duplicate fugu orthologs, if any, for the human genes selected for this study, I searched using a combination of data from Ensembl Biomart (fugu version 4 assembly) and INPARANOID analysis. INPARANOID has been used to identify duplicate fugu orthologs for human forebrain genes that may have been missed in Ensembl and these orthologs have been made available in the public domain on the human-fugu synteny viewer (http://humpback.bii.a-star.edu.sg/fugu-synteny/viewer.php) (Mohamad Hirwan and B. Venkatesh, unpublished).

The genomic sequences, comprising the entire 5’ and 3’ flanking regions, for each of the human, mouse and fugu genes were retrieved from Ensembl. The use of global alignment algorithms here would have more power in detecting weakly conserved regions than local alignments (Frazer et al., 2003) and among the global alignment programs, MALGAN has been shown to be adept in identifying CNEs with relatively high specificity (Prakash and Tompa, 2005).The sequences of the orthologous human, mouse and fugu gene loci were therefore aligned using the global alignment algorithm MLAGAN (http://genome.lbl.gov/vista/lagan/) using fugu as the reference sequence. Reverse complementation of sequences was necessary to harmonize sense and antisense sequences prior to upload. The conserved noncoding elements (CNEs) were predicted and visualized using VISTA (http://genome.lbl.gov/vista/index.shtml). Annotation files of the fugu reference sequence were obtained from Ensembl to achieve the VISTA plots showing exon structure of the reference gene. The CNEs between human and mouse are generally predicted using the criterion equal to or greater than 70% identity over 100 bp or more (Loots et al., 2000) or greater than 60% identity over 50 bp of sequence.

In order to exclude any coding sequences among the CNEs that were missed in the genome annotation, I searched the CNEs using BLASTX against NCBI's non-redundant protein database and the significant matches (E-value $<10^{-4}$) were eliminated. The remaining CNEs were searched using both BLASTN (E-value $<10^{-4}$) and INFERNAL searches against the Rfam database (Release 7.0) and miRNA registry (Release 8.0), and those containing RNA sequences were excluded from further analysis. The final set of CNEs should comprise mainly transcriptional enhancers, chromatin structural sequences and other regulatory elements.

The functional categories of the forebrain genes were determined by identifying the Gene Ontology (http://www.geneontology.org/) terms associated with them. The transcription factor (TF) binding sites in CNEs were predicted using the program TESS (transcription element search system) (http://www.cbil.upenn.edu/cgi-bin/tess/tess). TF binding sites are usually short (6-15bp) sequences with degeneracy at several positions, and hence TESS predictions may contain many false positives. In order to reduce the number of false positives, only those binding sites that showed 90% identity to known binding sites in the TRANSFAC database and were totally conserved in all the three genomes (human, mouse and fugu) were retained.

2.2 Generation of DNA constructs for microinjection

The functions of individual CNEs were assayed by linking them to their basal promoter and a β-galactosidase reporter. The β-galactosidase reporter vector, pnlacF (Mercer et al., 1991) was constructed by Jacques Peschon and kindly supplied by Richard Palmiter from Howard Hughes medical institute, Washington, USA. The basal promoter and individual or clusters of CNEs were amplified by PCR using mouse genomic DNA or fugu genomic DNA as a template. The PCR primers for the basal promoters (690 to 860 bp) contained restriction sites for *Xba*I upstream and *Sal*I downstream. The PCR amplicon was fractionated on a TAE agarose gel and excised from the gel and purified by the Geneclean II kit (Qbiogene, USA). The purified product was digested with *Xba*I and *Sal*I enzymes and cloned into the respective cloning sites of the pnlacF vector. The primers for amplifying CNEs (ranging from 220-740 bp in size) contained restriction site for *KpnI* upstream and downstream, and amplicons included about 100 bp of sequence on either side of the CNEs. The CNE-PCR amplicons were digested with *KpnI* and cloned into the *KpnI* site upstream of the basal promoter in the *lacZ*+promoter construct. The orientation and sequences of the inserts (promoter and CNEs) were verified by sequencing on an ABI3730xl (Applied Biosystems, USA) automated DNA sequencer using the big dye terminator chemistry.

2.3 Isolation and sequencing of fugu cosmid to map the orexin locus

The latest assembly of the fugu draft genome sequence (http://www.fugu-sg.org) contains 7213 scaffolds spanning 393 Mb. The fugu scaffold (#424; 131 kb) containing the orexin gene was identified by TBLASTX search using human prepro-orexin cDNA sequence. Searching the fugu scaffold sequence against the non-redundant protein database at the National Center for Biotechnology Information (NCBI) using BLASTX algorithm confirmed that it contained a homolog of human orexin (*ORX*) gene. The other genes on the scaffold were identified and annotated in a similar way based on their homology to sequences in the NCBI database. The information about the order and orientation of genes at the human and mouse *ORX* loci was obtained from the UCSC genome browser at http://www.genome.ucsc.edu. Alignment of Fugu and mammalian *ORX* sequences was performed by Megalign (DNAStar) using the ClustalW algorithm with Gonnet 250 protein weight matrix, and gap penalty of 10.

The fugu scaffold sequence surrounding the *ORX* gene contained several gaps. To fill the gaps and to make constructs for transgenic studies, I isolated fugu cosmids for this locus. A fragment of the fugu *ORX* locus (680 bp of exon 1, intron1 and exon2 amplified by PCR using the primer pairs forward: 5’- CAG AAA GGC ACG AGG ATG TCC-3’ and reverse: 5’- GTT TGCCCA GCG TGAGGA TGC -3’) was used to probe a fugu cosmid library (Greg Elgar, UK HGMP Resource Center). Altogether 10 positive cosmids (19J5, 33B9, 40H16, 106I3, 117J20, 132F24, 151I10, 173F1, 199L2 and 199L3) were isolated. The end sequences of the cosmids were determined using primers complementary to the cosmid arm on an automated DNA sequencer and compared to the fugu scaffold sequence. Cosmid 33B9 with an insert size of about 39 kb was selected as a representative clone for sequencing. The fugu scaffold sequence spanning this cosmid contained ten gaps ranging from 50 bp to 1 kb. The gaps were filled by directly sequencing the cosmid DNA using primers flanking the gaps. Transgenic constructs were made by deletions of this cosmid using appropriate restriction enzymes.

2.4 Generation of transgenic mice

The mouse strains used were as follows: the embryo donors were FVB/N F1 mice and pseudo pregnant recipients were B6CBA F1 mice. Mice were cared for in accordance with National Institutes of Health (NIH), USA guidelines. Transgenic mice were generated essentially as described by Murphy and Carter (1993). Briefly, clean linearized DNA was microinjected into single-cell embryos and implanted into the pseudo pregnant mothers. At different embryonic developmental stages, the mothers were sacrificed to harvest the embryos for analyzing the expression of the introduced transgene. Embryo sacs or yolk sacs were used for genotyping to identify founders. For orexin constructs, the transgenic mice from the microinjected embryos were allowed to mature to 3 weeks old, and then tail clipped and genotyped to identify founders. Founders were then crossed in a series of breeding experiments to generate transgenic lines of mice, individuals of which were then analyzed for expression of the introduced transgenes. At least 3 founder lines of transgenic mice were generated for each construct.

2.5 Preparation of DNA for microinjection

The DNA has to be cleaned of a number of contaminants that can kill the embryos, particularly excess EDTA and endotoxin (bacterial lipopolysaccharide). Transgenics are generated with linear DNA with the naked ends acting as substrate for nonhomologous end joining to be incorporated into the host genome. Cosmid and plasmid DNA meant for microinjection were prepared with silica-based miniprep columns (Promega, USA). Miniprepped DNA was linearized with appropriate restriction enzymes. Vector sequences were removed to minimize rearrangement of transgenes in the nuclei of the embryos. If the clone was a plasmid, it was fractionated on an agarose gel to separate the insert from the vector. The insert DNA was extracted from the agarose gel using sodium iodide solution (Geneclean II kit, Qbiogene, USA). If the clone was a cosmid, then the restriction digest was purified directly on a silica-based miniprep column (Qiagen, USA). The DNA was quantified and diluted to 4 ng/μl using ‘microinjection TE’ (10 mM Tris pH 8, 0.1 mM EDTA), filtered (0.2 μm disposable filter, Sartorius, Germany) and then submitted to the Biological Resource Centre of Biopolis, Singapore who carries out microinjection for researchers as a core service.

Briefly, the process leading up to microinjection as performed by the Biological Resource Centre is as follows. 3-week old FVB/N female mice are superovulated with 10 IU of pregnant mare’s serum (Sigma-Aldrich, USA) followed by 10 IU of human chorionic gonadotrophin (Sigma-Aldrich, USA) 46 to 48 hours later. They are then mated with 3 to 6-month old FVB/N stud males and the following morning the mated females are checked for ‘ovulatory plugs’. Presence of the ovulatory plugs on the mice is taken to be signs of successful mating. The plugged females are sacrificed using carbon dioxide and the oocyte-cumulus complex (OOC) is then surgically retrieved from the ampulla of the oviduct in M2 medium (Sigma-Aldrich, USA). The oocytes are further released from the oocyte-cumulus complex using hyaluronidase (Sigma-Aldrich, USA). The oocytes are then cultured in M16 medium (Sigma-Aldrich, USA) and covered with mineral oil (Sigma-Aldrich, USA) in a 5% CO_2 incubator at 37^{o}C. During microinjection of DNA, DNA is introduced into one of the pronucleus of two-pronuclear embryos under 400x magnification using Leica micromanipulator with Nikon 2000 Eclipse microscope. Embryos that have survived DNA injection are transferred into the oviducts of CBAB6F1 pseudo pregnant female mice. These mice are then handed over to me to sacrifice at the appropriate developmental stages to harvest transgenic embryos, or allowed to give birth to produce transgenic lines for mating.

2.6 Genotyping

Mice about three weeks old were ear clipped for identification and about 5 mm of tail taken for genotyping. If embryos were harvested, yolk sacs or embryos sacs of the same size were removed for genotyping. The tail clip or sac was digested in 300 μl of tail digestion buffer (50 mM Tris pH 8, 100 mM NaCl, 10 mM EDTA, 0.1% SDS, 0.4 mg/ml proteinase K) at 55^{o}C for at least three hours. An equal volume of isopropanol was then added and the mixture shaken vigorously for ten seconds. This would cause the DNA to quickly appear as a stringy white precipitate. The precipitate was immediately spun down at 14,000 g for two minutes in a bench top centrifuge, the supernatant discarded and 70% ethanol was added to wash the pellet. This was again spun down for five minutes, the supernatant tipped off and the remainder spun down for five seconds and pipetted out. The pellet was dried for three minutes in a vacuum, then dissolved with vigorous vortexing in 40 µl TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0). The DNA concentration was consistently about 1.0-1.2 µg/µl. The DNA can be genotyped either by PCR or Southern blotting. Initially I performed both procedures and because they gave consistent results, I continued with PCR as the preferred way of genotyping.

For Southern blotting, briefly 10 µl of each DNA sample was digested with 20-30U of a restriction enzyme for 1-2 hours at 37^{o}C, and then fractionated on a 0.5-1.0% TBE agarose gel (Invitrogen, USA). The gel was agitated in transfer / denaturation solution (1.5M NaCl, 0.25M NaOH) for 15 minutes. The DNA in the gel was then transferred (for three hours to overnight) from the gel to a HybondN nylon membrane (Amersham, United Kingdom) by capillary transfer. After the transfer, the position of the wells was marked on the membrane, and the membrane was rinsed in 2 x SSC (20 x SSC contains 3M NaCl and 0.3M TriNaCitrate) for 30 seconds. The DNA was then cross-linked in the UV cross linker (Stratalinker, Stratagene, USA). The membrane was prehybridized for 20 minutes in Church and Gilbert solution (0.25M Na_2HPO_4 adjusted to pH 7.4 with orthophosphoric acid; 7% SDS; 1mM EDTA). The purified probe was prepared by running either a PCR reaction or a digested plasmid on TAE gel, excising the desired band and extracting it using the Geneclean II kit. This probe is then melted to singlestranded DNA and labeled with [α-^{32}P]dCTP using the Random Labeling kit (Boehringer-Ingelheim, Germany). The labeled and purified probe was denatured by boiling at 100^{o}C for 5 minutes and hybridized to the membrane overnight. Next morning the membrane was rinsed with wash solution (2 x SSC, 0.1% SDS) to remove excess probe, then washed twice for 15 minutes each. This was usually sufficiently stringent for genotyping. The membrane was then blotted on paper towels, wrapped in Saran wrap and exposed to X-ray film.

PCR genotyping was conducted with *lacZ*-specific primers (forward primer: 5'-TTT CCA TGT TGC CAC TCGC -3'; reverse primer: 5'-AAC GGC TTG CCG TTC AGCA -3'). The conditions comprised a denaturation step at 95^{o}C for 2 min, 35 cycles of primer annealing and extension: 95^{o}C for 30 sec, 60^{o}C for half a min, 72^{o}C for 1 min, followed by a final elongation step at 72^{o}C for 5 min. The PCR reaction mixture was carried out in a volume of 20 μl consisting of 1x PCR buffer (i-DNA biotechnology, Singapore); 0.2 mM of each dNTP (Amersham, United Kingdom), 0.2 μM of each primer (1st Base, Singapore) and 1 U of *Taq* polymerase (i-DNA biotechnology, Singapore). The PCR products were analyzed by agarose gel electrophoresis using a 1 kb ladder (Invitrogen, USA) as a marker, and transgenic lines were identified by the presence of a 374 bp band.

2.7 In situ hybridization

2.7.1 Preparation of embryos and tissues for whole-mount or section in situ hybridization

Instead of β-galactosidase staining, I employed the more sensitive technique of *in situ* hybridization to detect the mRNA of the *lacZ* reporter transgene in mouse embryos and tissue sections. Mice were killed by cervical dislocation or carbon dioxide gassing. Mouse embryos were harvested from a range of 8.5 to 15.5 days post coitum (E8.5-E15.5) and dissected in 1 x DEPC-treated PBS, pH 7.4. Individual yolk sacs were dissected out and collected for genotyping while the embryos were transferred into freshly prepared 4% paraformaldehyde (Sigma-Aldrich, USA) in DEPC-treated PBS in 6-well or 12-well or 24-well Nunclon (Apogent, Denmark) cell culture plates using sterile disposable pipettes. Briefly, 4% paraformaldehyde solution was prepared by dissolving the powder in 1 x DEPC-treated PBS and heating it to about 60^{o}C on a hot plate with gentle stirring in a fume hood. A few drops of 10M NaOH were added until the solution is completely clear, and pH is adjusted to 7.4 with HCl. For convenience, 20% paraformaldehyde in PBS stock solutions were made and frozen at -20^{o}C, ready to be diluted for use. After transfer, the wells were examined to ensure the embryos (especially the smaller E9.5-E10.5 ones) were not floating on the surface of the fixative where they would be destroyed or damaged by surface tension; if so additional fixative was used to cause the embryos to sink. Embryos were fixed for 3 hours to overnight with gentle shaking at 4^{o}C. Smaller embryos (E8.5-E11.5) were then dehydrated through a methanol (MeOH) series: 25% MeOH / 75% PBT (1 x DEPC-treated PBS, 0.1% Tween 20), 50% MeOH / 50% PBT, 75% MeOH / 25% PBT, and then twice in 100% MeOH, each for 5 minutes. These were then stored in -20^{o}C in 100% MeOH until ready for hybridization.

After fixing, larger embryos (E13.5-E15.5) were transferred to 30% sucrose (BDH, Great Britain) dissolved in DEPC-treated water and incubated with gentle shaking until all the embryos sank to the bottom of the well; this may take up to two days. The rehydrated embryos were then beheaded and the heads embedded in OCT mounting medium (Sakura Finetek, USA) in plastic boats (Polysciences Inc, USA), and the boats were placed in dry ice/ethanol bath until they were frozen through. The same procedure was applied when harvesting tissues from adult mouse. Briefly, these tissues were first fixed in 4% paraformaldehyde, then rehydrated in 30% sucrose and finally embedded in OCT medium in plastic boats. The orientation in which the tissue is embedded depends on the kind of section that will be taken. These tissue containing OCT blocks in the plastic boats were then stored in the -80^{o}C freezer until ready for sectioning.

Sectioning of tissues was carried out in a cryostat (Leica, USA). 10 to 20 µm sections were cut from the OCT blocks at -18^{o}C to -20^{o}C, and thaw mounted onto polysine coated slides (Menzel-Glaser, Erie Scientific Company, USA). The slides were left to dry on a heat block at 37^{o}C for about an hour after sectioning, then stored in a slide box and kept in the -80^{o}C freezer until the hybridization.

2.7.2 Synthesis of RNA probes for in situ hybridization

To create a template for riboprobe synthesis, the probe insert of about 300-400 bp in size was cloned into a pBluescript vector (Stratagene, USA) with a ‘T’ overhang, that contained a T3 and T7 promoter sites, and the vector was linearized by *EcoR1* or *HindIII* respectively to create a 5’ overhang. After enzyme digestion, the DNA was precipitated with 0.1 volume 3 M NaOAc and 2.5 volumes 100% ethanol, washed with 70% ethanol and resuspended in 40 µl TE buffer (pH 8.0) at a final concentration of 1 µg/µl. Riboprobe synthesis was then carried out as follows using a non-radioactive label: To a final volume of 20 µl, add 1 µl template (1 µg), 4 µl of 5x transcription buffer (Roche, Germany), 2 µl of digoxygenin (DIG) or fluorescein (FITC) RNA labeling mix (Roche, Germany), 1 µl of RNase inhibitor (Roche, Germany) and 40 units of T3 or T7 polymerase (Roche, Germany). The mixture was incubated at 37^{o}C for 2-3 hours, then treated with 1 µl of RNase-free DNase I (10 units/ul; Roche, Germany) and incubated at 37^{o}C for 15 minutes. Probe synthesis was checked by running a 2 µl sample on a 1% agarose gel. RNA was then precipitated by the addition of 2 µl 0.5M EDTA (pH 8.0), 5 µl of 4M LiCl and 125 µl of 100% ethanol and incubated at -80^{o}C overnight. On the next day, RNA was spun down at 14,000 g for 30 minutes at 4^{o}C, the pellet washed with cold 70% ethanol and air-dried briefly. RNA was then resuspended in 30-50 µl DEPC-treated water (depending on the size of the pellet) and stored at -80^{o}C.

2.7.3 Pretreatment of embryos and sections

Embryos (8.5-11.5 dpc) were rehydrated through 75% MeOH / 25% PBT, 50% MeOH / 50% PBT, 25% MeOH / 75% PBT and finally twice in PBT, with 5 minutes for each wash. Using a 27 G needle, punctures were made in the head and trunk of E10.5 and E11.5 embryos to facilitate the penetration of the various reagents and probe. Embryos were then bleached with 6% H_2O_2 (diluted in PBT from fresh 30% concentrated stock) for 1 hour at room temperature, and washed three times with PBT for 5 minutes each. Embryos were then permeabilized with proteinase K (Roche, Germany) diluted in PBT at room temperature, for the appropriate length of time depending on their stage, and observed under the microscope. Proteinase K treatment was stopped by adding freshly prepared 2 mg/ml glycine in PBT for 10 minutes at room temperature, then washed twice with PBT for 5 minutes each. Embryos were then postfixed with freshly made 4% paraformaldehyde / 0.2% glutaraldehyde (Sigma-Aldrich, USA) in PBT for 20 minutes at room temperature, followed by two PBT washes, before being transferred into prehybridization solution in 1.5 ml or 2 ml eppendorf tubes, depending on the size and number of embryos used. The prehybridization solution comprised 50% formamide (Invitrogen, USA), 5x SSC pH 4.5, 50 μg/ml yeast tRNA (Roche, Germany), 1% SDS (BDH, Great Britain) and 50 μg/ml heparin (Sigma-Aldrich, USA) all dissolved in DEPC-treated water. Prehybridization was carried out at 70^{o}C in a water bath for at least an hour.

Slides were pretreated as follows: 4% paraformaldehyde in DEPC-PBS was prepared as described previously and added to Coplin jars (Analytical technology, USA; each can hold up to 10 slides and a liquid volume of 25-30 ml), and the slides were transferred directly from the freezer to the jars. Fixing took place for an hour at room temperature. At the end of this incubation, the solution was tipped out of the jar and the slides were washed with DEPC-treated PBS three times, 5 minutes each. The next solution 0.2% Triton X in DEPC-treated PBS was freshly made and added for an incubation of 10 minutes, followed by another three washes of PBS. Slides were then incubated a third time for 10 minutes with acetic anhydride solution of pH 8.0 that comprised 390 μl of triethanolamine (Sigma-Aldrich, USA) and 75 μl of acetic anhydride stock (Sigma-

Aldrich, USA) in 30 ml of DEPC-treated water. After washing three times with PBS, slides were arranged face up in a humidified chamber and sufficient (100-200 µl) pre-hybridization solution (Dako, Denmark) was added to cover all the sections in each slide. In addition, parafilm was added onto each slide to ensure the sections did not dry up. The chamber was then placed in an oven at 55-58^{o}C for at least two hours.

2.7.4 Hybridization, washing and antibody addition

The prehybridization solution in the Eppendorf tubes containing the embryos was replaced with fresh hybridization solution containing 1 µg/µl riboprobe (ratio is about 1 µl probe: 400 µl hybridization solution). Tubes were then inverted a few times gently to mix the probe, sealed with tape or parafilm, and submerged in a water bath at the same temperature of 70^{o}C overnight. On the next morning, embryos were rinsed with freshly made Solution 1 (50% formamide, 4x SSC at pH 4.5 and 1% SDS) and washed three times at 70^{o}C for 30 minutes each. Tubes were inverted often to mix and to ensure embryos did not stick to the walls. New tubes were used after the first round of wash. A second series of three washes with freshly made Solution 2 (50% formamide, 2x SSC at pH 4.5) was then carried out at 65^{o}C for 30 minutes each. Following this, embryos were transferred to a 6-well or a 12-well plate (depending on the number of tubes) and washed in TBST (1x TBS diluted from a 10x TBS stock that comprised 0.5M Tris Base and 9% NaCl at pH 7.6; 0.1% Tween 20; 2 mM levimasole (Sigma-Aldrich, USA)) three times at room temperature for 5 minutes each, with gentle shaking. Embryos were then incubated in blocking reagent (Boehringer Ingelheim, Germany) dissolved in TBST and containing 10% heat-inactivated sheep or horse serum (Gibco, Invitrogen, USA) for at least an hour at 4^{o}C with shaking, before being replaced by a new aliquot of blocking solution containing the pre-absorbed antibody labeled with alkaline phosphatase (AP) and directed to DIG (Roche, Germany) at 1:2000 concentration ratio and left to shake overnight at 4^{o}C.

After hybridization overnight in the humidified chamber, the parafilm was removed and the slides were transferred back into Coplin jars and washed twice in 1x SSC / 0.1% Tween 20 for 5 minutes at room temperature. The second round of washes was carried out using 1x SSC / 50% formamide for two washes at 20 minutes each at 60^{o}C. A third round of two washes used 0.2x SSC / 0.1% Tween 20 at 60^{o}C for 20 minutes each. The final round of washes was in 1x PBS / 0.1% Tween 20 at room temperature for two washes at 5 minutes each. The slides were then taken out of the jars, excess fluid was tipped off (without allowing the slides to dry), and were arranged face-up in the humidified chamber again. Blocking agent dissolved in 1x Maleic acid (Sigma-Aldrich, USA) was added in excess to each slide, covered with parafilm and left at room temperature for at least an hour. Following this, blocking agent was tipped off and fresh blocking buffer was added that contained antibody either labeled with AP or horseradish peroxidase (HRP) and directed to DIG or FITC (Roche, Germany) at a concentration of 1:500, and slides were incubated in the chamber at room temperature for about 2 hours or left overnight at 4^{o}C if detection is to be carried out on a third day.

2.7.5 Visualization

After antibody detection, embryos were washed with TBST at room temperature for the whole third day, initially three times for 5 minutes each, and then five times for 1 hour or more each time, before being left to shake overnight at 4°C. On the fourth day, embryos were washed in NTMT (100 mM NaCl, 100 mM Tris-HCl at pH 9.5, 50 mM $MgCl_2$, 0.1% Tween 20, 2mM levimasole) for three times at room temperature at 10 minutes each. Embryos were then transferred to cavity dishes and detection solution containing 20 μl NBT / BCIP (Roche, Germany) in 1 ml NTMT and 10% polyvinyl alcohol (Sigma-Aldrich, USA) was added to each dish and kept in the dark. NTMT containing 10% polyvinyl alcohol was heated to 70°C with stirring until the polyvinyl alcohol was completely dissolved, before being cooled down to room temperature and then the NBT/BCIP mix was added. Polyvinyl alcohol significantly increases detection sensitivity. The progress of the staining reaction was monitored at hourly intervals using a dissection microscope. Staining reaction was stopped by washing embryos twice with PBT at pH 5.5, then postfixing embryos for 1 hour with freshly made 4% paraformaldehyde / 0.1% glutaraldehyde in PBS, followed by washing three times with PBS. Embryos were then cleared for visualization of signals by shaking in 50% glycerol (Invitrogen, USA) / PBS for 30 minutes, followed by 80% glycerol / PBS for 30 minutes before storing in 100% glycerol at 4°C until ready for photography.

For detection of signals in tissue sections, the antibody mix covering each section was tipped off and slides were transferred back to the Coplin jars and washing was carried out using TBST for three washes at 10 minutes each. Slides were then rinsed with AP-detection buffer (100 mM NaCl, 100 mM Tris-HCl at pH 9.5, 50 mM $MgCl_2$) three times at 5 minutes each. Following this, NBT / BCIP was mixed with the detection buffer as described before and added to each slide in excess (150-200 µl per slide). The slides were allowed to develop in the dark in the humidified chamber, and the staining reaction which could take 10 minutes to several hours to develop, was monitored under a microscope at regular time intervals. Staining was stopped by rinsing in MilliQ water and slides were washed three times for 5 minutes each. Slides were then mounted using crystal mount, left to dry and stored at 4^oC in a humidified chamber until ready for photography.

2.8.6 Double in-situ hybridization

Double *in situ* hybridization was carried out to show colocalization of the endogenous gene and transgene in the same cell type. Slides were pretreated as described above, and during hybridization, both probes were added together on the same slide. Washing was carried out the next day as described. Both antibodies used for probe detection were then added together on the same slide. One was directed to DIG and labeled with AP (Roche, Germany), while the other antibody was directed to FITC and labeled with HRP (Roche, Germany). The visualization stage was carried out in two steps. First, the signal attributed to the FITC-specific probe was developed using DAB solution (0.02% (w/v) 3, 3'-diaminobenzidine tetrahydrochloride (DAB) and 0.005% (v/v) H_2O_2 in 0.05 M Tris-HCl buffer, pH 7.6) that produced a brown signal that is not soluble in water. Following this, slides were rinsed in AP-detection buffer described above for three times at 5 minutes each, and the DIG-labeled probe signal was developed using the Nuclear Fast Red reagent (Sigma-Aldrich, USA) or the Vector Blue reagent system (Vector Laboratories, USA). Slides were then post-processed as described above.

Chapter 3

Results

Identification of CNEs in forebrain genes

3.1 Introduction

Comparative genomics is a powerful approach for identifying functional noncoding sequences in the human genome. Orthologous noncoding sequences that are highly conserved over long evolutionary periods are most likely to be functional elements that are under purifying selection. Such elements include transcriptional enhancers, RNA genes; splicing regulatory elements; sequences conferring structural chromatin features; and sequences playing a role in chromosomal replication and recombination. The main focus of my project is to use a comparative genomics approach for identifying gene regulatory elements associated with forebrain genes in the human genome. I chose fugu as a model for comparison because of its maximal evolutionary distance from the human genome (~420 Myr) and its compact intergenic and intronic sequences. I first chose a set of 50 human genes at random that are known to express in the forebrain and whose regulatory mechanisms have not been well characterized. Although whole genome comparisons of human and fugu have been carried out, such comparisons can fail to identify and align all the exact orthologous sequences, particularly between distantly related genomes like human and fish. On the other hand, locus-by-locus comparison of orthologous sequences using a global alignment program is more effective in identifying all the conserved non-coding elements (CNEs) including the weakly conserved regions (Frazer et al., 2003). In this study, sequences for the orthologous gene loci from human, mouse and fugu genomes were systematically extracted and compared locus-by-locus using the global alignment algorithm MLAGAN, and the CNEs were predicted using VISTA.

3.2 Identification of human, mouse and fugu forebrain genes

Human forebrain genes for this study were selected by searching the Pubmed database (http://www.ncbi.nlm.nih.gov/sites/entrez) using a combination of key words such as “human”, “forebrain”, “gene expression” and “gene regulation”, and by reading the relevant publications. The list of the human forebrain genes selected is given in Annex I. The nucleotide sequences encompassing the entire 5’ and 3’-flanking sequences of these genes were retrieved from Ensembl (http://www.ensembl.org/index.html). The mouse and fugu orthologs for these genes were identified using reciprocal BLAST and INPARANOID (see Materials and Methods). Of the 50 human forebrain genes I started with, all 50 have orthologs in mouse and fugu, with fugu containing two orthologs (“co-orthologs”) for 8 of the genes. These co-orthologs in fugu are likely to be the result of a whole-genome duplication in the fish lineage (Christoffels et al., 2004; Vandepoele et al., 2004).

3.3 Prediction of CNEs

The orthologous genomic sequences, comprising the entire 5’ and 3’ flanking regions, for each of the human, mouse and fugu genes were aligned using the global alignment algorithm MLAGAN (http://genome.lbl.gov/vista/lagan/) using fugu as the reference sequence. Consistent with its compact genome size, the fugu loci were about one-eighth the size of the human and mouse loci. The CNEs between human and mouse are generally predicted as sequences that exhibit a minimum of 70% identity across 100 bp or more (Loots et al., 2000). However, in human and fish comparisons, considering the longer evolutionary distance between these two lineages, a less stringent criterion of a minimum 60% identity across 40 bp or more has been used for defining human-fish CNEs. Indeed, a significant proportion of human-fish CNEs predicted using these criteria has been found to function as transcriptional enhancers directing tissue-specific expression during *in vivo* assays (Sanges et al., 2006; Woolfe et al., 2005). Therefore I decided to use the threshold values of 60% identity and 50 bp window sizes for identifying mammal-fugu CNEs in the 50 forebrain genes.

Altogether a total of 206 CNEs with a combined length of 30 kb were identified (Table 1). The average length of these CNEs is 139 bp, with the largest CNE spanning 1024 bp. Slightly more than half of the CNEs (107 CNEs) were located in the 5’ intergenic regions while about a quarter (61 CNEs) were located in the 3’ intergenic region with the rest distributed in the introns. A representative VISTA plot of a MLAGAN alignment is shown in Figure 2.

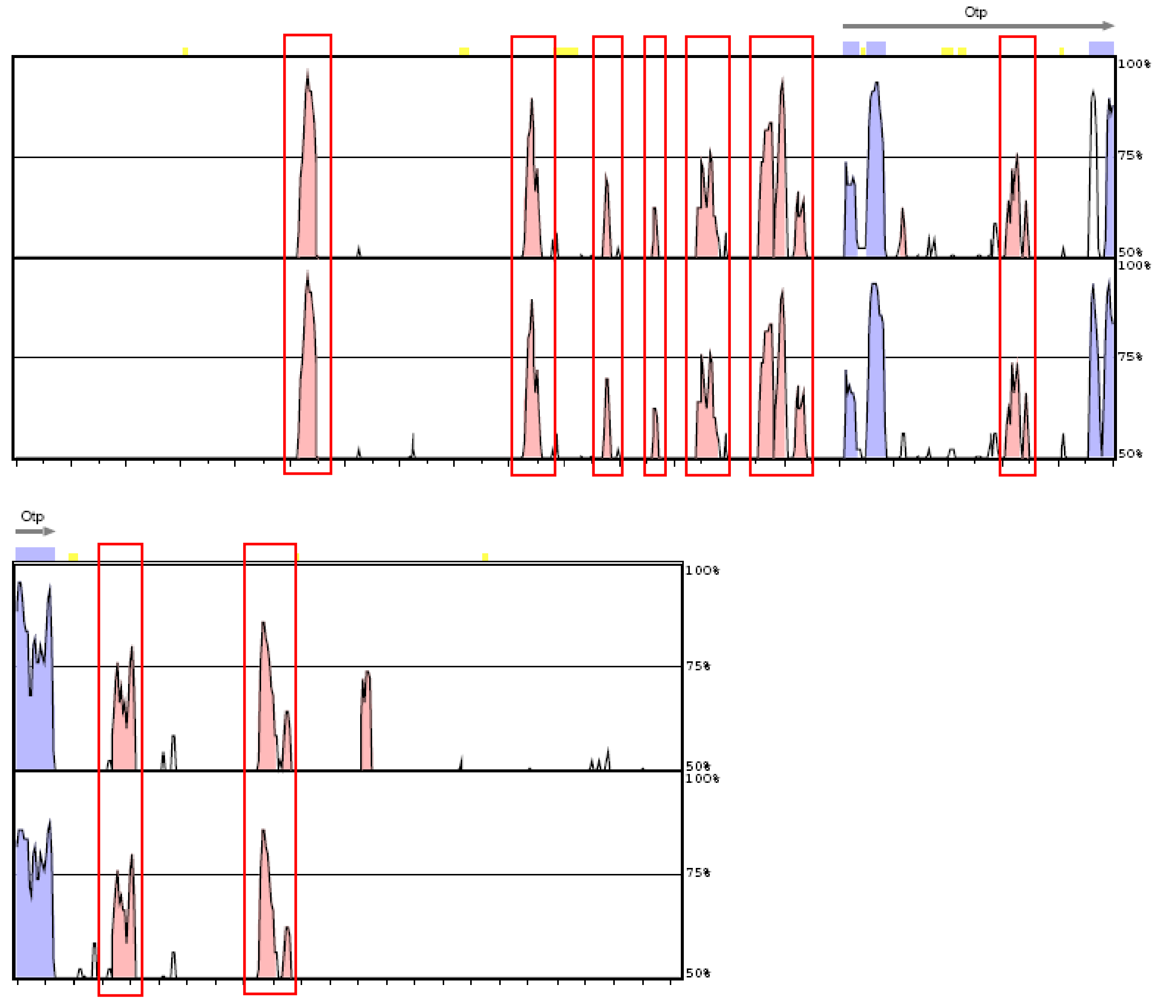


Figure 2: **Identification of CNEs in *Otp* locus in human, mouse and fugu.** VISTA plot of the MLAGAN alignment is shown. The human and mouse loci span 200 kb and 133 kb respectively while the fugu locus is 45 kb long. The MLAGAN alignment was generated using fugu as the base sequence. Peaks represent conserved sequences; coding sequences are shaded purple and non-coding sequences (CNEs) are shaded pink. The 9 CNEs identified are shown inside red rectangle boxes. X-axis represents fugu sequence and Y-axis represents percent identities (50%-100%).

Table 1: **List of 50 forebrain genes with the number and total length of CNEs associated with each gene**.

No	Gene description	Symbol	No. of CNEs	Length of CNEs (bp)
1	Empty spiracles homeobox 1	EMX1	0	-
2	Aristaless related homeobox	ARX	9	1054
3	Ventral anterior homeobox 1	VAX1	2	490
4	Orthodenticle homeobox 1	OTX1	3	265
5	Retina and anterior neural fold homeobox	RAX	1	58
6	Orthopedia homeobox	OTP	13	2020
7	GS homeobox 1	GSH1	4	1707
8	GS homeobox 2	GSH2	6	1509
9	Paired-like homeodomain 2	PITX2	16	2398
10	Sine oculis-related homeobox homolog 3	SIX3	14	1910
11	Sine oculis-related homeobox homolog 6	SIX6	3	278
12	Cartilage paired-class homeoprotein 1	CART1	0	-
13	LIM homeobox 2	LHX2	1	58
14	LIM homeobox 5	LHX5	6	627
15	LIM homeobox 6	LHX6	3	201
16	LIM homeobox 8	LHX7/ LHX8	1	50
17	POU class 3 homeobox 3	BRN1/ POU3f3	22	2765
18	POU class 3 homeobox 2	BRN2/ POU3f2	5	590
19	Transducin-like enhancer of split 1	TLE1	1	50
20	Single-minded homolog 1	SIM1	0	-
21	T-box brain gene 1	TBR1	1	70
22	Eomesodermin homolog	TBR2/ EOMES	1	81
23	Cellular nucleic acid binding protein isoform 1	CNBP1/ ZNF9	2	561
24	FEZ family zinc finger 2	FEZF2/ ZFP312	9	708
25	Zinc finger protein of the cerebellum 2	ZIC2	4	591
26	GLI-Kruppel family member isoform 2	GLI2	9	899
27	GLI-Kruppel family member isoform 3	GLI3	22	3258

28	Forkhead box G1	BF1/FOXG1	4	335
29	Forkhead box B1	FOXB1/FKH5	27	5237
30	Forkhead box H1	FOXH1/FAST1	0	-
31	Hypocretin (orexin) neuropeptide precursor	HCRT	0	-
32	Cholecystokinin preproprotein	CCK	0	-
33	Neuropeptide Y	NPY	0	-
34	Agouti related protein	AGRP	0	-
35	Thyrotropin-releasing hormone	TRH	0	-
36	Somatostatin	SST	0	-
37	Cocaine and amphetamine regulated transcript	CART	0	-
38	Pro-melanin-concentrating hormone	PMCH	0	-
39	Calcitonin-related polypeptide alpha	CGRP/CALCA	0	-
40	Proenkephalin	PENK	0	-
41	Nerve growth factor (beta polypeptide)	NGFB	0	-
42	Brain-derived neurotrophic factor	BDNF	10	1302
43	Insulin-like growth factor 1	IGF1	4	506
44	Vasoactive intestinal peptide	VIP	0	-
45	Cryptochrome 1 (photolyase-like)	CRY1	0	-
46	Cryptochrome 2 (photolyase-like)	CRY2	0	-
47	Ring finger protein 111 / Arkadia	RNF111 /ARK	3	312
48	Noggin	NOG	0	-
49	Chordin	CHRD	0	-
50	TGFB-induced factor homeobox 1	TGIF	0	-
	Total		206	29890

Of the 50 genes analyzed, about a third of the genes (17 of 50 genes) contained 4 or more conserved elements in their noncoding sequences, with 7 genes containing more than 10

CNEs per locus. On the other hand, no CNEs were detected in about 40% of the genes (21 of 50 genes). To ascertain the types of genes associated with CNEs, I identified the gene ontology (GO) terms (http://www.geneontology.org/) associated with the 50 genes. The GO terms associated with genes containing different spectrum of CNEs are shown in Table 2. CNEs tend to be associated with genes that encode transcription factors and those that play a role in development. This is consistent with the whole-genome comparisons of human and fugu in which a significant proportion of CNEs identified were found to cluster around transcription factor and developmental genes (Bejerano et al., 2004; Woolfe et al., 2005). However, about a quarter of the genes which lack CNEs (5 out of 18 genes) were found to encode transcription factors or developmental genes. Thus it seems that not all types of transcription factor and developmental genes contain evolutionarily constrained regulatory elements. It is possible that these genes are either regulated differently in mammals and fish, or regulated similarly but their regulatory elements are divergent, or that these genes may express ubiquitously and as such do not require tissue-specific enhancers. Indeed, verification of their expression patterns as indicated in the GNF SymAtlas (http://symatlas.gnf.org/SymAtlas/), showed that all these five transcription-factor genes are expressed in a wide range of tissues. These tissues may therefore lack tissue-specific enhancers. My analysis of CNEs also showed that some non-transcription factor genes, such as genes encoding hormones and metabolic regulators (lipid catabolism), contain 3 or more CNEs. The regulatory network of these non-transcription factor genes seems to be highly conserved during evolution. This implies that these genes might play a fundamental role in the physiology of vertebrates.

Table 2: **Number of CNEs identified and the functional categories of genes**. TF: transcription factor; MCH: melanin-concentrating hormone

No of CNEs	No of genes	Gene Ontology terms
>10	6	DNA binding; TF activity; regulation of transcription; development; nucleic acid binding; Zn ion binding
	1	Growth factor activity
4-10	9	DNA binding; TF activity; regulation of transcription; development; nucleic acid binding; Zn ion binding
	1	Hormone activity
3	3	DNA binding; TF activity; regulation of transcription; development
	1	Phospholipase A2 activity; calcium ion binding; lipid catabolism
2	2	DNA binding; TF activity; regulation of transcription; nucleic acid binding; Zn ion binding
1	6	DNA binding; TF activity; regulation of transcription; development; Zn ion binding
0	5	DNA binding; TF activity; regulation of transcription; development
	5	Signal transducer activity; signal transduction; hormone-mediated signaling; neuropeptide signaling pathway
	6	Hormone activity; MCH activity
	3	Growth factor activity; negative regulation of cell differentiation
	2	DNA photolyase activity; DNA repair

Of the 206 CNEs identified in my analysis of 50 human genes, only 22 overlap with the CNEs identified in whole-genome comparisons of human and fugu (Woolfe et al., 2005). Thus my analysis has identified a large number of novel CNEs associated with forebrain genes in the human genome.

3.4 Summary

By analyzing the noncoding regions of 50 human forebrain genes, I was able to identify 206 CNEs associated with 29 genes. These CNEs include a large number of novel CNEs that were not identified in previous comparisons of human and fish genomes. These

CNEs are likely to be functional noncoding sequences that are under selective pressure in human and fish genomes. In order to confirm that they are indeed transcriptional enhancers, I proceeded to systematically test the functions of some selected CNEs in transgenic mouse using a β-galactosidase reporter gene. For this purpose, I selected two transcription factor genes that function at different levels in the hierarchy of the gene regulatory network of forebrain development, and assayed the functions of 13 CNE regions associated with them at different stages of mouse embryo development. The two genes I selected are *Six3* and *Foxb1*. *Six3* is a master regulator essential for the early specification and development of the forebrain and eye. *Foxb1* is a transcription factor required for the normal development of the diencephalon and mammary glands. In addition, I also decided to analyze the regulatory region of a gene that shows tissue-specific expression but does not contain any CNE. The gene I selected was orexin (*ORX*), which codes for a neuropeptide hormone and functions as an effector downstream in the gene-regulatory network. The objective of this experiment is to determine how in the absence of a conserved regulatory element the tissues-specific expression of a gene is achieved. To ascertain this I first generated transgenic mice carrying regulatory sequences of fugu to determine if the fugu *ORX* gene is regulated in the same way as the mouse gene, and then I localized the regulatory region in the fugu gene by progressive deletion of its regulatory region. I will present the results of the analysis of the three genes (*Six3*, *Foxb1* and *ORX*) in the next three chapters.

Chapter 4

Results

Regulation of *Six3*

4.1 Introduction

The transcription factor *Six3* is a member of the *Six/sine oculis* family of homeobox transcription factors, which also contain a SIX domain that binds to cofactors and participate in transcriptional activation (Lopez-Rios et al., 2003). The first vertebrate *Six3* gene was cloned and characterized in mouse (Oliver et al., 1995). In early embryonic development *Six3* expression is restricted to the anterior neural plate, including regions that will later give rise to ectodermal and neural derivatives, suggesting that this gene is involved in establishing positional information at the anterior boundary of the developing mouse embryo (Oliver et al., 1995). Vertebrate *Six3* gene has been subsequently isolated from chicken (Bovolenta et al., 1998), medaka (Loosli et al., 1998), zebrafish (Seo et al., 1998), *Xenopus laevis* (Zhou et al., 2000), and human (Granadino et al., 1999). Sequence comparisons show extensive identity within the homeodomain and the SIX domain. In all vertebrates, *Six3* is expressed from the neurala stage first in the anteriormost neural plate and then in its derivatives: the developing eyes and olfactory placodes, the hypothalamic pituitary regions and the ventral telencephalon.

A number of studies that manipulated *Six3* expression in fish and mouse have demonstrated its essential role in eye and forebrain development (Kobayashi et al., 1998; Lagutin et al., 2001; Loosli et al., 1999). Targeted inactivation of *Six3* in medaka fish embryos by morpholino knock-down resulted in the lack of forebrain and eye development (Carl et al., 2002). Similarly *Six3*-null mice displayed anterior truncation of the forebrain and died at birth, lacking most of the head structures anterior to the midbrain including the eyes (Lagutin et al., 2003). Conditional deletion of mouse *Six3* in the presumptive lens ectoderm (PLE) disrupted lens formation, and showed that *Six3* is essential in the PLE to activate *Pax6* expression for lens induction (Liu et al., 2006). In addition, mutations in *Six3* have been found in humans affected by holoprosencephaly (Wallis et al., 1999), which is a severe malformation of the brain that involves separation of the central nervous system into left and right halves. Many gene loci have been implicated in the aetiology of holoprosencephaly including *Six3*. Mutational analysis in holoprosencephaly patients has identified four different mutations in the homeodomain of *Six3* that are predicted to interfere with the transcriptional activation of *Six3* (Wallis et al., 1999).

How the activity of *Six3* with multiple functions and several expression domains during embryo development is regulated remains to be fully elucidated. To identify *cis*-regulatory elements that mediate expression of human *Six3* to different domains during development, I chose to identify conserved noncoding elements in the *Six3* loci of human, mouse and fugu and validate their function in transgenic mice during embryonic development.

4.2 *Six3* loci in human, mouse and fugu, and identification of CNEs

Identification of *Six3* ortholog in fugu using Ensembl Biomart annotation and INPARANOID showed that fugu contains a single ortholog unlike duplicate copies of *Six3* discovered in zebrafish (Seo et al., 1998) and medaka (Conte and Bovolenta, 2007). The genomic sequences for the human, mouse and fugu *Six3* genes were retrieved from Ensembl (see Materials and methods). Scanning of the genes located upstream and downstream of *Six3* in the human, mouse and fugu genome assemblies indicated that the synteny of the genes in this locus is highly conserved in the three genomes (Figure 3). The protein sequence of *Six3* is encoded by a single exon in human, mouse and fugu. The 5’ and 3’ flanking sequences of *Six3* span 580 kb and 50 kb in human and 420 kb and 50 kb in mouse, respectively. The flanking sequences of *Six3* is unusually large (76 kb and 15 kb; Figure 3) in fugu in which the overall gene density is only one gene/~18 kb and the intergenic regions are generally short (http://www.fugu-sg.org/). The vast noncoding sequences flanking the *Six3* gene indeed poses a challenge for identifying *cis*-regulatory elements regulating the expression of this gene.

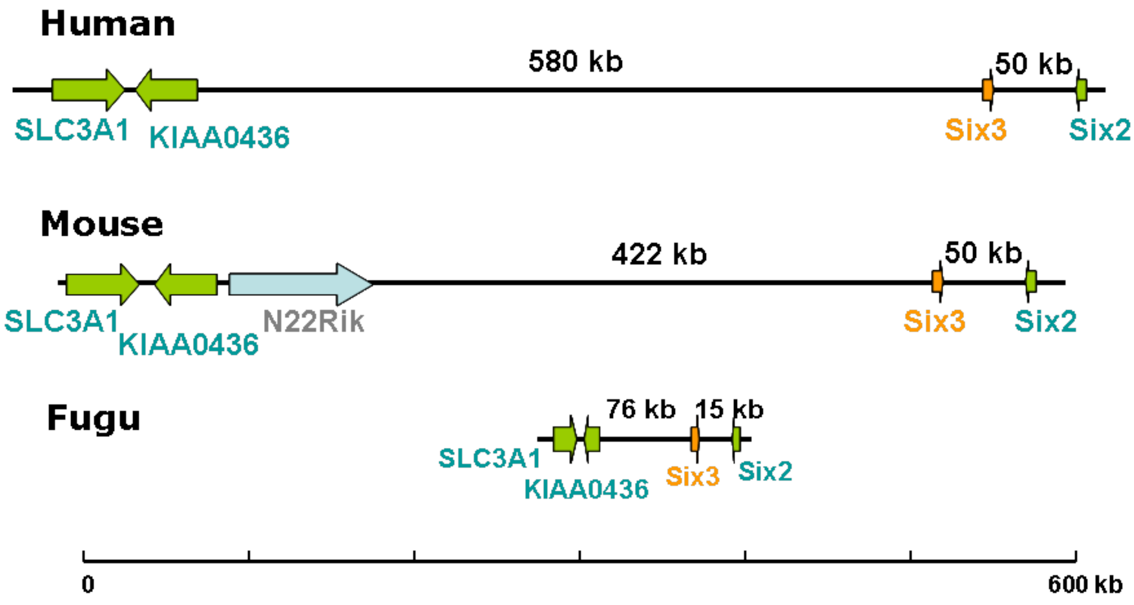


Figure 3: ***Six3* loci of human, mouse and fugu**. Arrows represent genes and the direction in which they point indicate the direction of transcription.

The orthologous genomic sequences comprising the entire 5’ and 3’ flanking regions of the human, mouse and fugu *Six3* genes were aligned using the global alignment algorithm

MLAGAN (http://genome.lbl.gov/vista/lagan/) using fugu as the reference sequence, and CNEs (>60% identity and larger than 50 bp) were predicted using VISTA. A total of 14 mammal-fugu CNEs were predicted in the *Six3* loci (Figure 4). While one of these CNEs (CNE14) is located in the 3'-flanking region, CNE13 (224 bp) is located just 200 bp upstream of the transcription initiation site and thus likely to overlap with the upstream basal promoter of *Six3*. The remaining CNEs (CNE1 to CNE12) are located in the 5'-flanking region. In the human genome, these CNEs are spread over a region of 163 kb with the most upstream CNE located 140 kb from the transcription start site. For the purpose of testing these CNEs in transgenic mice, CNEs located in clusters were grouped together and their combined sequences were amplified by PCR while the rest were amplified as individual CNEs (Figure 4). One drawback of the clustering is that the expression patterns of the cluster reflect the combined expression pattern of the CNEs clustered and as such, the contribution of individual CNEs is not clear. The details of the CNEs tested are given in Table 3. In all, eight noncoding sequences were tested in transgenic mice. I first tested the function of the basal promoter (that includes CNE13) alone by cloning it upstream of a β-galactosidase reporter, and then tested the functions of other CNEs by cloning each of them upstream of the basal promoter. Expression of *lacZ* was assayed by in situ hybridization using a *lacZ*-specific riboprobe on transgenic embryos or embryo sections at various developmental stages. The resulting expression profiles from each construct shared by at least three transgenic embryos for each developmental stage would be described in the subsequent sections.

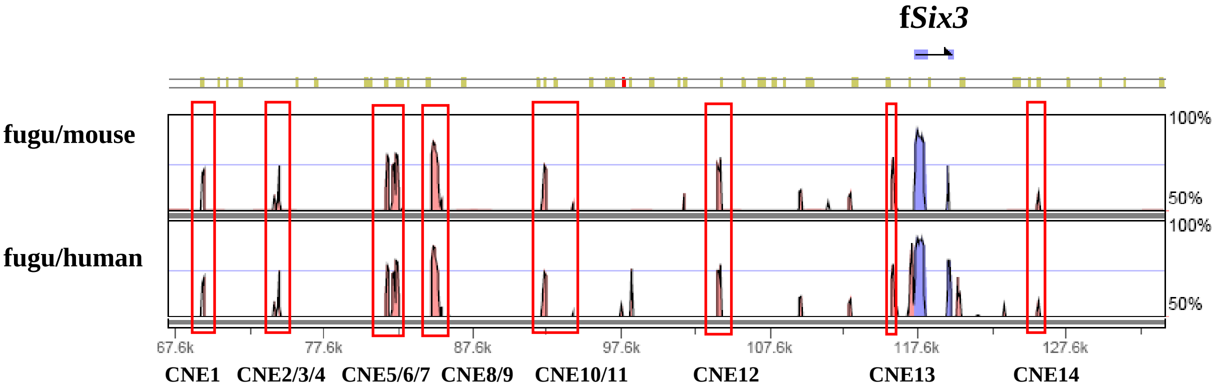


Figure 4: **Conserved noncoding elements in the *Six3* locus**. VISTA plot of the MLAGAN alignment of the fugu, mouse and human *Six3* loci is shown. Fugu was used as the base sequence. CNEs were predicted as noncoding sequences that are ≥60% identity across 50 bp or longer. Peaks represent conserved sequences; coding sequences are shaded blue and non-coding sequences (CNEs) are shaded pink. The arrow indicates the direction of transcription of *Six3*. The 14 CNEs are highlighted by red rectangle boxes. CNE13 likely overlaps with the promoter sequence. The pink peaks outside the red boxes overlapped with NCBI EST sequences and hence were not counted as CNEs. X-axis represents fugu sequence and Y-axis represents percent identities (50%-100%).

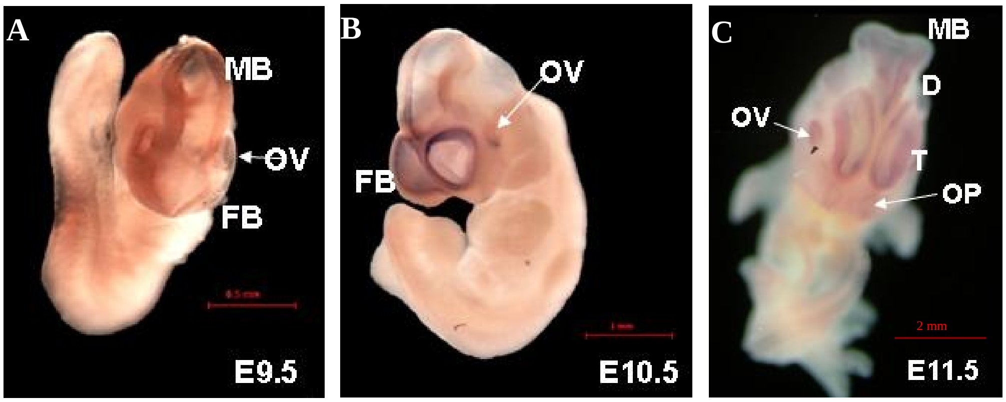
CNE	**Number of CNEs merged**	**Combined length (bp)**	**% identity**	**Approximate distance from TSS of human *SIX3***	**Length (bp) amplified by PCR**
CNE1	1	80	76.3	-140 kb	220
CNE2/3/4	3	203	68.4	-138 kb	480
CNE5/6/7	3	496	71.3	-103 kb	740
CNE8/9	2	473	75.1	-100 kb	610
CNE10/11	2	149	79.5	-63 kb	445
CNE12	1	88	85.2	-42 kb	290
CNE13	1	224	75.0	-200 bp	860
CNE14	1	78	66.7	+20 kb	380

Table 3: ***Six3* CNEs tested in transgenic mice**. The length, percent identity and location of the CNEs are shown. The actual size of the mouse noncoding sequence amplified and cloned into a *lacZ* reporter construct is shown in the last column. TSS: transcription start site.

In the course of my work, a similar comparative genomics approach was used to identify the regulatory elements of *Six3* gene in the teleost fish, medaka (Conte and Bovolenta, 2007). Like zebrafish, medaka contains duplicate copies of *Six3*. Conte and Bovolenta aligned ~20 kb sequence upstream of translational start site of one of the medaka *Six3* genes (*olSix3.2*) with corresponding sequences from the single copy *Six3* in fugu and Tetraodon, and the duplicated *Six3* in zebrafish, and identified 10 blocks of conserved noncoding sequences (>75% identity over 100 bp) contained within the first 4.5 kb genomic region flanking the translation start site. Functional assay of these conserved blocks of noncoding sequences (designated blocks A to L) in transgenic medaka revealed that they include two enhancer modules, D and I, that recapitulate expression of medaka *olSix3.2* at early (stages 16-21) and late (stages 24-40) stages of brain development, respectively, and two 'silencers' and two 'silencer blockers' that restrict the spatial expression of the two enhancers. In addition, a combination of five different modules (spread between modules D and H) was found to be responsible for the expression of olSix3.2 in the lens ectoderm and in the differentiating retina during stages 22-23. Thus, the conserved modules within the first 4.5 kb genomic sequence were found to be adequate for driving expression of *olSix3.2* during early and late stages of brain development. Interestingly, only one of these conserved noncoding sequence blocks (block L) overlapped with a fugu-mammal CNE (CNE13) identified by me. This block did not exhibit any function when assayed in transgenic medaka (Conte and Bovolenta, 2007). Although another medaka conserved sequence (block G) overlapped with a conserved noncoding sequence in my MLAGAN alignment, it was excluded from my set of CNEs as it showed high similarity to mouse *Six3* opposite strand transcript sequence from the NCBI EST database. It should be noted that Conte and Bovolenta were able to identify rearranged fragments of medaka blocks H and I in human and Xenopus *Six3* loci and these rearranged sequences were able to recapitulate part of the combined expression patterns of H and I in transgenic medaka. Since these rearranged conserved fragments fall below the conservation criteria used for identifying fugu-mammal CNEs, they were not predicted as CNEs by VISTA in my study. Thus, most of the conserved noncoding blocks of medaka sequences identified by Conte and Bovolenta (2007) appear to be specific to teleost fishes. In contrast, the CNEs identified by me are evolutionarily constrained sequences that are conserved all the way from teleost fishes to mammals, and are likely to have functions in teleost fishes and mammals.

4.3 Expression pattern of mouse *Six3*

The expression pattern of *Six3* in the mouse embryos during the early stages of development (embryonic days E9.5 to E11.5) and the late stages of development (embryonic days E13.5 to E15.5) has been previously investigated (Oliver et al., 1995). However, for the sake of comparison of the expression pattern of the mouse *Six3* gene with the expression pattern of the reporter gene driven by mouse *Six3* CNEs, I determined the expression pattern of the *Six3* gene in the FVB/N mouse strain used for testing the CNEs. The expression in the early stages of development (E9.5-E11.5) was analyzed by a whole-mount in situ hybridization using an antisense RNA probe that binds specifically to a 380 bp fragment of the mouse *Six3* coding region whereas expression during the late stages of development was investigated by in situ hybridization of sagittal (E13.5) or coronal (E15.5) sections of the mouse embryos using the same probe. At E9.5, the expression of *Six3* was detected in the ventral forebrain and optic vesicles, with some signal observed in the midbrain tegmentum (Figure 5A). By E10.5, the expression was localized predominantly in the forebrain, midbrain tegmentum and optic vesicle (Figure 5B). At E11.5, the expression persisted in the differentiated telencephalon and diencephalon of the forebrain, and the optic vesicle with some expression in the midbrain tegmentum. Low levels of expression were also detected in the olfactory placodes (Figure 5C). At later stages of development (E13.5), *Six3* labeling was clearly seen in the ventral thalamus, hypothalamus and Rathke's pouch (Figure 5D) which are all derived from the diencephalon; as well as in the olfactory epithelium of the nasal cavities (Figure 5E) and the neural retina (localized in the inner neuroblastic layer) and lens of the differentiated eye (Figure 5F). *Six3* mRNA was also seen in the telencephalon, specifically in the striatum (Figure 5G). At E15.5, expression remained in the thalamus and telencephalon (Figure 5H) and in the hypothalamus and eye (Figure 5I). These expression patterns are in general agreement with the expression patterns of mouse *Six3* previously reported by Oliver et al. (1995).



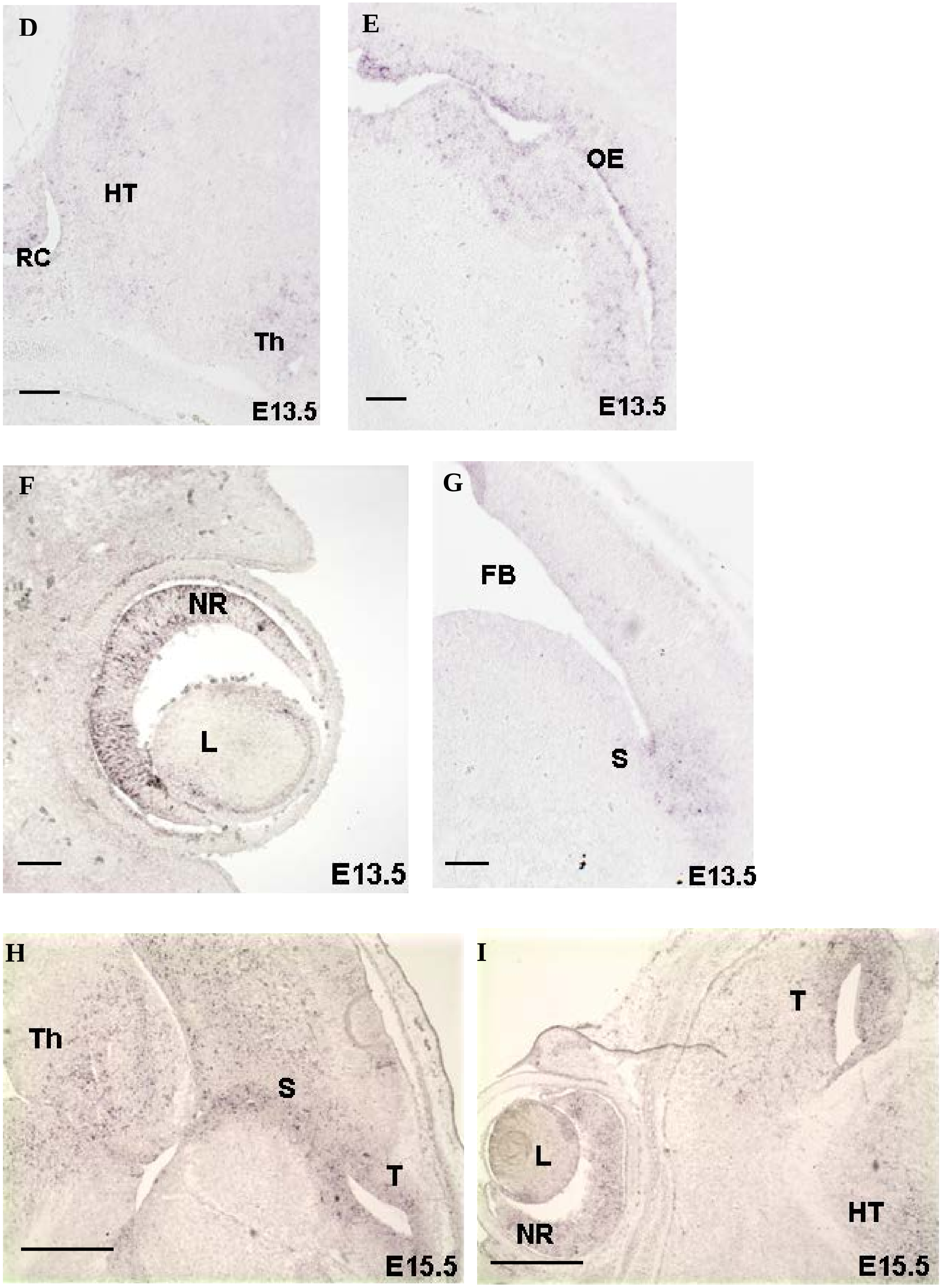


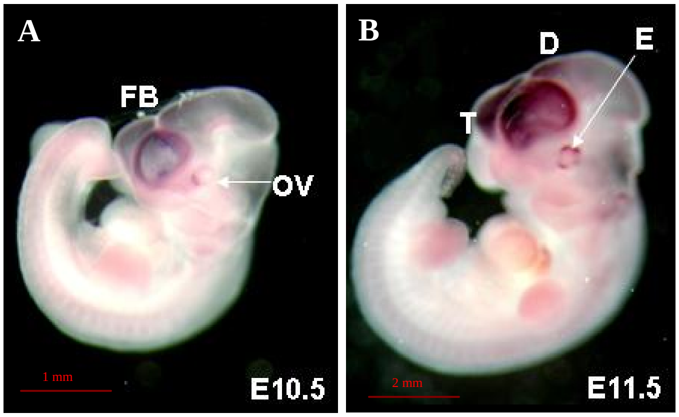
Figure 5: **Expression patterns of *Six3* in the developing mouse embryo**. (A-C) Whole mount in situ hybridization of wild type embryos showing expression of mouse *Six3*. At E9.5, mRNA accumulation was seen in the forebrain, midbrain tegmentum and optic vesicle (A). At E10.5 expression was predominant in the forebrain and optic vesicle (B).

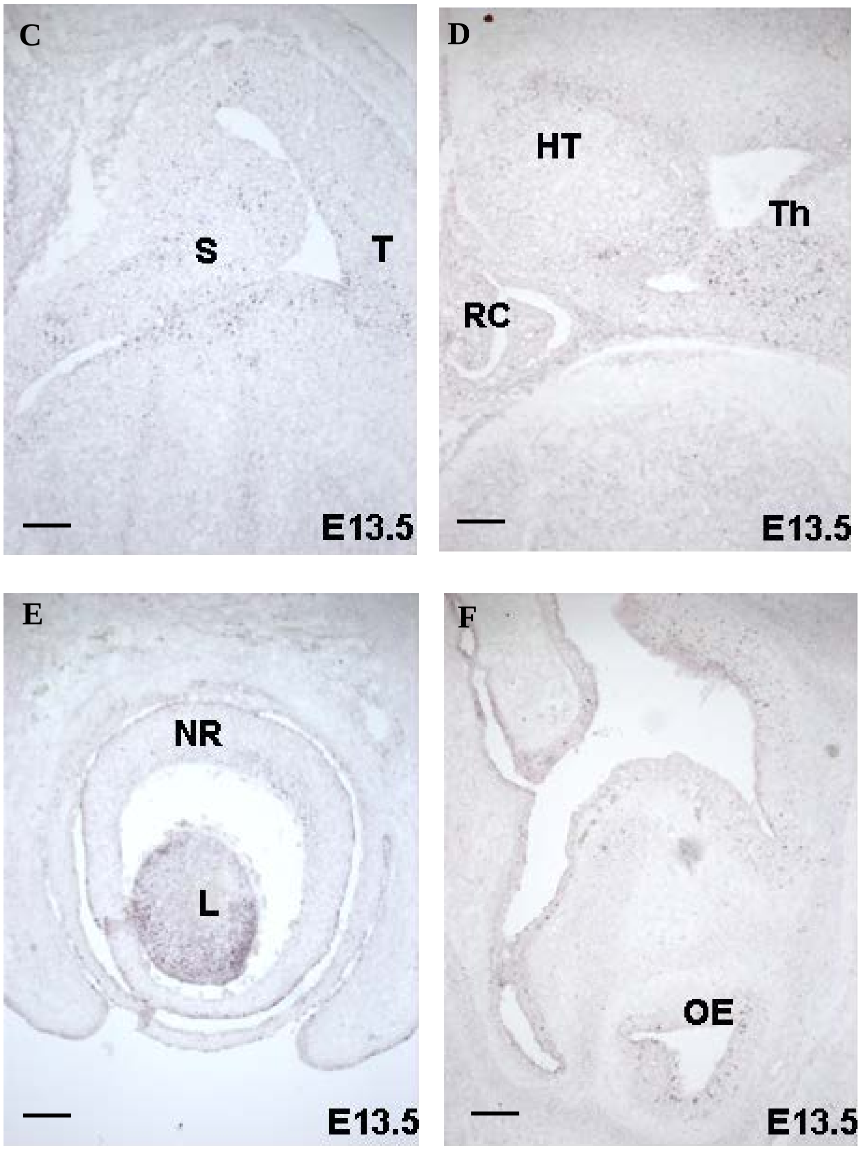
At E11.5, expression persisted in the differentiated telencephalon and diencephalon of the forebrain, the optic vesicle and weakly in the olfactory placodes (C). (D-I) In situ hybridization of cryosections of the head of wild-type embryos with a 380 bp fragment of the mouse *Six3* exon as a probe at E13.5 (D-G) and E15.5 (H-I) show that mouse *Six3* expression is detected at E13.5 in the thalamus, hypothalamus and Rathke’s pouch of the diencephalon (D); olfactory epithelium (E); neural retina and lens of the eye (F); and striatum of the telencephalon (G) and at E15.5, expression was maintained in the telencephalon and thalamus (H); and hypothalamus and neural retina and lens of the differentiated eye (I). D: diencephalon; FB: forebrain; HT: hypothalamus; L: lens; MB: midbrain; NR: neural retina; OE: olfactory epithelium; OP: olfactory placodes; OV: optic vesicle; RC: Rathke’s pouch; S: striatum; T: telencephalon; Th: thalamus. Scale bar = 100 µm unless otherwise indicated.

4.4 Functional assay of *Six3* CNEs

4.4.1 Basal promoter region (includes CNE13) of mouse *Six3* is sufficient to recapitulate most aspects of expression in the forebrain and eye during early and late stages of development

I first tested the function of the promoter region alone to determine its contribution to the expression pattern of mouse *Six3* gene. A 860 bp fragment of the mouse *Six3* promoter region, spanning from -450 bp to +410 bp in relation to the transcriptional start site, was amplified and cloned upstream of a β-galactosidase reporter gene. This region of the promoter includes the highly conserved CNE13 which spans from -200 to +24 bp. This construct did not show expression of the reporter gene in E9.5 embryos. However at E10.5, it showed consistent expression in the forebrain and optic vesicle (Figure 6A). This expression pattern persisted at E11.5 (Figure 6B). This restricted expression profile is similar that of the endogenous mouse *Six3* expression at E10.5 and E11.5 stages and indicates that the 860-bp promoter region is capable of recapitulating the expression pattern of mouse *Six3* at stages E10.5 and E11.5. At later stages of E13.5 and E15.5, expression of the transgene was detected in the striatum of the telencephalon (Figure 6C and 6G); the hypothalamus and thalamus of the diencephalon (Figure 6D and 6H); the neural retina where the signal was localized to the inner neuroblastic layer, the lens of the eye (Figure 5E and 5I); the olfactory epithelium of the nasal cavity (Figure 6F); as well as the midbrain tegmentum (Figure 6J). In addition, ectopic expression was observed in the hindbrain (Figure 6J). Overall, this construct directed expression in all the domains in which mouse *Six3* shows expression except in the Rathke's pouch (Figure 6D). These results indicate that the 860-bp promoter region is capable of reproducing expression of *Six3* in almost all domains in the differentiated forebrain during late embryo development. Thus the 860-bp promoter region alone seems to be capable of reproducing expression pattern of *Six3* in the early and late stages of development of mouse embryos.




C
S
T
E13.5
D
HT
Th
RC
E13.5
E
NR
L
E13.5
F
OE
E13.5

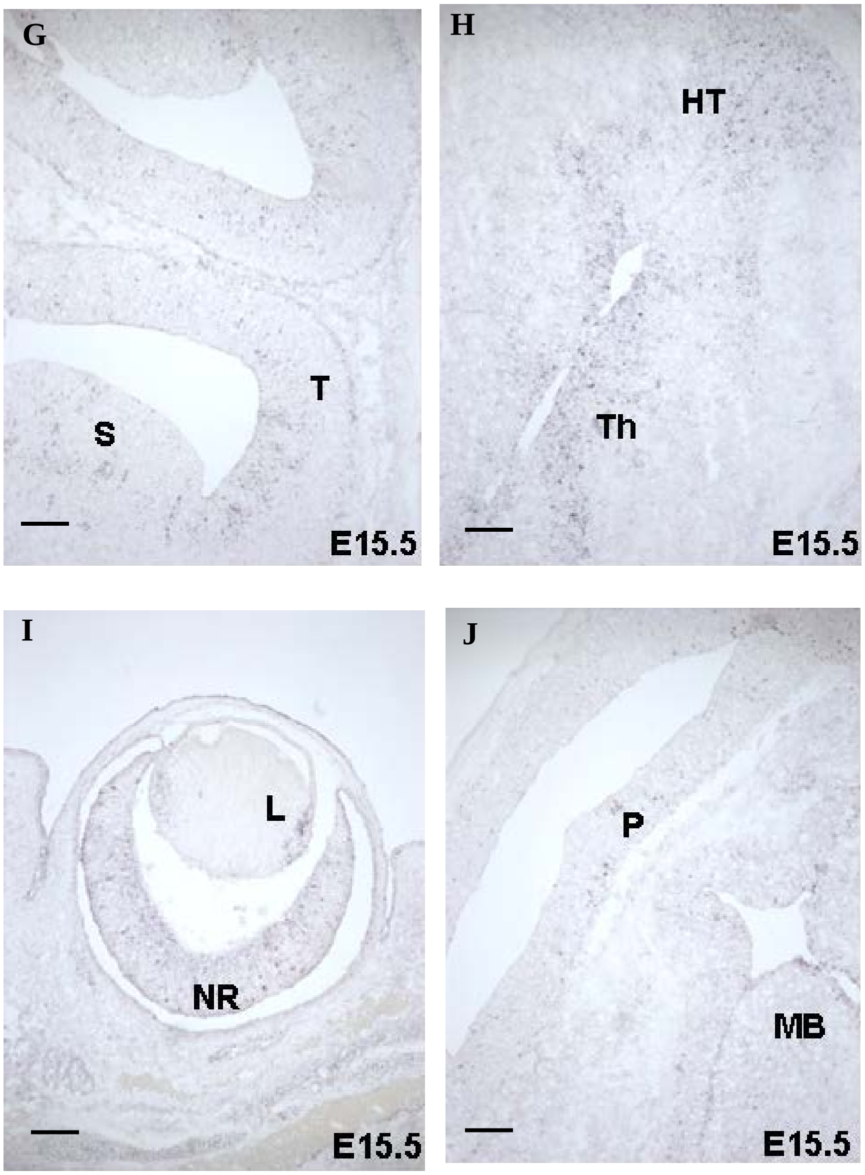


Figure 6: **A 860-bp promoter region of mouse *Six3* directs expression of *lacZ* mRNA to the forebrain and eye during embryonic development**. (A-B) Whole mount in situ hybridization of transgenic embryos show that the 860-bp promoter region directs expression of reporter gene in the forebrain and optic vesicle at E10.5 (A) and at E11.5 (B). (C-J) In situ hybridization of cryosections of the head of transgenic embryos at E13.5 (C-F) and E15.5 (G-J) show that the 860-bp promoter region directs *lacZ* expression to the telencephalon (C, G); the hypothalamus and thalamus of the diencephalon (D, H); the neural retina and lens of the differentiated eye (E, I); and the olfactory epithelium of the nose (F). Ectopic transgene expression was also detected in the hindbrain (J). D: diencephalon; E: eye; FB: forebrain; HT: hypothalamus; L: lens; MB: midbrain; NR: neural retina; OE: olfactory epithelium; OV: optic vesicle; P: Pons; S: striatum; T: telencephalon; Th: thalamus. Scale bar = 100 µm unless otherwise indicated.

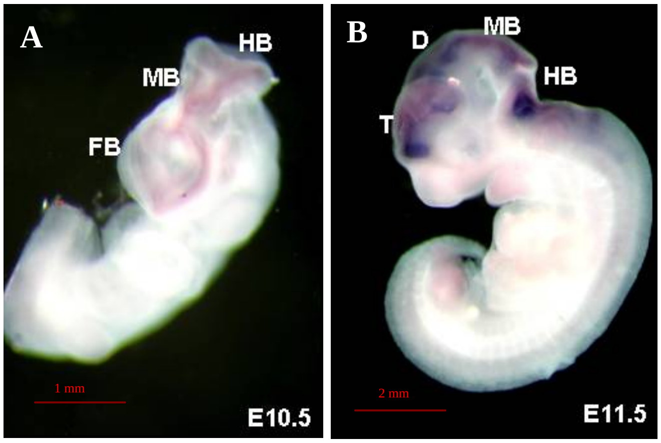
4.4.2 Expression patterns directed by CNE1, CNE2/3/4 and CNE5/6/7

CNE1 spans 80 bp and was amplified as a 220 bp fragment and cloned upstream of the mouse *Six3* promoter construct. CNE2/3/4 was made up of three conserved fragments spanning a total of 203 bp and were amplified together as a 480 bp fragment and cloned upstream of the mouse *Six3* promoter. CNE5/6/7 comprised 3 conserved sequences spanning almost 500 bp of genomic sequence, and was cloned as a 740 bp fragment upstream of the promoter. Expression of these transgenes was then individually assayed at the same time points using *lacZ* RNA probes.

At E9.5 no transgene expression was detected for CNE1. However at E10.5, *lacZ* mRNA expression directed by CNE1 and the promoter was found to express in the forebrain (Figure 7A), with similar intensity to the endogenous gene expression. Expression however was more intense in the midbrain. In addition, ectopic expression was found in the hindbrain (Compare Figure 7A with Figure 5B), and these domains of expression intensified at E11.5 (Figure 7B). No expression was detected in the optic vesicle at these stages. Therefore the effect of CNE1 on the promoter was to inhibit optic vesicle expression at E10.5 and E11.5. Similarly for CNE2/3/4, no expression was detected in transgenic embryos at E9.5. At E10.5, expression of the transgene was seen in the forebrain and the optic vesicle, as was observed with the “promoter alone” construct (Figure 7C). However at E11.5, expression in the optic vesicle was abolished while that in the forebrain intensified (Figure 7D). This showed CNE2/3/4 had no effect on the promoter at E10.5 and acted only at E11.5 to repress expression in the optic vesicle. For CNE5/6/7, there was no signal detected at E9.5. The expression at E10.5 was similar to that of the promoter alone, where expression was detected in the forebrain and optic vesicle. However at E11.5, expression was abolished in the optic vesicle, and ectopic expression was detected in the midbrain and hindbrain (Figure 7E), which was also observed for CNE1 at that particular stage of development. Therefore the effect of CNE5/6/7 on the promoter was similar to that of CNE1 in silencing optic vesicle expression but it occurred only at a single time point of E11.5.

At later stages of E13.5 and E15.5, all three constructs (CNE1, CNE2/3/4 and CNE5/6/7) directed expression to the differentiated telencephalon; hypothalamus and thalamus; neural retina and lens; olfactory epithelium; midbrain tegmentum; as well as ectopically in the hindbrain, indicating CNE1 and,CNE2/3/4 had no observable phenotypic effect on the promoter at these stages. However it was noticeable that by E15.5, lens expression in CNE5/6/7 positive embryos (Figure 7F) had diminished considerably as compared to the expression in the lens driven by the basal promoter construct (Figure 6I). CNE1 and CNE2/3/4 therefore had no effect on the promoter while CNE5/6/7 seemed to function as a silencer of lens expression during the later stages of embryonic development.

CNE1



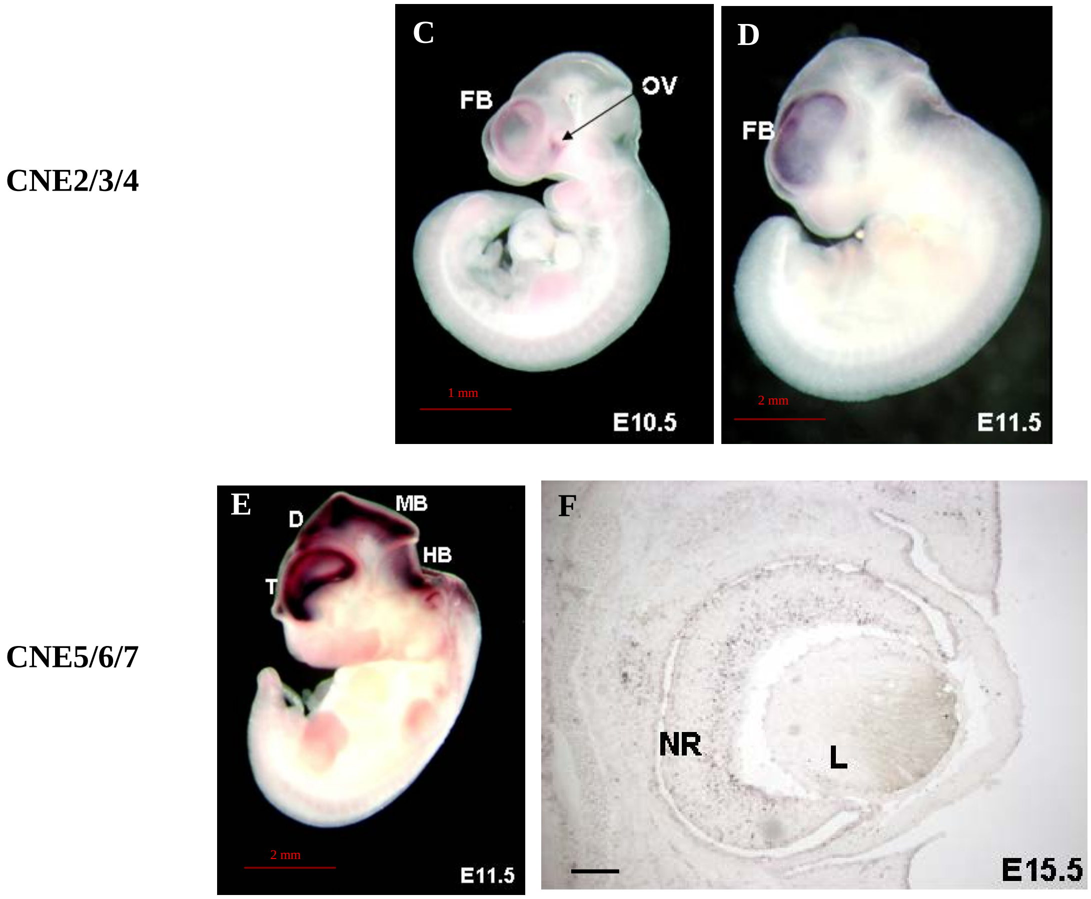


Figure 7: **Expression patterns directed by CNE1, CNE2/3/4 and CNE5/6/7.** (A-E) Whole mount in situ hybridization stained transgenic embryos that show *lacZ* mRNA labeling. CNE1 drove *lacZ* mRNA expression to the forebrain, midbrain and ectopically in the hindbrain, and inhibited optic vesicle expression at E10.5 (A) and E11.5 (B). CNE2/3/4 had no effect on the promoter-driven expression at E10.5 (C), but at E11.5 it inhibited expression in the eye while maintaining expression in the forebrain (D). CNE5/6/7 directed expression in the same way as CNE1 but only at the particular stage of E11.5 (E). (F) Coronal section of the head of a CNE5/6/7 positive embryo showed inhibition of lens expression at E15.5. D: diencephalon; FB: forebrain; HB: hindbrain; L: lens; .MB: midbrain; NR: neural retina; OV: optic vesicle; T: telencephalon. Scale bar = 100 µm unless otherwise indicated.

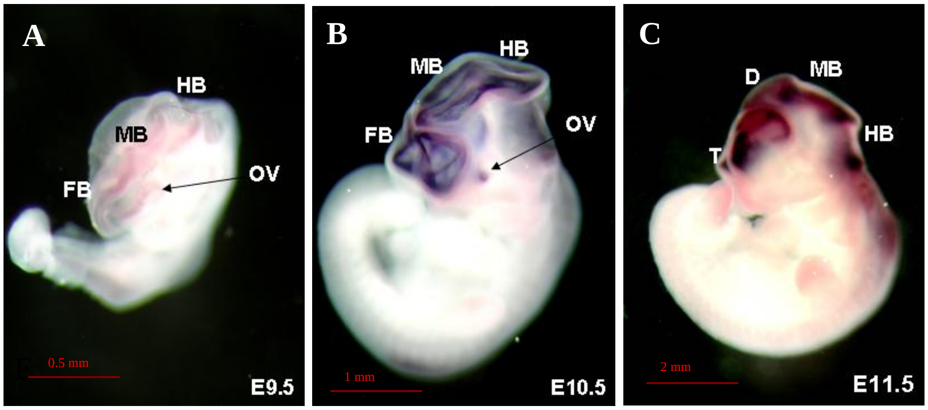
4.4.3 Expression patterns directed by CNE8/9 and CNE12

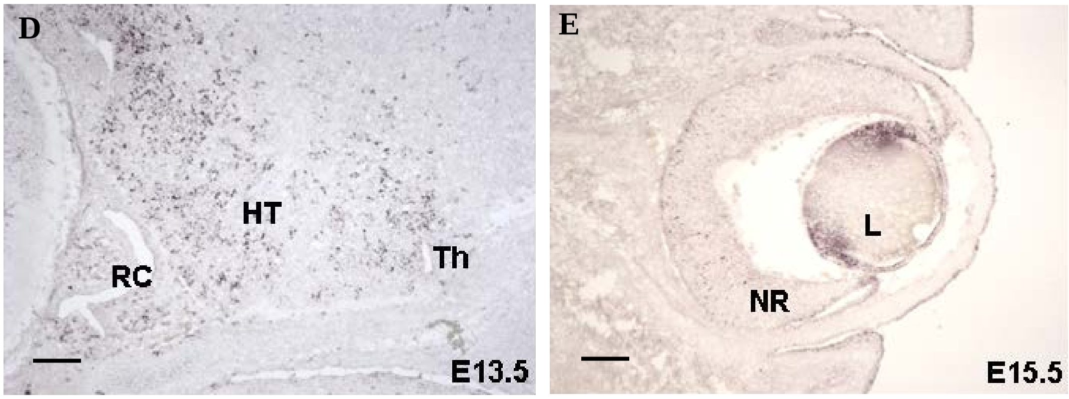
CNE8/9 spanning 473 bp of genomic DNA had one of the longest sequences conserved between fugu and mammals. This was amplified as a 610 bp fragment and cloned upstream of the mouse *Six3* promoter, and tested for enhancer activity in transgenic mice embryos. Expression of the transgene was detected as early as E9.5 in the forebrain and optic vesicle, similar to the endogenous gene expression (Figure 8A). However there was more intense expression in the midbrain and ectopic expression in the hindbrain as well (Figure 8A) and expression at these domains intensified at E10.5 (Figure 8B). However at E11.5, optic vesicle expression was abolished while expression remained in the forebrain, midbrain and ectopically in the hindbrain (Figure 8C). This suggested CNE8/9 acted as both an enhancer and a silencer of the basal promoter: it activated forebrain and optic vesicle expression at E9.5 to recapitulate the endogenous gene expression, and it silenced eye expression at E11.5. At later stages of development (E13.5-E15.5), eye expression was restored while forebrain and midbrain expression were maintained together with ectopic expression in the hindbrain region, similar to the expression profile observed with that of the basal promoter construct. However there were two important differences. Firstly, CNE8/9 also directed expression to the Rathke's pouch (Figure 8D) that was not specified by the promoter alone, and secondly retinal expression at E15.5 was severely diminished (Figure 8E) as compared to that directed by the promoter alone. This suggested that CNE8/9 also acted as a Rathke's pouch enhancer and a retina silencer in the later stages of embryo development.

CNE12 was much smaller in size, comprising of a single conserved sequence of 88 bp that was cloned as a 290 bp fragment upstream of the mouse *Six3* promoter. Its enhancer potential was assayed in transgenic mice embryos. CNE12 together with the basal promoter acted early at E9.5 to drive *lacZ* mRNA expression to the forebrain, midbrain and optic vesicle (Figure 8F). This expression persisted at relatively high levels at E10.5

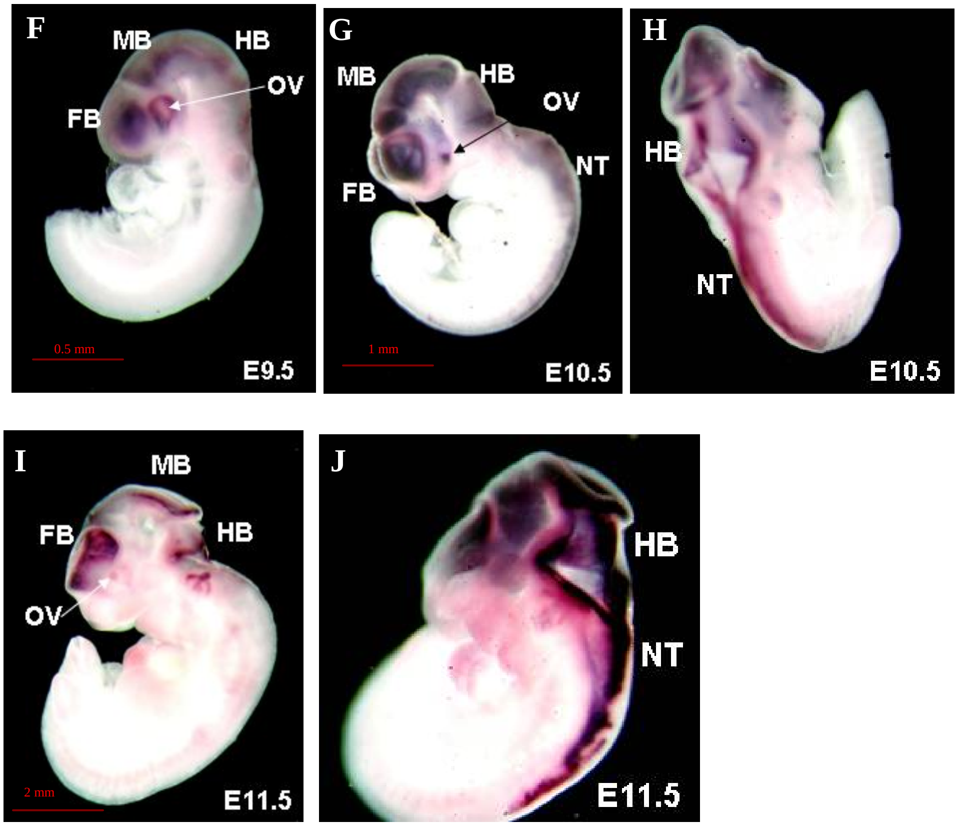
(Figure 8G) and E11.5 (Figure 8H). In addition, there was strong ectopic expression in the hindbrain and neural tube during these stages (Figure 8I and 8J). Expression in the differentiated forebrain, midbrain and ectopically in the hindbrain persisted through to E15.5, and included the Rathke's pouch (Figure 8K). However at E13.5, retinal expression was abolished (Figure 8L) while at E15.5, retinal expression was restored but scattered throughout the retina and was not localized to the inner neuroblastic layer like with the endogenous gene expression, and lens expression was abolished (Figure 8M). Therefore CNE12 was similar to CNE8/9 in being an enhancer that directed expression to the forebrain, midbrain and optic vesicle from E9.5, as well as an enhancer to direct expression in the Rathke's pouch at later stages of development. However it also functioned as a silencer of expression in both the retina and lens at E13.5 and E15.5 respectively and therefore differed with CNE8/9 in the spatial and temporal inhibition of eye expression.

CNE8/9




D
HT
Th
RC
E13.5
E
L
NR
E15.5

CNE12


F
MB
HB
OV
FB
0.5 mm
E9.5
G
MB
HB
OV
NT
FB
1 mm
E10.5
H
HB
NT
E10.5
I
MB
FB
HB
OV
2 mm
E11.5
J
HB
NT
E11.5

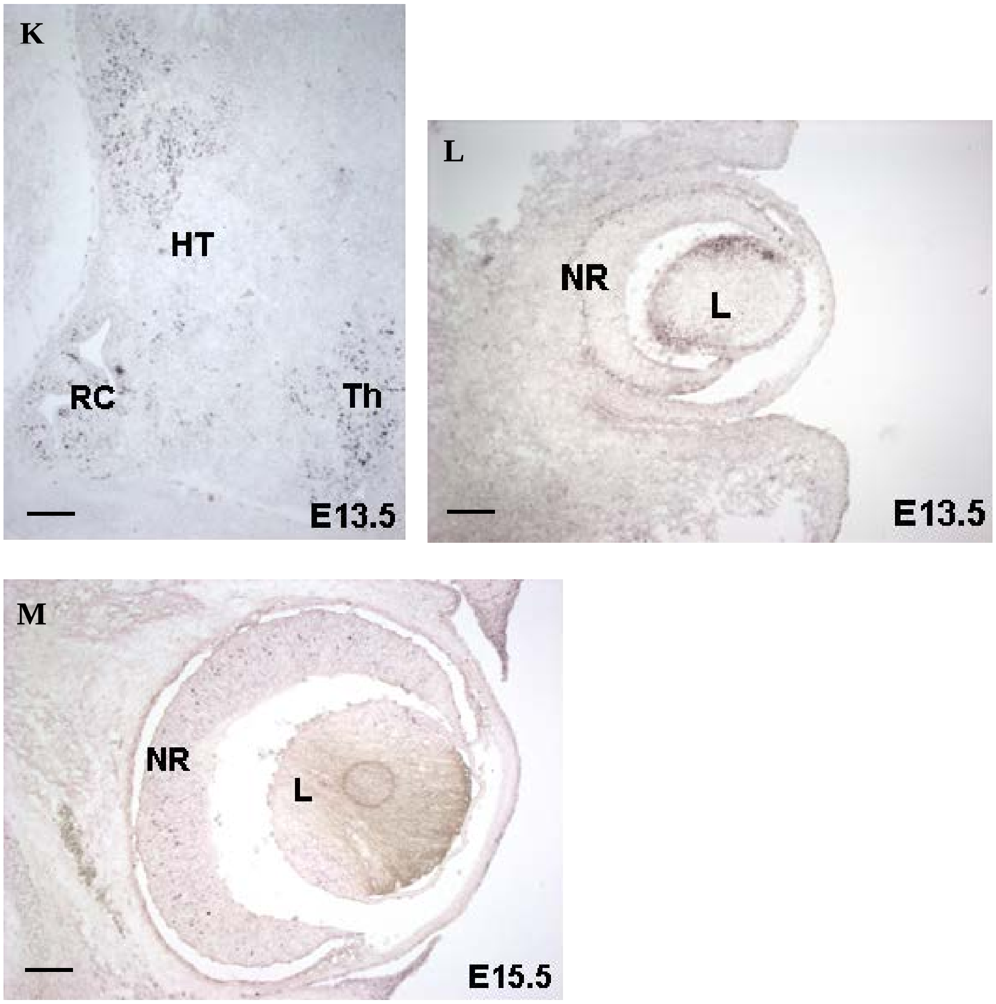


Figure 8: **Expression patterns directed by CNE8/9 and CNE12.** (A-C) Whole mount in situ hybridization stained E9.5 (A), E10.5 (B) and E11.5(C) CNE8/9 positive embryos showed *lacZ* mRNA labeling in the forebrain, midbrain, optic vesicle (A, B only) and ectopically in the hindbrain. (D) Sagittal section of the head of an e13.5 CNE8/9 positive embryo showed *lacZ* mRNA labeling in the Rathke's pouch. (E) Coronal section of the head of an E15.5 CNE8/9 transgenic embryo showed diminished labeling in the neural retina. (F-J) Whole-mount in situ hybridization stained E9.5 (F), E10.5 (G, H) and E11.5 (I, J) CNE12 positive embryos showed *lacZ* mRNA labeling in the forebrain, midbrain, optic vesicle and ectopically in the hindbrain and neural tube. (K) Sagittal section of the head of an e13.5 CNE12 positive embryo showed *lacZ* mRNA labeling in the Rathke's pouch. (L, M) Coronal section of the head of CNE12 positive embryos showed diminished labeling in the neural retina at E13.5 (L) and abolished lens expression at E15.5 (M). D: diencephalon; FB: forebrain; HB: hindbrain; HT: hypothalamus; L: lens; MB: midbrain; NR: neural retina; NT: neural tube; OV: optic vesicle; RC: Rathke's pouch; T: telencephalon; Th: thalamus. Scale bar = 100 µm unless otherwise indicated.

4.4.4 CNE10/11 silences the mouse *Six3* promoter at all developmental stages

CNE10/11 consisted of two conserved peaks spanning 150 bp that were amplified as a single fragment of size 445 bp and cloned upstream of the mouse *Six3* promoter. Mouse embryos containing the CNE5-promoter construct were harvested at different time points as before, and *lacZ* mRNA was detected using whole-mount in situ hybridization or mouse embryo heads were sectioned prior to hybridization with *lacZ* RNA probes. In total I analyzed 33 transgenics including more than 3 transgenic embryos per time point, and remarkably no *lacZ* mRNA labeling was detected in any tissue at any of the time points. Therefore CNE10/11 seemed to function as a complete silencer that repressed expression by the mouse *Six3* promoter in all tissues and at all stages of embryonic development.

4.4.5 Expression pattern directed by CNE14

CNE14 was the least conserved sequence spanning 78 bp and amplified as a 340 bp fragment. At E9.5, CNE14 directed *lacZ* mRNA expression to the early forebrain and optic vesicle of the mouse embryo (Figure 9A), consistent with the endogenous gene expression. CNE14 then directed ectopic *lacZ* mRNA expression to the hindbrain and inhibited expression in the optic vesicle at E10.5 (Figure 9B). At E11.5, CNE14 restricted expression to only the forebrain (Figure 9C). Therefore the effects of CNE14 on the promoter were the most dynamic. It acted as an enhancer to direct expression to the forebrain and optic vesicle at E9.5. In addition it silenced optic vesicle expression at E10.5 and E11.5 as well as mediated ectopic expression to the hindbrain at E10.5.

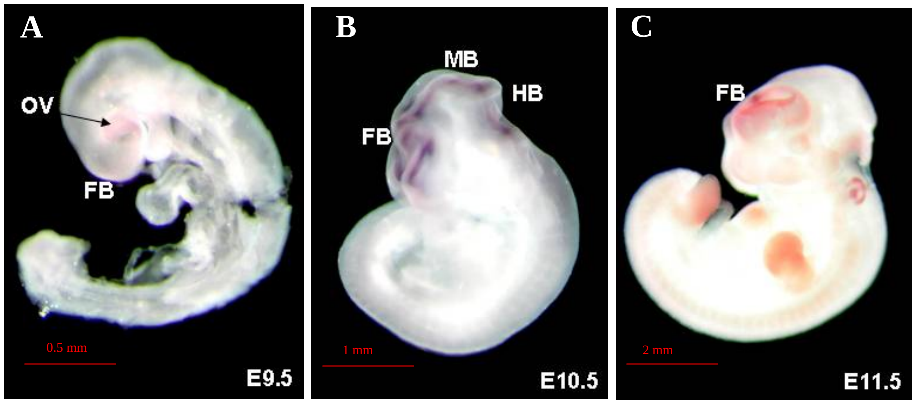


Figure 9: **Expression pattern directed by CNE14 at E9.5-E11.5**. (A-C) Whole mount *in situ* hybridization stained E9.5 (A), E10.5 (B) and E11.5(C) CNE14 positive embryos showedCNE14 enhanced expression of *lacZ* mRNA in the forebrain and optic vesicle at E9.5 (A). At E10.5, CNE14 inhibited optic vesicle expression and activated ectopic expression in the hindbrain (B). At E11.5, ectopic expression was abolished but inhibition of expression in the optic vesicle persisted (C). FB: forebrain; HB: hindbrain; MB: midbrain; OV: optic vesicle.

4.4.6 Summary of the regulatory potential of mouse *Six3* CNEs

The CNEs I analyzed from the mouse *Six3* locus all functioned as tissue-specific enhancers and / or silencers in transgenic mouse embryos. CNE13 and its flanking sequences likely represent the basal promoter that directs expression to almost all domains of the endogenous mouse *Six3* expression in both the early and late stages of embryo development, while the other CNEs work in concert with the basal promoter to silence or enhance particular domains of expression at particular stages of development. Table 4 provides a summary of the effect of each CNE on the expression pattern driven by the mouse *Six3* promoter and clearly shows their overlapping roles in modulating the basal promoter.

CNE	FB	HB	Optic vesicle	Others
13-P	Enhancer from E10.5	Ectopic expression from E13.5	Enhancer from E10.5	Enhancer of T, HT, Th, OE, NR, L in the differentiated FB and eye
13-P +1		Ectopic expression at E10.5-E11.5	Silencer at e10.5-E11.5	
13-P +2/3/4			Silencer at E11.5	
13-P +5/6/7		Ectopic expression at E11.5 only	Silencer at e11.5 and E15.5 (lens)	
13-P +8/9	Enhancer at E9.5	Ectopic expression at E9.5-e11.5	Silencer at e11.5 and E15.5 (retina)	Enhancer of Rathke's pouch at E13.5
13-P +10/11	Silencer at all developmental stages			
13-P +12	Enhancer at E9.5	Ectopic expression at E9.5-E11.5 including neural tube	Silencer at E13.5 (retina) and at E15.5 (lens)	Enhancer of Rathke's pouch at E13.5
13-P +14	Enhancer at E9.5	Ectopic expression at E10.5 only	Enhancer at e9.5; silencer at E10.5-E11.5	

Table 4: **Enhancer function of mouse *Six3* CNEs across different developmental stages and in different tissues.** CNE13 is part of the basal promoter tested. Other CNEs were cloned upstream of the basal promoter construct (CNE13-P) and expression pattern was assayed using in situ hybridization that detected *lacZ* mRNA labeling and compared with that driven by the basal promoter alone. The differences in the expression pattern conferred by each CNE are then tabulated. FB: forebrain; HB: hindbrain; HT: hypothalamus; L: lens; NR: neural retina; OE: olfactory epithelium; T: telencephalon; Th: thalamus.

4.5 Discussion

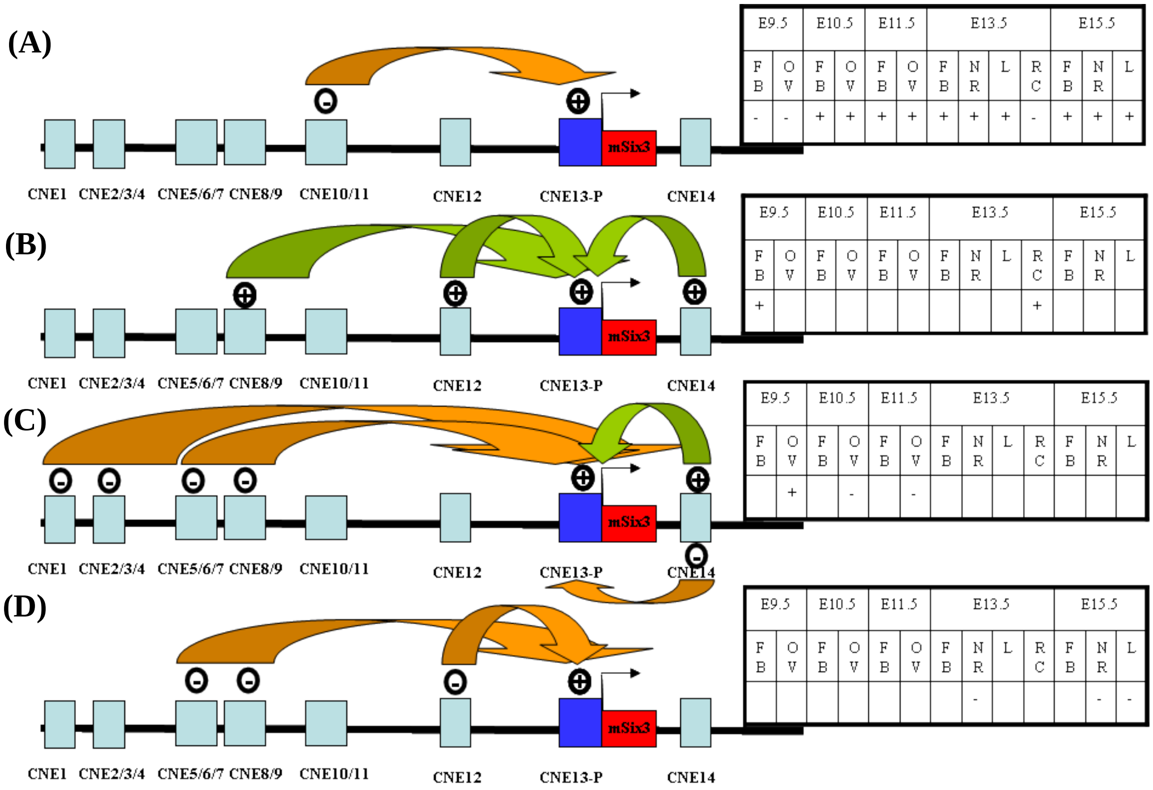
Six3 is an important regulator of vertebrate forebrain development and defining the precise gene regulatory network that controls its spatiotemporal expression would help in elucidating the mechanisms by which it regulates forebrain and eye development. This study has allowed a better understanding of the transcriptional mechanisms responsible for regulating *Six3* expression in a tissue and developmental-stage specific manner. Firstly, CNE13 together with the flanking sequences representing the basal promoter was able to specify forebrain and optic vesicle expression from E10.5, maintain it at E11.5, and by later stages (E13.5-E15.5) direct expression to the differentiated forebrain (telencephalon, striatum, hypothalamus and thalamus), midbrain tegmentum, eye (neural retina and lens), olfactory epithelium. This meant the basal promoter could recapitulate the majority of the endogenous mouse *Six3* expression domains. However, it is not clear if this expression pattern is due to the CNE13 sequence or the sequences flanking it in the basal promoter construct tested or a combination of both, although the high level of conservation of CNE13 suggests that it is most likely playing a role in the expression patterns observed with this construct. In the early developmental stages of E10.5-E11.5, the basal promoter alone was sufficient to reproduce endogenous gene expression levels in the forebrain and optic vesicles. However in the later stages of development, the promoter was insufficient to direct expression to Rathke's pouch, and it did not down-regulate expression in the hindbrain, as according to the endogenous gene expression level. A silencer located in this locus is likely to be involved in suppressing the expression of the basal promoter in the hindbrain during development to ensure the correct expression of *Six3* during development.

Secondly the remaining CNEs in the *Six3* locus acted as enhancers and/or silencers with overlapping functions. Having determined the function of the basal promoter in directing transgene expression at several time points, the function of individual enhancers that were 'added on' to the basal promoter construct can be deduced by a comparison of expression patterns resulting from the effect of individual CNEs on the basal promoter, with that observed due solely to the action of the basal promoter. I found at least three early enhancers (CNE8/9, CNE12 and CNE14) that directed expression to the forebrain and optic vesicle at E9.5, reproducing endogenous gene expression at that stage and this complemented the action of the promoter from E10.5. There were at least 5 enhancers that gave rise to ectopic expression in the hindbrain at levels greater than that seen for the endogenous gene expression and at earlier stages of development compared to the basal promoter (see Figures 7B, 7E, 8C, 8H, 8J and 9B). This ‘leaky’ expression was present either early from E9.5 (CNE8/9 and CNE12) or only at a particular time point (CNE1, CNE5/6/7 and CNE14). It would appear the activity of some of the identified regulatory elements was not properly turned off when these elements were studied outside their normal genomic context. In the genomic context of the *Six3* locus, these enhancers would have had to be modulated or repressed by other *cis*- and/or *trans*-acting agents, so as not to give rise to ‘leaky’ expression levels in the hindbrain. Strangely, many of these enhancers functioned simultaneously as silencers of expression in the optic vesicle and/or in the differentiated retina or lens, either over a few developmental stages (CNE1, CNE8/9 and CNE12) or at a particular time point (CNE2/3/4, CNE5/6/7 and CNE14). CNE5/6/7, CNE8/9 and CNE12 were the only three enhancers that helped to modulate the action of the endogenous promoter at later developmental stages, by firstly specifying expression in Rathke’s pouch (CNE8/9 and CNE12) as well as inhibiting expression in the differentiated retina (CNE8/9 and CNE12) and lens (CNE5/6/7 and CNE12). Being one of the primary domains of *Six3* expression, it was surprising to find the presence of many silencer elements of eye expression. It is likely that in the context of the whole locus, these repressors worked in combination to modulate the level of *Six3* expression and were required to help maintain a physiologically appropriate dosage of *Six3* expression in the eye at all times.

4.5.1 Comparison of results from *Six3* regulation in medaka

The *cis*-regulatory elements required for *Six3* expression in medaka was found to be contained in a 4.5 kb genomic region upstream of the transcription start site of the medaka locus. They comprised two enhancer modules that directed early and late stages of brain development, two silencer modules and two silencer blocker modules, which together control *Six3* expression in the lens and early retina of medaka (Conte and Bovolenta, 2007). Unfortunately the regulatory organization of the medaka *Six3* locus was poorly conserved in vertebrates other than fishes. Out of the ten conserved blocks in fish, only two were highly conserved in human and mouse (Conte and Bovolenta, 2007), and were promptly detected in my MALGAN analysis. However, one of them (block G) matched the opposite strand transcript sequence of mouse *Six3* in the NCBI EST database and was excluded from further analysis. Since this sequence in medaka (block G) functioned as a repressor of the late brain enhancer (block I) and is know to code for a transcript in mouse, it is possible that this sequence may code for a RNA gene that itself may be acting as a silencer. The other conserved module (block L) overlaps with CNE13 in the *Six3* basal promoter region, but unlike the basal promoter construct that directed reporter expression to almost all the mouse *Six3*-expressiong domains in both the early and late stages of development, it had no detectable function in directing reporter expression in transgenic medaka (Conte and Bovolenta, 2007). The difference in the function could be due to the additional sequences flanking CNE13 in the basal promoter construct (CNE13-P). While the conserved module in medaka is only 224 bp, the basal promoter construct (860 bp long) tested by me contained about 240 bp region upstream and 400 bp downstream of the medaka conserved module. Another possibility is that the conserved module may be differentially utilized in medaka and mice, and that *Six3* may be regulated differently in fishes and mammals. In medaka, each conserved module played a unique role as an enhancer or silencer or blocker in specific tissues with little overlap in function between the modules, while in mouse the basal promoter has taken on the function of primary enhancer in directing *Six3* expression in almost all the required domains while the other conserved modules (CNEs) spread over a large region of 163 kb act in concert with the basal promoter to direct specific expression by modulating the promoter activity spatially or temporally and these enhancers show redundancy in their roles as secondary enhancers or silencers. I could not detect silencer-blocker activity as I tested the CNEs individually and only in conjunction with the promoter. The presence of CNE10/11 as a complete silencer however would suggest the need for silencer-blockers to neutralize this silencer when physiologically appropriate levels and correct domains of *Six3* expression were required to be activated during the development of the forebrain and eye, and it is likely some of the CNEs I tested would have this silencer-blocker activity. This can be confirmed by testing CNEs in various combinations in transgenic mice.

My study has shown that the regulation of *Six3* in vertebrates is more complex than previously thought based on the identification of *cis*-regulatory elements in medaka. The *cis*-regulatory elements are actually spread over much larger region than the proximal 4.5 kb characterized by Conte and Bovolenta (2007) in medaka. Since the CNEs identified in my study are conserved in fugu, they are most likely to be present in medaka either in the sequences upstream of 4.5 kb proximal promoter sequences in *olSix3.2* or in the 5' and 3' flanking regions of its duplicate copy *olSix3.1* that remain to be characterized. Altogether the data I present here provide a more comprehensive picture of the regulatory code that governs *Six3* expression during the development of the forebrain and eye in vertebrates. The regulatory code as revealed by the transgenic mice reporter gene assay is summarized in Figure 10. Firstly the basal promoter alone directs expression to the forebrain and optic vesicle in the early stages of development from E10.5; as well as to most of the *Six3*-expressing domains in the differentiated forebrain and the differentiated eye during the later stages of development (Figure 10A). This action of the promoter is dependent upon the silencer activity of CNE10/11. Secondly the other CNEs act to modulate the basal promoter. For example *Six3* expression in the early forebrain at E9.5 is mediated by at least three enhancers (CNE8/9, CNE12 and CNE14), two of which also mediate expression to Rathke's pouch (Figure 10B). In addition most of the CNEs function as silencers of expression in the optic vesicle (Figure 10C) and in the retina and lens of the developing eye (Figure 10D) to help keep the expression of *Six3* in the eye under strict control at all times. Thus, spatio-temporal code of *Six3* regulation is provided by the combined use of at least 14 different modules, all conserved in fish and mammals, with distinct and overlapping roles as enhancers and silencers, but all working in concert to modulate the basal promoter.



E9.5		E10.5		E11.5		E13.5				E15.5		
FB	OV	FB	OV	FB	OV	FB	NR	L	RC	FB	NR	L
-	-	+	+	+	+	+	+	+	-	+	+	+

E9.5		E10.5		E11.5		E13.5				E15.5		
FB	OV	FB	OV	FB	OV	FB	NR	L	RC	FB	NR	L
+									+			

E9.5		E10.5		E11.5		E13.5				E15.5		
FB	OV	FB	OV	FB	OV	FB	NR	L	RC	FB	NR	L
	+		-		-							

E9.5		E10.5		E11.5		E13.5				E15.5		
FB	OV	FB	OV	FB	OV	FB	NR	L	RC	FB	NR	L
							-				-	-

Figure 10: Summary of the regulatory code that controls the expression of *Six3* in mouse. (A) The basal promoter alone directs expression to the forebrain and optic vesicle in the early stages of development from E10.5; as well as to most of the *Six3*-expressing domains in the differentiated forebrain and the neural retina and lens of the differentiated eye during the later stages of development. This action of the promoter is dependent upon the silencer activity of CNE10/11 at all stages of development. (B) *Six3* expression in the early forebrain at E9.5 is mediated by at least three enhancers (CNE8/9, CNE12 and CNE14), the former two of which also mediate expression to Rathke's pouch. (C)Expression of *Six3* in the optic vesicle during early development is mediated by at least one enhancer (CNE14) that activates expression at e9.5 and at least 5 silencers that repress expression at E10.5-E11.5. (D) Late eye expression of *Six3* is in turn mediated by at least three silencers (CNE5/6/7, CNE8/9 and CNE12) that repress expression in either the lens or the neural retina at E13.5 or E15.5. FB: forebrain; L: lens; NR: neural retina; OV: optic vesicle; RC: Rathke's pouch.

Chapter 5

Results

Regulation of *Foxb1*

5.1 Introduction

The forkhead (Fox) gene family encodes more than one hundred transcription factors, each characterized by a “winged helix” configuration in their DNA-binding domain. These transcription factors play key roles in development, metabolism, aging, cancer and immunoregulation (Lehmann et al., 2003). *Foxb1* was originally described under the names of *HFH-5.1* and *Fkh5* as an early expressing gene in the brain as well as in the neural plate and early mesoderm in primitive streak stage embryos (Ang et al., 1993; Kaestner et al., 1996). By midgestation in mouse, *Foxb1* is restricted to specific populations of cells in the thalamus and hypothalamus, midbrain tegmentum, hindbrain and spinal cord; as well as in the mammary gland epithelium (Kloetzli et al., 2001; Wehr et al., 1997). Late in gestation and in newborn mice, the predominant region of expression for *Foxb1* is the most caudal region of the hypothalamus, within the mammillary bodies, indicating its likely role in the growth and differentiation of a specific segment of the anterior forebrain (Labosky et al., 1997). *Foxb1* has also been identified in zebrafish as playing an important role in the induction and patterning of the forebrain by specifying the posterior domain of the presumptive neuroectoderm in the developing embryo through its expression in the prospective diencephalon, mesencephalon and posterior hindbrain/spinal cord, before any morphological subdivision (Grinblat et al., 1998). The similarities in the expression patterns of *Foxb1* in mouse and zebrafish suggest *Foxb1* to be an important regulator of development and maintenance of these structures, and that its function is conserved among vertebrates.

To address its function, *Foxb1* has been knocked out in a variety of ways, and the mutant phenotype in mice included increased perinatal mortality and growth retardation in the mutant embryos and pups that survive (Labosky et al., 1997); impaired differentiation of regions in the midbrain and hypothalamus that would compromise spatial memory formation (Wehr et al., 1997); impaired differentiation of neural progenitors in the spinal cord resulting in motor weakness (Dou et al., 1997); and incomplete lobuloalveolar development of the mammary glands resulting in a failure to generate the milk ejection reflex and an inability to lactate (Kloetzli et al., 2001). Therefore *Foxb1* has multiple roles to play during embryogenesis and adult life, acting as an important regulator to fine-tune development of the diencephalon, brainstem, spinal cord, mammary glands and other regions that regulate the milk-ejection reflex. To date, the regulatory mechanisms underlying the complex expression pattern of *Foxb1* have not been elucidated. In order to identify conserved *cis*-regulatory elements of *Foxb1*, I aligned the fugu *Foxb1* locus with the corresponding sequences in mouse and human, and characterized individual CNEs in transgenic mice embryos at different developmental stages.

5.2 Comparison of *Foxb1* loci in human, mouse and fugu

Identification of *Foxb1* ortholog in fugu using Ensembl Biomart annotation and INPARANOID showed that fugu contains a single ortholog. The genomic sequences for the human, mouse and fugu *Foxb1* genes were retrieved from Ensembl (see Materials and methods). Scanning of the genes located upstream and downstream of *Foxb1* in the human, mouse and fugu genome assemblies indicated that the synteny of the genes in this locus is highly conserved in the three genomes (Figure 11). The protein sequence of

Foxb1 is encoded by a single exon in human, mouse and fugu. The 5' and 3' flanking sequences of *Foxb1* span 325 kb and 350 kb in human, 230 kb and 270 kb in mouse and 24 kb and 27 kb in fugu respectively (Figure 11). Therefore as with *Six3*, *Foxb1* is situated in a vast sea of noncoding DNA and it can be a challenge to identify *cis*-regulatory elements directing expression of *Foxb1* gene in these vertebrates. The orthologous genomic sequences, comprising the entire 5' and 3' flanking regions, for each of the human, mouse and fugu genes were aligned using the global alignment algorithm MLAGAN (http://genome.lbl.gov/vista/lagan/) using fugu as the reference sequence and CNEs (>60% identity and larger than 50 bp) were predicted using VISTA.

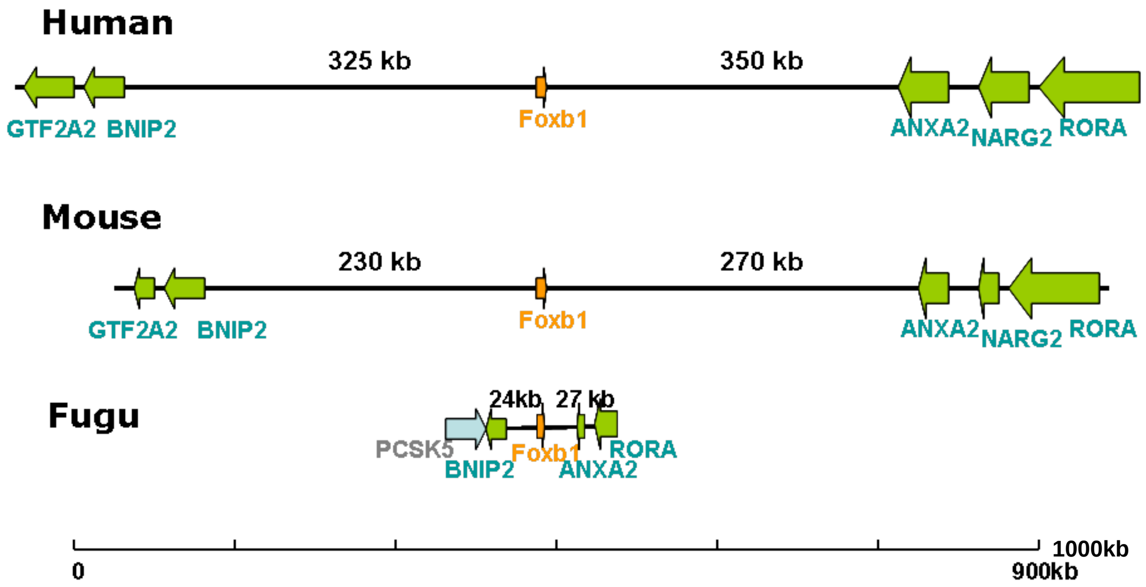


Figure 11: ***Foxb1* loci of human, mouse and fugu**. Arrows indicate the direction of transcription. *Foxb1* gene is indicated in orange while conserved syntenic genes are indicated in green.

A total of 30 mammal-fugu CNEs were predicted in the *Foxb1* loci (Figure 12). There are 13 CNEs located in the 5'-flanking region, including one CNE (235 bp) situated just 200 bp upstream of the transcription initiation site and thus overlapping with the basal promoter and 5’UTR of *Foxb1*, while 17 CNEs are located in the 3’-flanking region. In the human genome, these CNEs are spread over a region of 440 kb with the most upstream and downstream CNE located about 125 kb and 315 kb respectively from the transcription start site. Due to the large number of CNEs and the enormous time constraints on testing them individually, I decided to pick only the CNEs most proximal to the gene for testing. For the purpose of testing these CNEs in transgenic mice, CNEs located in clusters were grouped together and their combined sequences were amplified by PCR while the rest were amplified as individual CNEs. The CNEs tested included the basal promoter which includes CNE3; CNE1 and CNE2 located upstream; and CNE4 and CNE5 located downstream of the coding sequence (Figure 13). The details of the CNEs tested are given in Table 5. In all, five conserved noncoding sequences were tested in transgenic mice. I first tested the function of the basal promoter (CNE3-P) alone by cloning it upstream of a β-galactosidase reporter, and then tested the functions of other CNEs by cloning each of them upstream of the basal promoter.

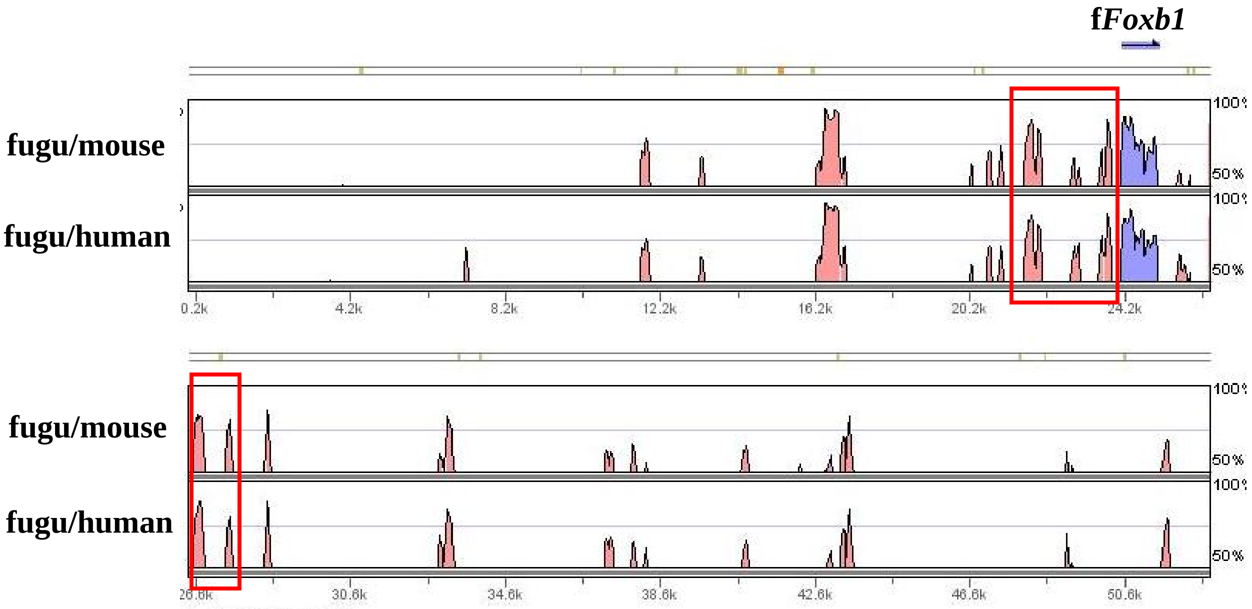


Figure 12: **Conserved noncoding elements in the *Foxb1* locus**. VISTA plot of the MLAGAN alignment of the fugu, mouse and human *Foxb1* loci is shown. Fugu was used as the base sequence. CNEs were predicted as noncoding sequences that are ≥60% identity across 50 bp or longer. Peaks represent conserved sequences; coding sequences are shaded blue and non-coding sequences (CNEs) are shaded pink. The arrow indicates the direction of transcription of *Foxb1*. There are 30 CNEs in total but only those enclosed in the red rectangles were analyzed for regulatory potential (see Figure 13 for a close-up view). X-axis represents fugu sequence and Y-axis represents percent identities (50%-100%).

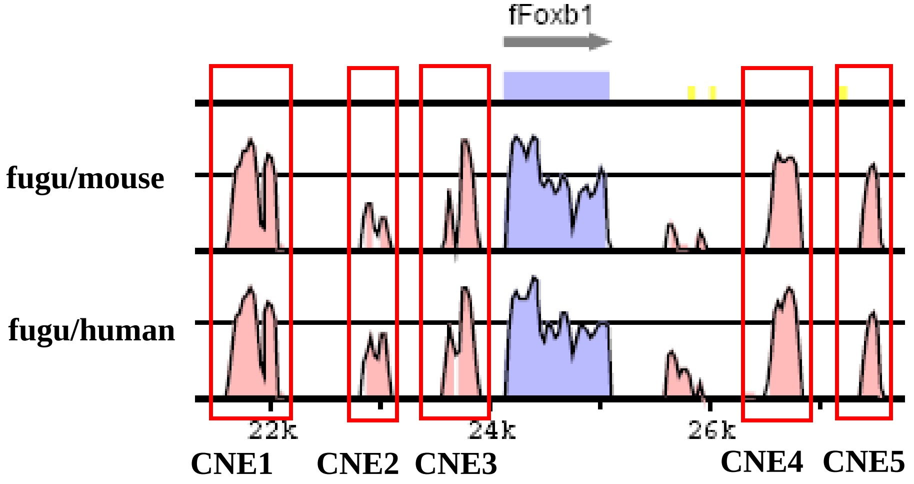


Figure 13: **CNEs selected for testing in transgenic mice.** The conserved noncoding peaks are shaded in pink and highlighted in red rectangle boxes. Three of the conserved peaks (designated CNE1-3) are located in the 5' flanking region, including one that overlaps the basal promoter and 5'UTR. The other two conserved peaks (designated CNE4-5) are in the 3' flanking region. The pink peaks outside the red boxes overlapped with NCBI EST sequences and were not counted as CNEs.

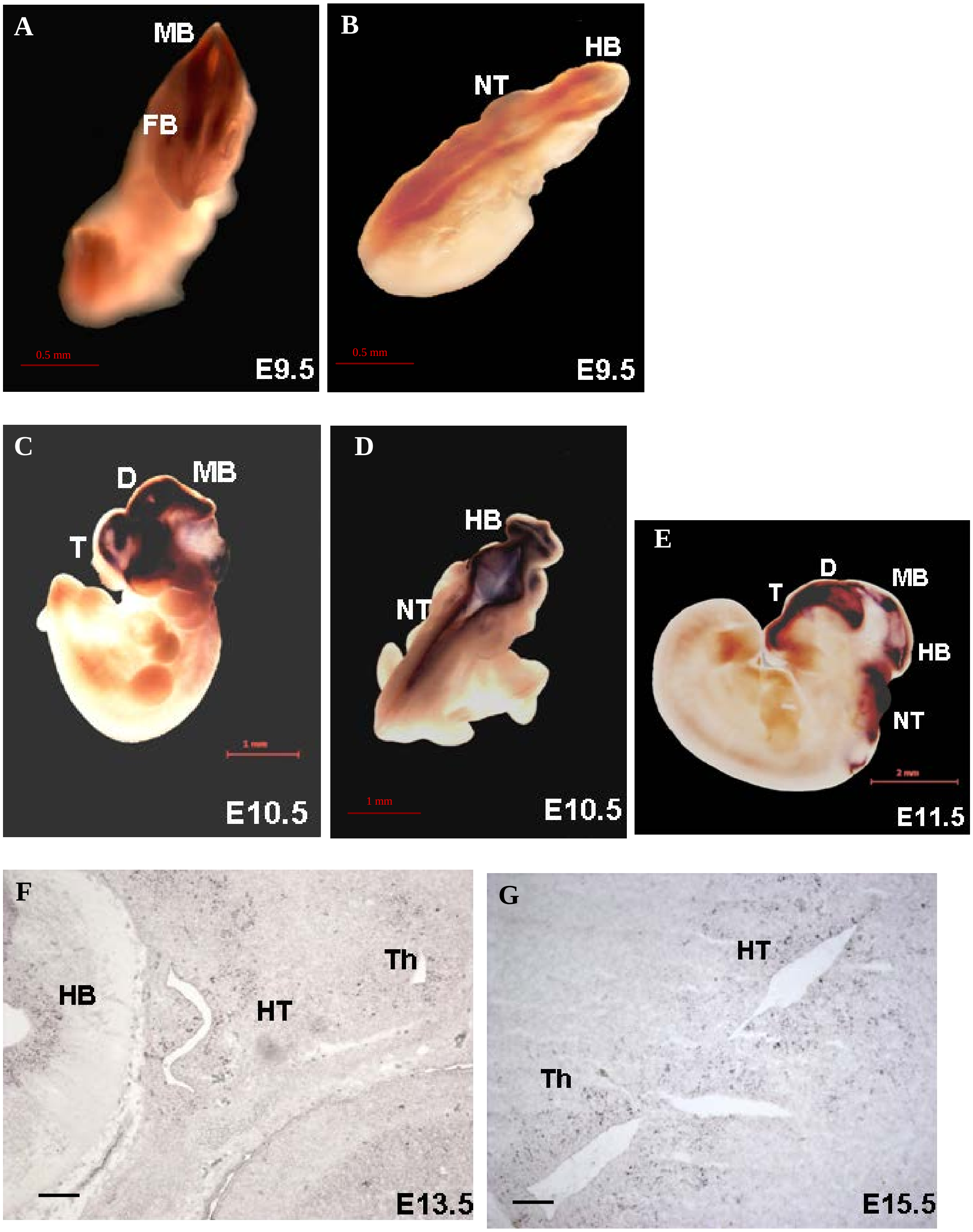
CNE	Number of CNEs merged	Combined length (bp)	% identity	Approximate distance from TSS of human *Foxb1*	Length (bp) amplified by PCR
CNE1	2	455	76.5	-6.7 kb	680
CNE2	1	235	68.1	-4 kb	290
CNE3	2	234	80.8	-200 bp	400
CNE4	1	246	82.1	+3 kb	460
CNE5	1	187	72.7	+34 kb	340

Table 5: ***Foxb1* CNEs tested in transgenic mice**. The length, percent identity and location of the CNEs are shown. The actual size of the mouse noncoding sequence amplified and cloned into a *lacZ* reporter construct is shown in the last column. TSS: transcription start site.

5.3 Expression pattern of mouse *Foxb1*

The expression pattern of *Foxb1* in the mouse embryo during the early stages of development (embryonic days E9.5 to E11.5) and the late stages of development (embryonic days E13.5 to E15.5) has been previously investigated (Labosky et al., 1997; Wehr et al., 1997). However, for the sake of comparison of the expression pattern of the mouse *Foxb1* gene with the expression pattern of the reporter gene driven by mouse *Foxb1* CNEs, I determined the expression pattern of the *Foxb1* gene in the FVB/N mouse strain used for testing the CNEs. The expression in the early stages of development (E9.5-E11.5) was analyzed by a whole-mount in situ hybridization using an antisense

RNA probe that binds specifically to a 450 bp fragment of the mouse *Foxb1* coding region and the expression during the late stages of development was investigated by *in situ* hybridization of sagittal (E13.5) or coronal (E15.5) sections of the mouse embryos using the same probe. At E9.5, the expression of *Foxb1* was detected in the presumptive diencephalon of the forebrain, midbrain, hindbrain and neural tube (Figures 14A and 14B), in agreement with the expression patterns observed in previous studies (Labosky et al., 1997; Wehr et al., 1997; Zhao et al., 2007). At E10.5, expression intensified in the same domains but has also spread to include the developing telencephalon (Figures 14C and 14D). At E11.5, expression remained in the diencephalon, midbrain, hindbrain, and neural tube (Figure 14E) with more intense staining in the differentiated telencephalon. *Foxb1* has not been known to express in the telencephalon in the handful of expression studies carried out so far, and previous staining observed in that region has been attributed to unspecific background (Wehr et al., 1997). At later stages of development (E13.5-E15.5), *Foxb1* expression was predominantly seen in the hypothalamus and thalamus of the diencephalon (Figure 14F and 14G), with the transcripts localized to the differentiated nuclei of the mammillary bodies of the posterior hypothalamus (Labosky et al., 1997; Wehr et al., 1997) or in cells of the lateral hypothalamus (Kloetzli et al., 2001). In addition expression was also observed in the midbrain tegmentum and hindbrain as previously observed by Wehr et al (1997) and Zhao et al (2007) (Figures 14H and 14I). Thus, the expression profile I observed for *Foxb1* in the hypothalamus and thalamus, the midbrain, hindbrain and spinal cord (neural tube) is generally consistent with that in the literature in both the early and late stages of development, with the mammillary bodies of the hypothalamus being the major site of expression after midgestation, and additional but fainter signals persisting in the thalamus, midbrain and hindbrain.



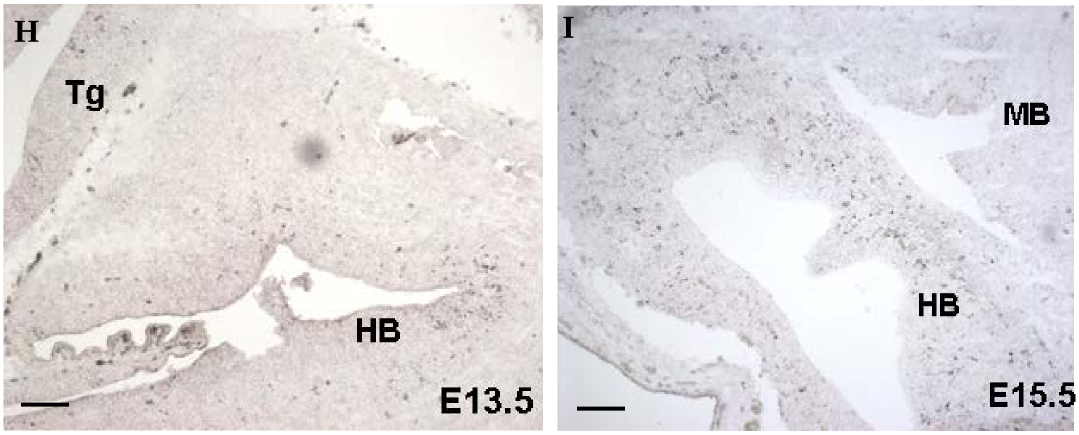


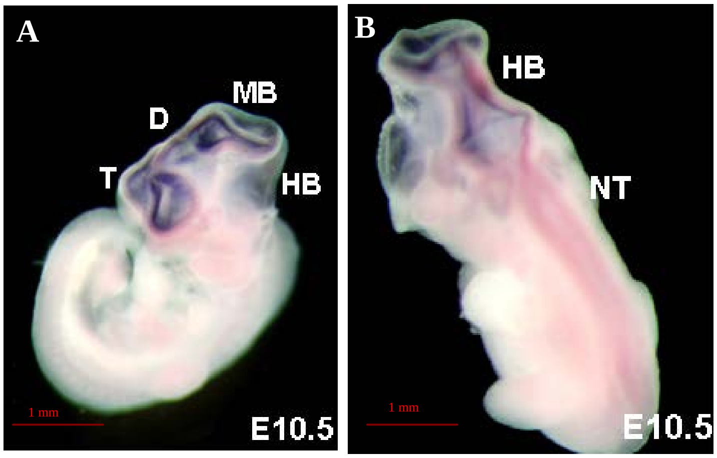
Figure 14: **Expression patterns of *Foxb1* in the developing mouse embryo**. (A-E) Whole mount in situ hybridization of wild type embryos showing expression of mouse *Foxb1*. At E9.5, mRNA accumulation was seen in the presumptive diencephalon of the forebrain, midbrain and hindbrain (A), as well as in the neural tube (B). At E10.5 expression intensified in the diencephalon, midbrain, hindbrain (C), and neural tube (D). At E11.5, expression persisted in the diencephalon, the midbrain, the hindbrain and neural tube (E). Staining in the telencephalon is due to unspecific background. (F-I) In situ hybridization of sagittal (e13.5) and coronal (e15.5) sections of the head of wild-type embryos with a 450 bp fragment of the mouse *Foxb1* exon as a probe at E13.5 (F, H) and E15.5 (G, I) show that mouse *Foxb1* expression is primarily detected in the hypothalamus and thalamus of the diencephalon (F, G); as well as more weakly in the midbrain tegmentum and hindbrain (H, I). D: diencephalon; FB: forebrain; HT: hypothalamus; MB: midbrain; Th: thalamus. Scale bar = 100 µm unless otherwise indicated.

5.4 Functional assay of *Foxb1* CNEs

5.4.1 Basal promoter region (includes CNE3) of mouse *Foxb1* is sufficient to recapitulate most aspects of endogenous expression during early and late stages of development

I first tested the function of the promoter region alone to determine its contribution to the expression pattern of mouse *Foxb1* gene. A 400 bp fragment of the mouse *Foxb1* promoter region, spanning from -250 bp to +150 bp in relation to the transcriptional start site, was amplified and cloned upstream of a β-galactosidase reporter gene. This region of the promoter includes the highly conserved CNE3-P which spans from -200 to +34 bp.

This construct did not show expression of the reporter gene in E9.5 embryos. However at E10.5, it showed reproducible expression in the prospective diencephalon, midbrain and hindbrain (Figure 15A) and this expression pattern persisted at E11.5 (Figure 15C). Ectopic expression was observed in the telencephalon at both stages. This expression profile is similar that of the endogenous mouse *Foxb1* expression at E10.5 and E11.5 stages, except that the promoter did not direct expression to the neural tube (Figures 15B and 15D). At later stages of E13.5 and E15.5, expression of the transgene was detected in the hypothalamus and thalamus of the diencephalon (Figure 15E and 15F); as well as in the midbrain tegmentum and hindbrain (Figure 15G). In addition, ectopic expression was observed in the striatum region of the telencephalon (Figure 15H). Overall, this construct directed expression in all the domains in which mouse *Foxb1* is expressed during late embryo development, but it included ectopic expression in the telencephalon. These results indicate that the 400-bp promoter region is capable of reproducing expression of *Foxb1* in almost all domains in the early and late stages of development of mouse embryos.



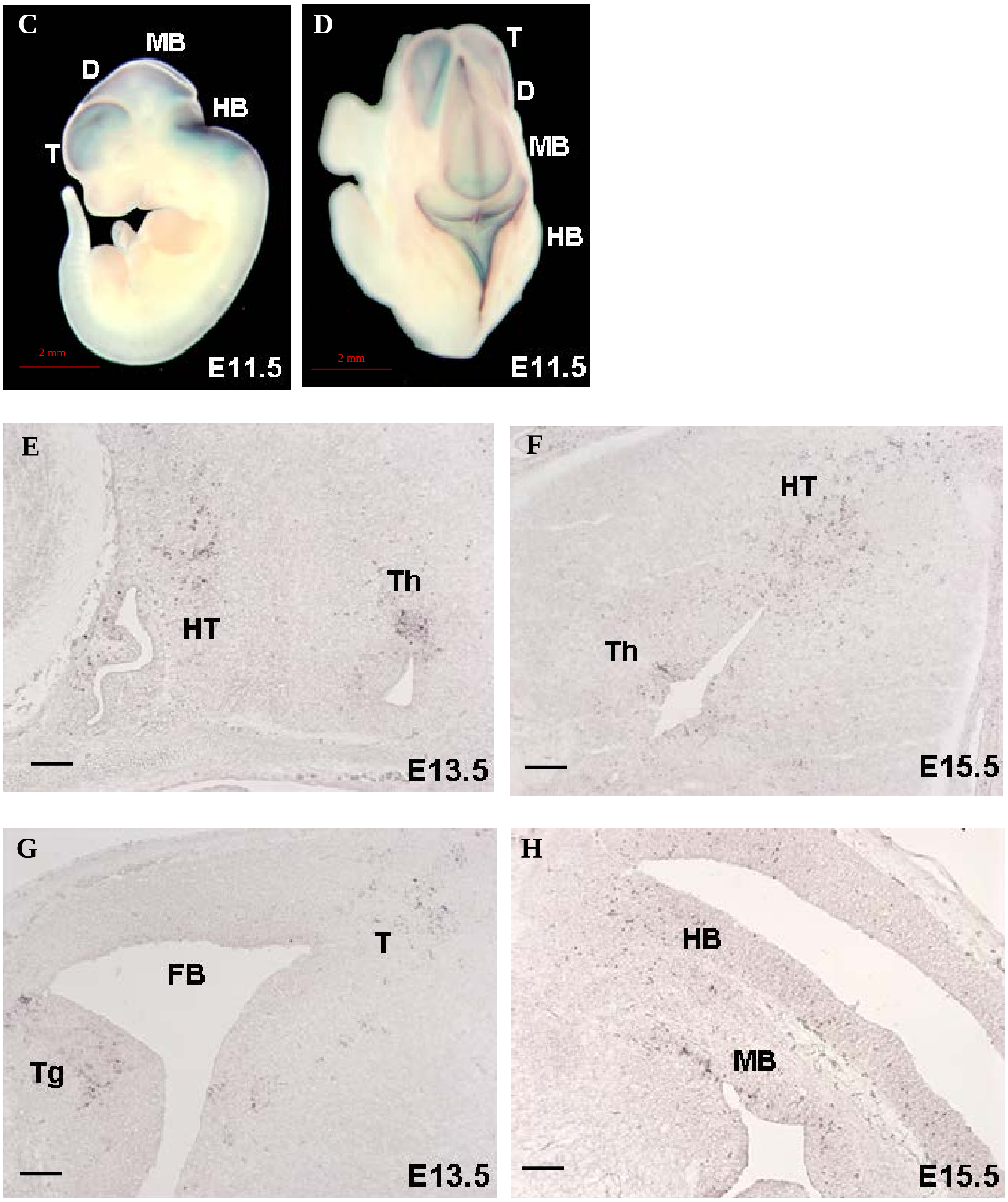


Figure 15: **A 400-bp basal promoter region of mouse *Foxb1* directs expression of *lacZ* mRNA to the diencephalon, midbrain and hindbrain during embryonic development**. (A-D) Whole mount in situ hybridization of transgenic embryos show that the 400-bp promoter region directs expression of reporter gene in the diencephalon of the forebrain, midbrain and hindbrain at E10.5 (A) and at E11.5 (C), with ectopic expression detected in the telencephalon. No expression was directed to the neural tube at these stages (B, D). (E-J) *In situ* hybridization of sagittal (e13.5) and coronal (e15.5) sections of the head of transgenic embryos at E13.5 (E, G) and E15.5 (F, H) show that the 400-bp

promoter region directs *lacZ* expression to the hypothalamus and thalamus of the diencephalon (E, F); and more weakly in the midbrain tegmentum and hindbrain (H). Ectopic expression was also detected in the telencephalon of the forebrain (G). D: diencephalon; FB: forebrain; HT: hypothalamus; MB: midbrain; T: telencephalon; Tg: tegmentum; Th: thalamus. Scale bar = 100µm unless otherwise indicated.

5.4.2 Expression patterns directed by CNEs 1, 2, 4 and 5

CNE1 consisted of 2 conserved fragments spanning a total of 455 bp and was amplified as a 680 bp sequence and cloned upstream of the mouse *Foxb1* promoter (CNE3-P) construct. CNE2 was made up of one conserved 235 bp fragment and amplified as a 290 bp fragment and cloned upstream of the mouse *Foxb1* promoter. CNE4 comprised one conserved fragment of 246 bp of genomic sequence, and was cloned as a 460 bp fragment upstream of the promoter. CNE5 spanned 187 bp and was amplified as a 340 bp sequence cloned upstream of the promoter. Expression of these transgenes was then individually assayed across different time points using *lacZ* RNA probe in situ hybridization.

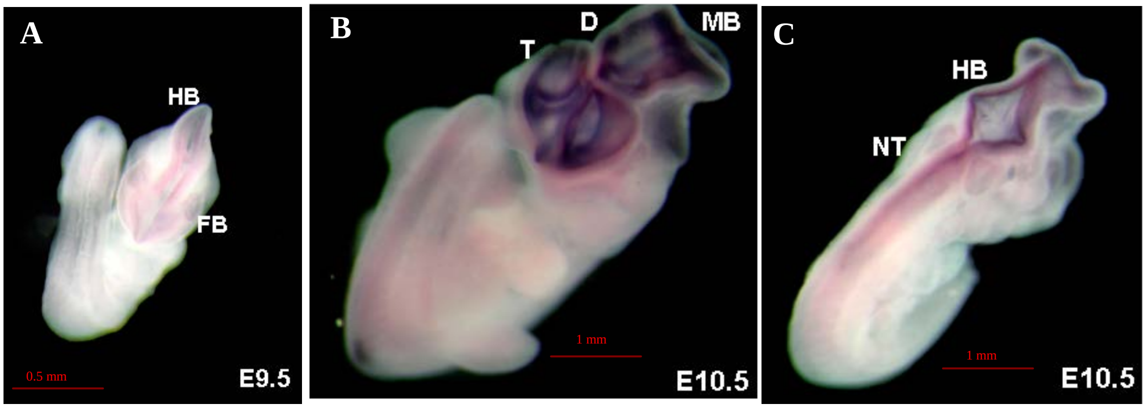
At E9.5 CNE1 directed *lacZ* expression to the presumptive diencephalon, the midbrain and hindbrain (Figure 16A), but not in the neural tube. Expression at this stage was weaker compared to the endogenous gene expression (compare with Figure 14A). There was weak ectopic expression in the presumptive telencephalon region of the forebrain (Figure 16A). At E10.5, *lacZ* mRNA expression had intensified in the same domains (Figure 16B), and had started expressing in the neural tube (Figure 16C). Expression remained in these domains of expression at E11.5 (Figures 16D and 16E). Therefore the effect of CNE1 on the promoter during early embryonic development was to direct early expression at E9.5 to the diencephalon, midbrain and hindbrain, as well as to act as an enhancer for neural tube expression from E10.5 onwards.

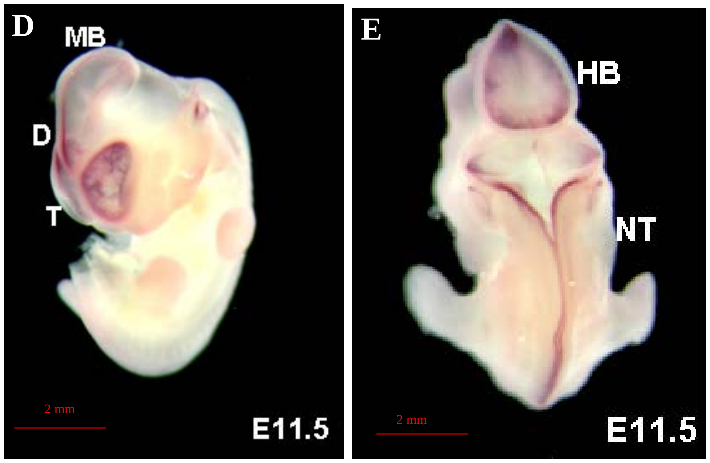
For CNE2, no expression was detected in transgenic embryos at E9.5. At E10.5, expression of the transgene was seen in the prospective diencephalon, the midbrain, hindbrain and neural tube (Figures 16F and 16G). There was ectopic expression in the prospective telencephalon that persisted till later developmental stages. This was similar to the basal promoter construct at the same stage (Figure 15A), except the additional neural tube expression and therefore CNE2 is likely to be a neural tube enhancer working in concert with the promoter at E10.5. However at E11.5, expression in the midbrain, hindbrain and neural tube were almost completely abolished (Figure 16H) while that in the diencephalon remained (Figure 16I). These results showed that CNE2 acted as an enhancer of the neural tube only at E10.5 and was a silencer of expression in the midbrain and hindbrain at E11.5.

CNE4 directed transgene expression as early as E9.5 to the presumptive diencephalon, midbrain and hindbrain with ectopic expression observed in the presumptive telencephalon (Figure 16J). However at E10.5 expression was detected only in the diencephalon and ectopically in the telencephalon (Figure 16K). It therefore acted as an early enhancer for *Foxb1* expression at E9.5, as well as a midbrain and hindbrain silencer at E10.5 (Figure 16L). At E11.5, expression was detected in the diencephalon and restored in the midbrain and hindbrain. In addition, ectopic expression was observed in the telencephalon (Figure 16M). There was no neural tube expression (Figure 16N), which meant the expression at E11.5 completely overlapped with that of the basal promoter and CNE3 had no observable effect on the promoter at this particular stage.

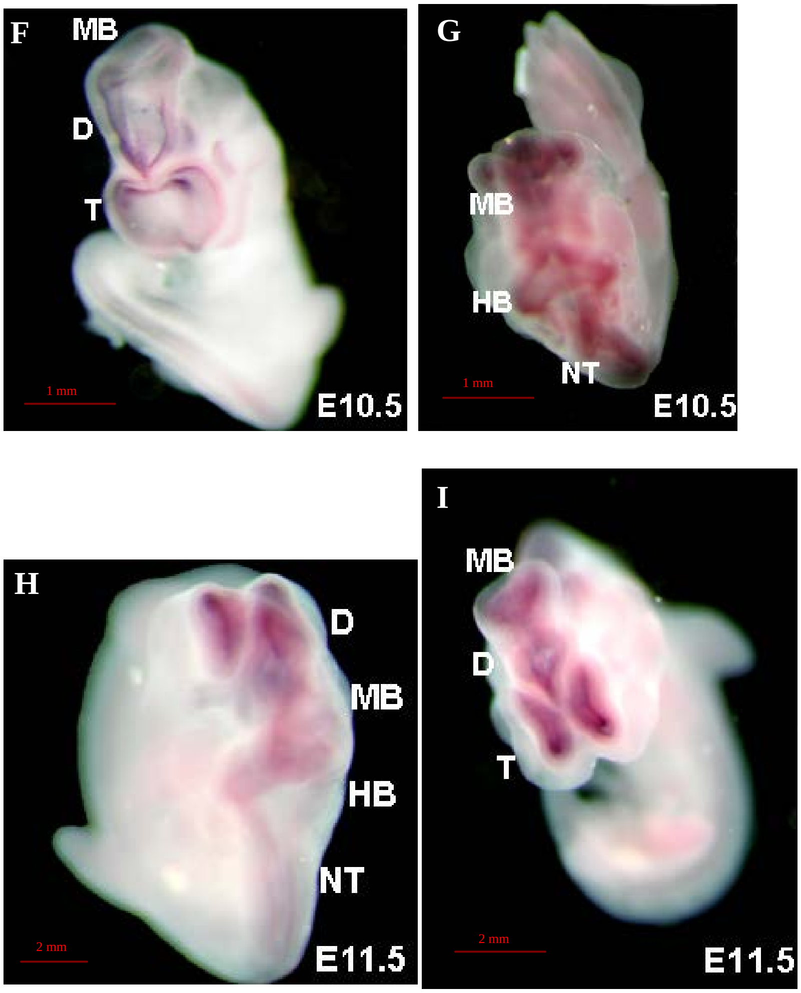
CNE5 directed no visible expression in E9.5 transgenic embryos but at E10.5 expression was detected in the presumptive diencephalon, the midbrain, hindbrain and neural tube (Figures 16O and 16P), with ectopic expression in the presumptive telencephalon. At E11.5, expression was abolished in the neural tube but remained in the other domains (Figure 16Q), indicating CNE5 acted solely as a neural tube enhancer at E10.5 in its interaction with the promoter. At later stages of E13.5 and E15.5, all four constructs (CNE1, CNE2, CNE4 and CNE5) directed expression to the hypothalamus and thalamus; midbrain tegmentum; and hindbrain; as well as ectopically in the telencephalon (data not shown). This overlapped completely with the basal promoter expression, and showed that CNE1, CNE2, CNE4 and CNE5 had no observable phenotypic effect on the promoter during these late developmental stages.

CNE1

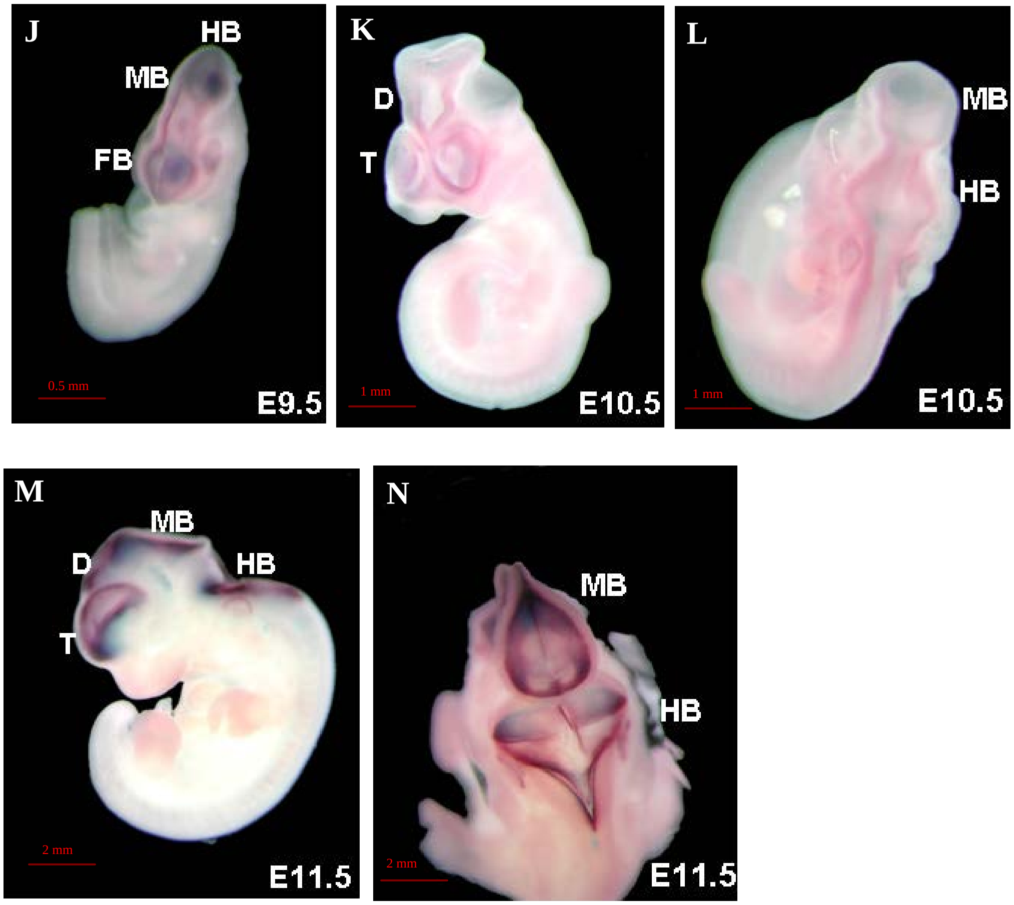



D
MB
D
T
2 mm
E11.5
E
HB
NT
2 mm
E11.5

CNE2


F
MB
D
T
1 mm
E10.5
G
MB
HB
NT
1 mm
E10.5
H
D
MB
HB
NT
2 mm
E11.5
I
MB
D
T
2 mm
E11.5

CNE4


J
HB
MB
FB
0.5 mm
E9.5
K
D
T
1 mm
E10.5
L
MB
HB
1 mm
E10.5
M
MB
D
HB
T
2 mm
E11.5
N
MB
HB
2 mm
E11.5

CNE5

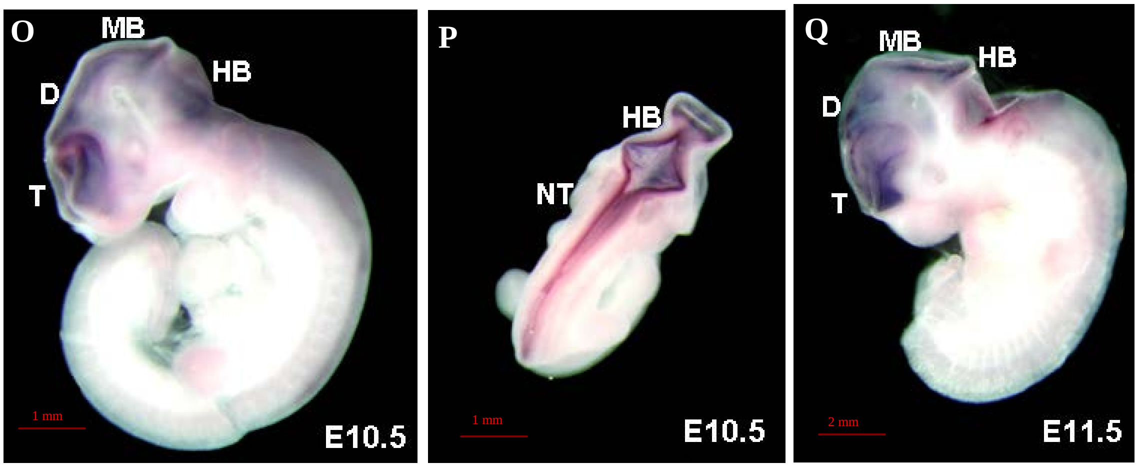

O
MB
HB
D
T
1 mm
E10.5
P
HB
NT
1 mm
E10.5
Q
MB
HB
D
T
2 mm
E11.5

Figure 16: **Whole mount in situ hybridization showing expression patterns directed by m*Foxb1* CNE1, CNE2, CNE4 and CNE5.** (A-E) CNE1 and m*Foxb1* basal promoter drove *lacZ* mRNA expression to the diencephalon, midbrain and hindbrain at E9.5 (A), E10.5 (B) and E11.5 (D), and activated neural tube expression at E10.5 (C) and E11.5 (E). (F-I) CNE2 and basal promoter directed expression to the diencephalon, midbrain, hindbrain (F) as well as in the neural tube (G) at E10.5, but at E11.5 it diminished expression in the midbrain and hindbrain (H) while maintaining expression in the diencephalon (I). CNE4 with the basal promoter gave rise to early expression in the presumptive diencephalon, midbrain and hindbrain at E9.5 (J), silenced expression partially in the midbrain and hindbrain at E10.5 (K, L), and restored expression in the same domains again at E11.5 (M, N). (O-Q) CNE5 with the promoter directed expression to the diencephalon, midbrain and hindbrain (O), as well as to the neural tube at E10.5 (P). Neural tube expression was abolished while the other expression domains remained at E11.5 (Q). Ectopic expression was observed in the presumptive telencephalon for all of the constructs. D: diencephalon; FB: forebrain; HB: hindbrain; MB: midbrain; T: telencephalon.

5.4.3 Summary of the regulatory potential of mouse *Foxb1* CNEs

The CNEs I analyzed from the mouse *Foxb1* locus all functioned as tissue-specific enhancers and / or silencers in transgenic mouse embryos. CNE3-P and its flanking sequences that represent the basal promoter tested directs expression to almost all domains of the endogenous mouse *Foxb1* expression in both the early and late stages of embryo development, while the other CNEs work in concert with the basal promoter to silence or enhance expression in certain domains at particular stages of development. Table 6 provides a summary of the effect of each CNE on the expression pattern driven by the mouse *Foxb1* promoter and clearly shows their overlapping roles in modulating the basal promoter.

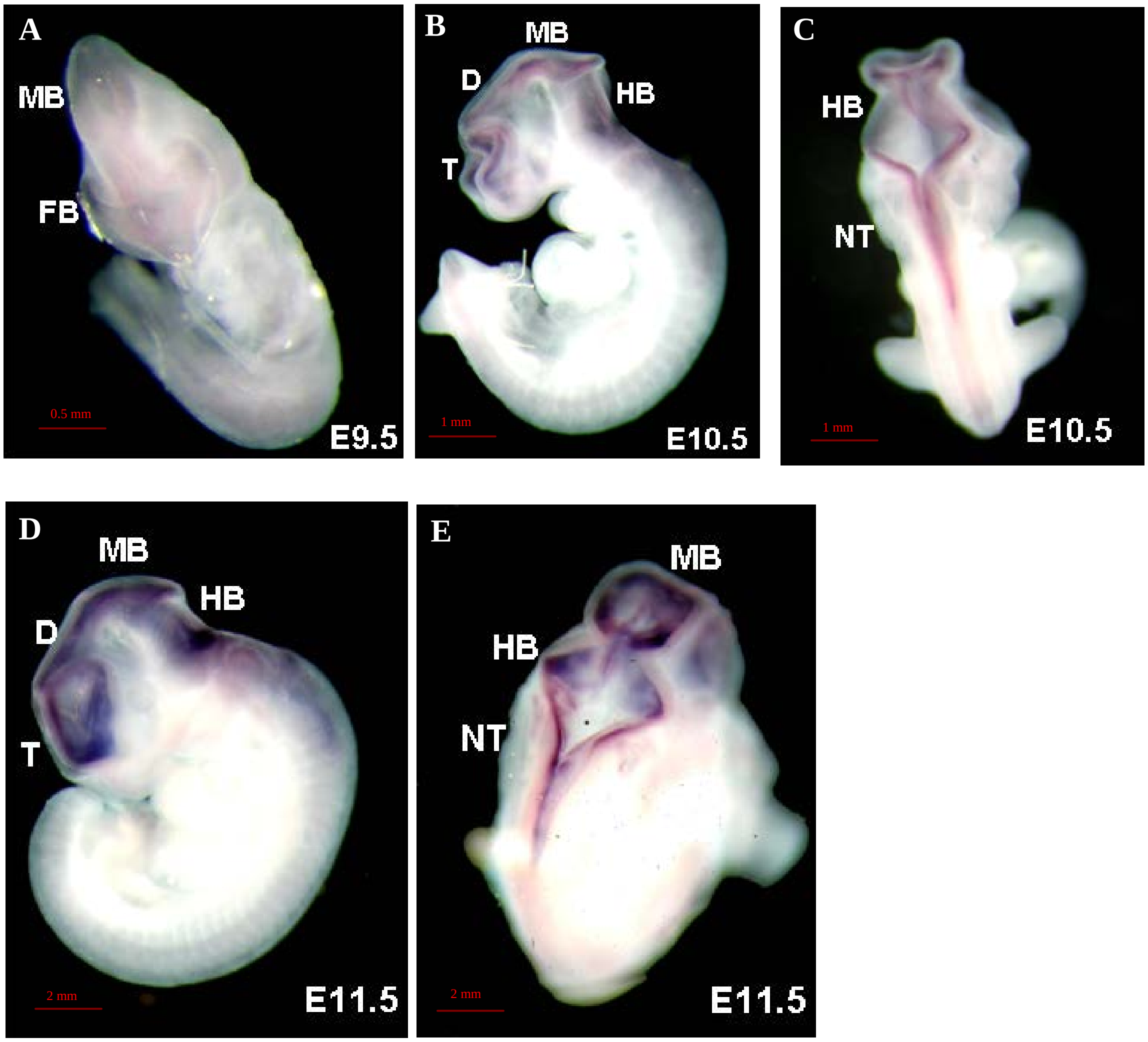
CNE	D	MB & HB	Neural tube	Others	Ectopic
3-P	Enhancer from E10.5	Enhancer from E10.5	-	Enhancer of HT and Th in differentiated D	T from E10.5
3-P +1	Enhancer at E9.5	Enhancer at E9.5	Enhancer at e10.5-E11.5		T from E9.5
3-P +2		Silencer at E11.5	Enhancer at E10.5		
3-P +4	Enhancer at E9.5	Enhancer at E9.5; Silencer at E10.5			T from E9.5
3-P +5			Enhancer at E10.5		

Table 6: **Enhancer function of mouse *Foxb1* CNEs across different developmental stages and in different tissues.** CNE3 is part of the basal promoter tested. Other CNEs were cloned upstream of the basal promoter (CNE3-P) and expression pattern was assayed using *in situ* hybridization that detected *lacZ* mRNA and compared with that driven by the basal promoter alone. The differences in the expression pattern conferred by each CNE are then tabulated. D: diencephalon; HB: hindbrain; HT: hypothalamus; T: telencephalon; Th: thalamus.

5.4.4 Conservation of regulation of *Foxb1* between fugu and mouse

A prediction from Table 6 would be if the basal promoter (CNE3-P) was working in concert with CNEs1, 2, 4 and 5 at the same time, *Foxb1* expression in the diencephalon, midbrain and hindbrain would be detected by E9.5 (activated by CNEs1 and 4), and expression in the midbrain and hindbrain modulated by CNEs 2 and 4 at E10.5-E11.5 so that the primary expression domain would be the diencephalon that would differentiate to form the hypothalamus and thalamus where Foxb1 would be primarily expressed in the later stages of embryonic development (directed by basal promoter). In addition, neural tube expression would be detected from E10.5 (activated by CNEs1, 2 and 5), while ectopic expression in the telencephalon would be observed from E9.5. I decided to validate the combined regulation of *Foxb1* by the 5 CNEs including the basal promoter by using the orthologous fugu DNA sequences to determine if this regulation was conserved between mouse and fish in spite of the slight differences in their sequences (the identity between mouse and fugu CNEs is 68% to 82%). Since the intergenic regions in fugu were much shorter than in mice, I could amplify CNEs 1 and 2 as a single 1.2 kb fragment, and CNEs 4 and 5 as a single 0.9 kb fragment. These fragments were cloned upstream of the fugu basal promoter (CNE3-P) linked to the β-galactosidase reporter such that the construct contained CNEs 4+ CNE5 +CNE1 +CNE2 upstream of the fugu basal promoter. This construct was then tested in transgenic mice and expression was assayed at various stages of development using in situ hybridization of *lacZ* riboprobe as previously described.

Expression was detected from E9.5 in the prospective diencephalon, midbrain and hindbrain, with ectopic expression observed in the prospective telencephalon (Figure 17A). At E10.5, expression intensified in the above-mentioned domains (Figure 17B), and was first detected in the neural tube (Figure 17C). At E11.5, expression remained in the diencephalon, midbrain, hindbrain and neural tube, as well as ectopically in the telencephalon (Figures 17D and 17E). At later stages of development (E13.5-E15.5), transgene expression was observed primarily in the hypothalamus and thalamus (Figure 17F), as well as in the midbrain and hindbrain (Figure 17G). Ectopic expression was also observed in the striatum of the telencephalon during these stages. Therefore this fugu construct that combined all 5 CNEs has reproduced almost all of the endogenous gene expression in both early and late stages of development. The exceptions were the lack of neural tube expression and the lack of inhibition of ectopic expression in the telencephalon both at stage E9.5. These results show that the mouse transcriptional machinery was able to interact with the fugu basal promoter and enhancers and direct tissue-specific expression similar to the expression pattern of the endogenous mouse elements, despite the differences in the sequences (68% to 82%) between the fugu and mouse elements.



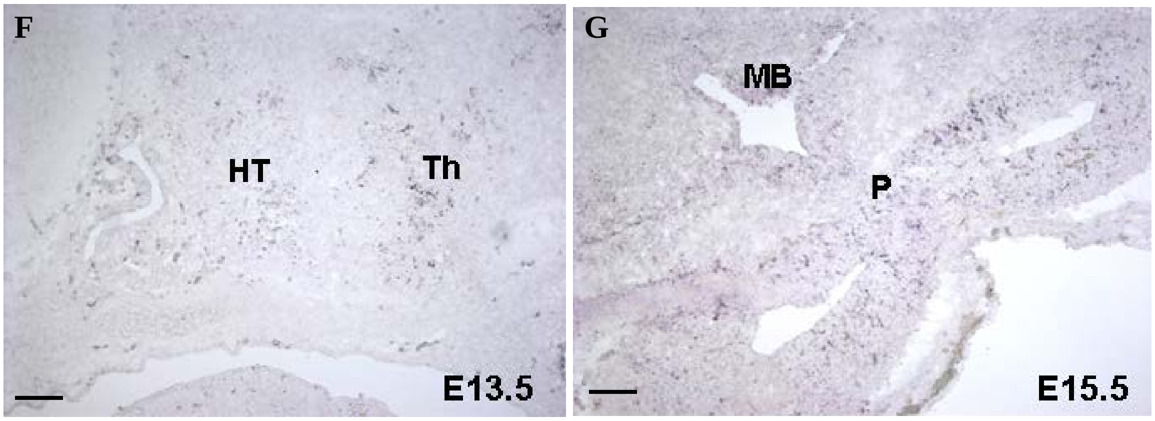


Figure 17: **A fugu construct containing CNEs 1, 2, 4 and 5 upstream of the basal promoter containing CNE3 reproduces mouse endogenous *Foxb1* expression in the diencephalon, midbrain and hindbrain**. (A-E) Whole mount *in situ* hybridization of transgenic embryos show that the fugu construct directed expression of reporter gene in the diencephalon of the forebrain, midbrain and hindbrain from E9.5 (A). Ectopic expression was also detected in the telencephalon (A). No expression was directed to the neural tube at this stage. Expression intensified at E10.5 (B) and in addition was detected in the neural tube (C). At E11.5 expression remained in the above mentioned domains (D, E). (F-G) *In situ* hybridization of sagittal (e13.5) and coronal (e15.5) sections of the head of transgenic embryos at E13.5 (F) and E15.5 (G) showed that the fugu construct directed *lacZ* expression to the hypothalamus and thalamus of the diencephalon (F); and more weakly in the midbrain and hindbrain (G). Ectopic expression was also detected at these stages in the telencephalon. D: diencephalon; FB: forebrain; HT: hypothalamus; MB: midbrain; T: telencephalon; Th: thalamus. Scale bar = 100 µm unless otherwise indicated.

5.5 Discussion

Foxb1 is an important regulator of the organization of the diencephalon during vertebrate forebrain development (Alvarez-Bolado et al., 2000). Defining the precise gene regulatory network that controls its spatiotemporal expression would help in elucidating the mechanisms by which it regulates forebrain development. This study has allowed a better understanding of the transcriptional mechanisms responsible for regulating *Foxb1* expression in a tissue and developmental-stage specific manner. Firstly, CNE3 together with the flanking sequences representing the basal promoter was able to specify the prospective diencephalon, midbrain and hindbrain expression from E10.5, maintain it at E11.5, and by later stages (E13.5-E15.5) direct expression to the differentiated diencephalon (hypothalamus and thalamus), midbrain tegmentum and hindbrain. This implies that the basal promoter can direct expression to the majority of the endogenous mouse *Foxb1* expression domains. However, it is not clear if this expression pattern is due to the CNE3 sequence or the sequences flanking it in the basal promoter construct tested or a combination of both, although the high level of conservation of CNE3 suggests that it is most likely playing a role in the expression patterns observed with this construct. In the early developmental stages of E10.5-E11.5, the basal promoter alone was sufficient to reproduce endogenous gene expression levels in the diencephalon, midbrain and hindbrain; but it did not specify neural tube expression that was part of the early expressing *Foxb1* domains in mouse. In addition, it did not down-regulate expression in the telencephalon at E10.5-E11.5, which was consistent with what I observed for the endogenous gene expression pattern at these stages of development. In the later stages of development, the promoter was sufficient to direct expression to the hypothalamus and thalamus, as well as to the differentiated midbrain and hindbrain, reproducing completely the endogenous gene expression level. However there was ectopic expression in the striatum of the telencephalon, which was not down-regulated at these later stages. Therefore a silencer located in this locus is likely to be involved in suppressing the expression of the basal promoter in the telencephalon to ensure the correct expression of *Foxb1* during embryonic development.

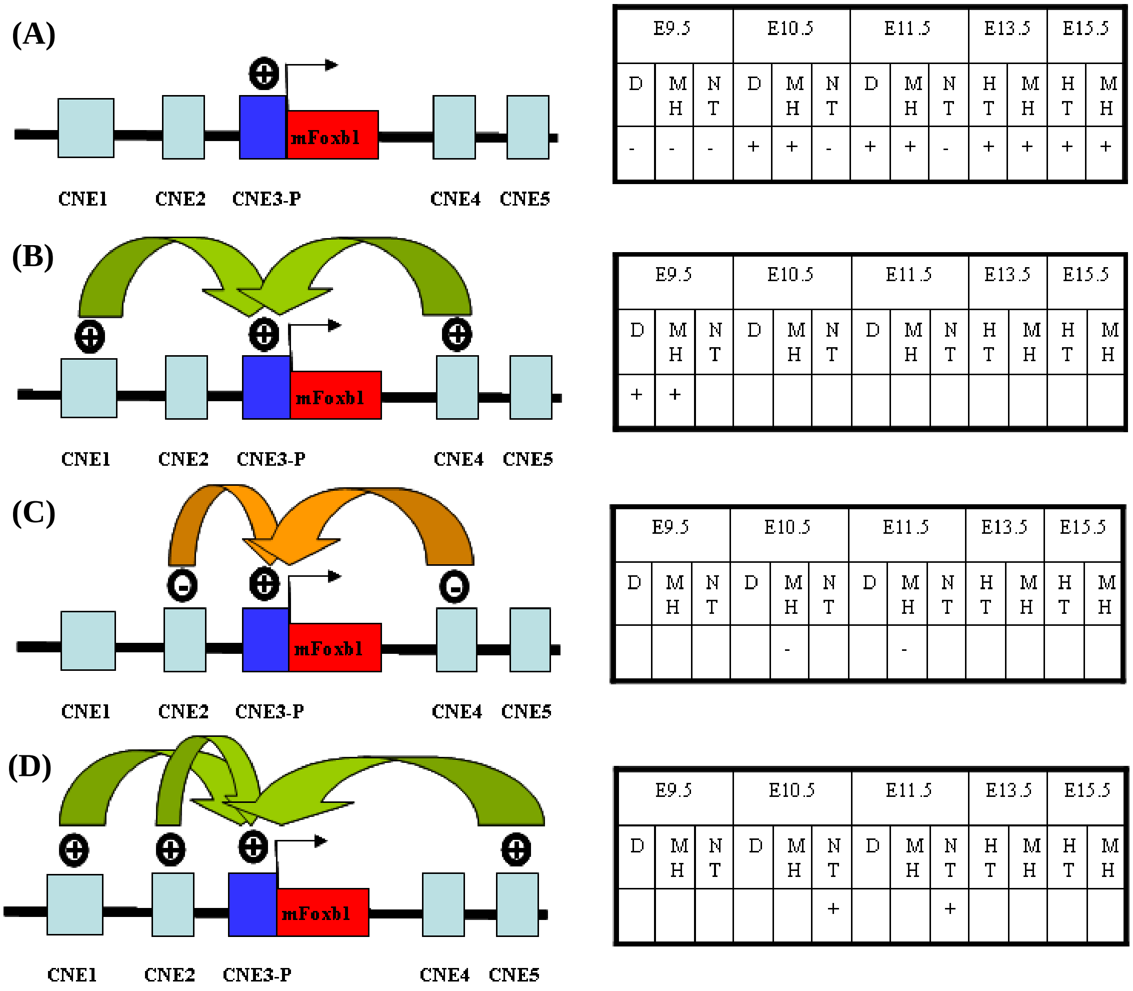
Secondly, the four CNEs in the *Foxb1* locus acted as enhancers and/or silencers with overlapping functions. The functions of individual enhancers that were ‘added on’ to the basal promoter construct can be deduced by a comparison of expression patterns resulting from the effect of individual CNEs on the basal promoter, with that observed due solely to the action of the basal promoter. I found two early enhancers (CNEs 1 and 4) that directed expression to the presumptive diencephalon, midbrain and hindbrain at E9.5, reproducing endogenous gene expression at that stage. Again there was ectopic expression in the presumptive telencephalon at this stage that was not present in the endogenous gene expression pattern. In the genomic context of the *Foxb1* locus, these early enhancers would have had to be modulated or silenced by other *cis*- and/or *trans*-acting agents, so as not to give rise to ‘leaky’ expression in the telencephalon. I found such silencers or modulators for midbrain and hindbrain expression in CNEs 2 and 4 at E11.5 and E10.5 respectively. It is likely they are required to inhibit midbrain and hindbrain expression incompletely so as to establish a gradient of expression from the rostral end to the caudal end in the midbrain and hindbrain observed during these stages that is crucial for midbrain development (Wehr et al., 1997). In addition, there were three enhancers that gave rise to neural tube expression and this also complemented the action of the promoter to specify the endogenous *Foxb1* expression domains at the appropriate stages. CNE1 specified neural tube expression from E10.5 to E11.5 (CNE1) while CNE2 and CNE5 were activated only at E10.5. Neural tube expression of *Foxb1* is observed in early to mid gestation (E9.5-E11.5) before expression becomes restricted to the hypothalamus, midbrain and hindbrain (Labosky et al., 1997; Wehr et al., 1997). Therefore the neural tube enhancers are only required for a short time in development, and this could explain their functional segregation from the promoter which specifies more persistent expression that will last through embryonic development till after birth.

The validation of a transgene containing the fugu orthologs of CNEs 1 to 5 linked to the fugu promoter reproduced almost all the effects of individual mouse CNEs in concert with the mouse basal promoter in specifying most domains of embryonic mouse *Foxb1* expression. However the silencing effects of mouse CNEs 2 and 4 in modulating expression in the midbrain and hindbrain could not be observed. This could be due to the presence of additional enhancers present in the fugu CNEs that also directed midbrain and hindbrain expression; or the presence of silencer blockers that neutralized the silencing effects of CNEs2 and 4 at E11.5 and E10.5 respectively. Alternatively, the function of these silencers may not be conserved in the fugu CNEs. This conservation of function of fugu and mouse CNEs despite the slight differences in their sequences indicated that the core sequence that is totally conserved in the two species are important for the expression of the CNEs.

The *in vivo* validation of CNEs proximal to the *Foxb1* locus reveals some clues on the regulatory logic of *Foxb1*. Firstly, the basal promoter containing CNE3 specifies the primary domains of *Foxb1* both in the early and late stages of development (Figure 18A). This includes the diencephalon that later develops to form the hypothalamus and thalamus; the midbrain and the hindbrain. Secondly, the CNEs act to modulate the basal promoter. For example they can act as temporal enhancers to the basal promoter (CNEs 1 and 4, Figure 18B); or they can act as modulators of the primary expression domains

(CNEs 2 and 4, Figure 18C); or they can act as enhancers of secondary expression domains like the neural tube (CNEs 1, 2 and 5, Figure 18D); or they can act as silencers of ectopic expression by the basal promoter like in the telencephalon. For the latter, none of the CNEs I tested had this function. Such a silencer would likely to be present in the more distant CNEs since the presence of this silencer is needed to ensure correct expression of *Foxb1*. Thirdly, the conservation of the regulatory organization and information of *Foxb1* in mouse and fugu through the validation of a handful of CNEs indicate this regulatory logic is likely to be conserved in all vertebrates and is crucial for the proper expression of *Foxb1* in embryonic development.

This study has focused primarily on the developing embryonic brain as the site of *Foxb1* expression studies. It must be pointed out that *Foxb1* also expresses in the mammary gland epithelium from embryonic to adult stages. Expression starts off at about E12.5 days and is restricted to the epithelial cells of the embryonic gland, before expanding after birth to the epithelial cells in the nipple anlage and those lining the ducts of the mature mammary gland (Kloetzli et al., 2001). As such, expression during late embryonic development in the embryonic gland was not readily observable and time constraints did not permit me to analyze the stronger mammary gland expression in the adult stage. Expression of *Foxb1* in the mammillary bodies in the hypothalamus during embryonic development and in the mammary gland epithelium during adult development is likely to be crucial for the regulation of the milk ejection reflex and the ability to lactate in mammals. However fishes do not have mammary glands, so it is doubtful if regulation of *Foxb1* in mammary gland is conserved in fish. Nevertheless, the use of fugu to identify *cis*-regulatory elements of *Foxb1* expression in the developing mouse embryonic brain and the ability of fugu CNEs to direct similar expression as the mouse CNEs to the different brain domains has shown that regulation is conserved at the level of hypothalamus, thalamus, midbrain and hindbrain expression, and the organization and development of the forebrain is likely to be conserved in mouse and fish.



E9.5			E10.5			E11.5			E13.5		E15.5	
D	MH	NT	D	MH	NT	D	MH	NT	HT	MH	HT	MH
-	-	-	+	+	-	+	+	-	+	+	+	+

E9.5			E10.5			E11.5			E13.5		E15.5	
D	MH	NT	D	MH	NT	D	MH	NT	HT	MH	HT	MH
+	+											

E9.5			E10.5			E11.5			E13.5		E15.5	
D	MH	NT	D	MH	NT	D	MH	NT	HT	MH	HT	MH
				-			-					

E9.5			E10.5			E11.5			E13.5		E15.5	
D	MH	NT	D	MH	NT	D	MH	NT	HT	MH	HT	MH
					+			+				

Figure 18: Summary of the regulatory code that controls the expression of *Foxb1* in mouse. (A) The basal promoter (CNE3-P) alone directs expression to the diencephalon, midbrain and hindbrain in the early stages of development from E10.5; as well as to the differentiated hypothalamus, thalamus, midbrain and hindbrain during the later stages of development. (B) *Foxb1* expression in the presumptive diencephalon, midbrain and

hindbrain at E9.5 is mediated by at least two enhancers (CNE1 and CNE4). (C) Expression of *Foxb1* in the midbrain and hindbrain during early development is modulated by at least two silencers (CNE2 and CNE4) that incompletely repress expression at E10.5-E11.5. (D) Neural tube expression of *Foxb1* is mediated by at least three enhancers (CNE1, CNE2 and CNE5) during early development at e10.5-E11.5. D: diencephalons; HT: hypothalamus and thalamus; MH: midbrain and hindbrain; NT: neural tube.

Chapter 6

Results

Regulation of Orexin

6.1 Introduction

In vertebrates, the hypothalamus plays a key role in the regulation of nutritional status and energy homeostasis through the coordination of many neurotransmitter systems. One such system is the recently discovered orexin-A and orexin-B (also known as hypocretin-1 and hypocretin-2) and their family of receptors. Orexin-A and -B are proteolytically derived from a single precursor protein encoded by the prepro-orexin gene (*ORX*). They were first discovered as ligands that bound to orphan G-protein coupled receptors in the rat brain, and their cDNAs were subsequently cloned (de Lecea et al., 1998; Sakurai et al., 1998). Since then, *ORX* gene has been cloned from several mammals including human (Sakurai et al., 1999), mouse (Chemelli et al., 1999), dog (Lin et al., 1999), pig (Dyer et al., 1999) and sheep (Archer et al., 2002). *ORX* mRNA and immunoreactive orexin-A are highly localized to distinct neurons in the lateral hypothalamic area (LHA) that has been regarded as the 'feeding center' (Nambu et al., 1999). Orexin neurons, however, innervate most regions of the central nervous system including various regions in the cerebral cortex, limbic system, and brain stem (Peyron et al., 1998). Besides mammals, *ORX* gene has also been cloned from *Xenopus laevis* (Shibahara et al., 1999), chicken (Ohkubo et al., 2002) and zebrafish (Kaslin et al., 2004). It has been found that the general organization of the *ORX* system of the brain, which includes a hypothalamic cell cluster and widespread fiber projections, seems to be conserved among vertebrates (Kaslin et al., 2004).

Orexins play a key role in regulating feeding behavior and states of sleep and wakefulness. Intracerebroventricular administration of orexins to rats led to a significant increase in food consumption (Sakurai et al., 1998). Furthermore, the expression levels of rat *ORX* are upregulated in response to fasting (Sakurai et al., 1998) and insulin-induced hypoglycemia (Moriguchi et al., 1999) indicating a role for orexins in regulating feeding behavior. Interestingly, the targeted disruption of *ORX* in mouse resulted in an autosomal recessive phenotype with characteristics similar to human narcolepsy (Chemelli et al., 1999). Subsequent studies in genetically narcoleptic dogs identified a mutation in the *ORX_2R* gene (Lin et al., 1999). Although no mutations were found either in the *ORX* or *ORX_2R* gene in human narcoleptics, a significant reduction in the number of *ORX* neurons and reduced *ORX* content in the cerebrospinal fluid was noted in such individuals (Thannickal et al., 2000). Conversely direct injection of *ORX* protein into the brain increases locomotor activity and decreases sleep for a few hours in mice (Mieda et al., 2004) and overexpression of *ORX* induces an insomnia-like phenotype in zebrafish (Prober et al., 2006). These findings show that orexin neuropeptide system plays a significant role in the regulation of sleep-wakefulness in mammals besides regulating the feeding behavior, and it has been proposed that the *ORX* system drives the aminergic and cholinergic system to control sleep and wakefulness states because of its widespread projections to the aminergic and cholinergic cell clusters (Kaslin et al., 2004). In fact the extensive projections of orexinergic neurons in the entire central nervous system (Cutler et al., 1999; Nambu et al., 1999) suggest that *ORX* may be involved in many other physiological functions, including the control of neuroendocrine systems and the autonomic nervous system (Johren et al., 2001).

The highly specific expression of *ORX* in the ‘orexinergic’ neurons of the LHA indicates a tight mechanism of regulation in the brain. To date there have been two main studies carried out in mice and zebrafish that have helped to elucidate the regulatory mechanisms of *ORX*. A 3.2 kb 5’flanking region of human *ORX* has been shown to be sufficient to direct expression to orexinergic neurons in the LHA of transgenic mice (Moriguchi et al., 2002) while in another study, a 1 kb 5’ flanking region of zebrafish *ORX* has been shown to be sufficient for driving cell-specific expression in the LHA in transgenic zebrafish (Faraco et al., 2006). In my efforts to identify CNEs in genes that express in the brain, I did not identify any CNEs in the human, mouse and fugu *ORX* loci. Since the expression pattern is conserved in these vertebrates, I was interested to see how the regulation is conserved despite of absence of CNEs in this locus. To address this I expressed the fugu *ORX* gene in transgenic mice to determine if the regulatory mechanism is conserved between mammals and fish and then I made deletions in the promoter region to identify the regulatory elements common to mammals and fish.

6.2 Comparison of *ORX* loci in human, mouse and fugu

I annotated all the genes present on the fugu *ORX* scaffold #424 (131 kb) based on their homology to known sequences in the NCBI database (see Materials and Methods). The fugu scaffold contains six genes besides *ORX* gene. These six genes are: signal transducer and activator of transcription 5 gene (*STAT 5*); phospholipase C-ε2 gene (*PLC-E2*); potassium voltage-gated channel protein subfamily H4 gene (*KCNH4*); a member of RAS oncogene family (*RAB5C*); transcriptional adaptor for general control of amino acid synthesis gene (*GCN5L2*), and a hypothetical protein *LGP2* gene (Figure 19). The fugu

ORX gene is flanked by *PLC-E2* at 2.2 kb upstream (from the polyadenylation signal of *PLC-E2* to the transcription start site of *ORX*) and *KCNH4* at 1.38 kb downstream (from the polyadenylation signal of *ORX* to the first codon of *KCNH4*). Comparisons of the fugu *ORX* locus with the human and mouse *ORX* loci show that the order and orientation of six of the fugu genes are conserved in the human and mouse loci (Figure 19). The order of genes in the human and mouse loci are totally conserved. Interestingly, the human and mouse loci contain two *STAT5* (5A and 5B) genes as compared to the single *STAT5* gene present at the 5' end of the fugu scaffold. The two *STAT5*s in the human and mouse loci are linked head to head suggesting that they arose through tandem duplication in the mammalian lineage. Alternatively, following the duplication of *STAT5* gene in a common ancestor of mammals and fishes, one copy may have been lost in the fugu lineage.

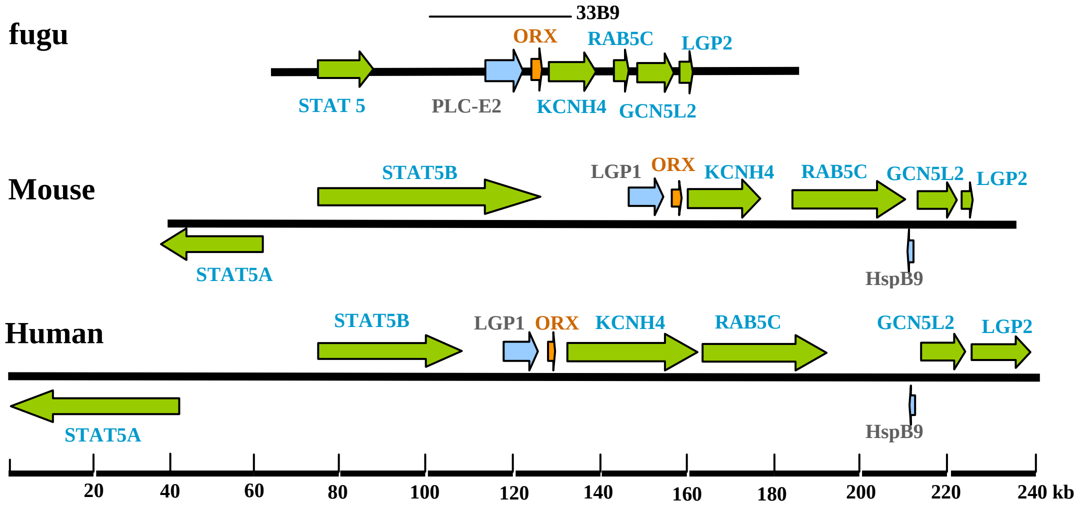


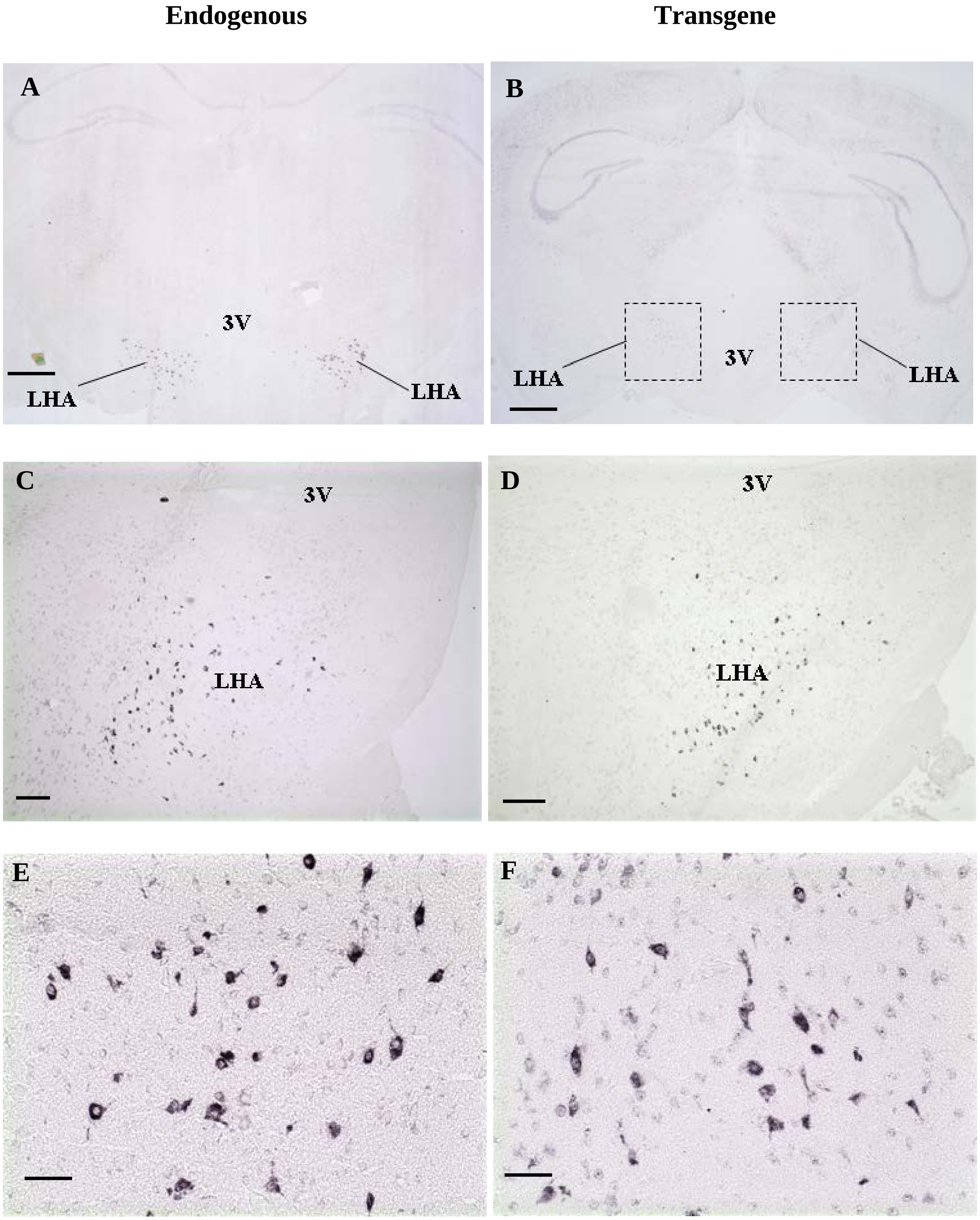
Figure 19**: *ORX* locus in fugu, mouse and human.** Arrows represent genes and indicate the direction of transcription. The gene order at the human and mouse loci were obtained from the UCSC Human Genome Browser (http://genome.ucsc.edu). *ORX* gene is indicated in orange while other syntenic genes are indicated in green. Fugu cosmid 33B9 which was used for filling gaps and to generate transgene constructs is indicated as a line above.

Comparison of non-coding sequences across evolutionarily distant vertebrates is a powerful strategy for identifying conserved *cis*-regulatory elements. Unfortunately aligning the noncoding regions of *ORX* in human, mouse and fugu using MLAGAN at a sensitive threshold of 60% identity and 50 bp window size did not identify any CNEs. In order to determine how in the absence of a conserved regulatory element the tissue-specific expression of *ORX* is achieved, I decided to find out firstly if the fugu *ORX* gene is regulated in the same way as the mouse gene, after which I could then localize the regulatory region in the fugu gene by progressive deletion of its locus.

6.3 Expression of fugu *ORX* in mouse

The fugu scaffold sequence surrounding the *ORX* gene contained several gaps. To fill the gaps and to make constructs for transgenic studies, I isolated fugu cosmids for this locus (see Materials and Methods). Cosmid 33B9 with an insert size of about 39 kb was selected as a representative clone for sequencing and gap filling. Annotation of the cosmid showed it contained the fugu *ORX* locus with its complete flanking sequences, together with the full coding region of *PLC-E2* upstream and part of the coding region of *KCNH4* downstream (Figure 19). Transgenic mice were then generated using the full cosmid sequence as described in Materials and Methods. Because the level of *ORX* expression gradually increases during postnatal development, analysis was performed using transgenic founder mice at 8-10 weeks of age (Moriguchi et al., 2002). The brains of the transgenic mice were removed and frozen, and coronal sections of the hypothalamus were taken. *In situ* hybridization was then carried out using an antisense RNA probe that binds specifically to a 320 bp fragment of the fugu *ORX* coding region.

The endogenous mouse *ORX* expression was detected by *in situ* hybridization using a RNA probe that binds to a 380 bp fragment of the mouse *ORX* coding region. This endogenous gene expression pattern was consistent with previous analyses (Moriguchi et al., 2002; Sakurai et al., 1999) in labeling specific neurons in the LHA (Figures 20A, C and E) and was carried out for the sake of comparison of the expression pattern of the mouse *ORX* gene with the expression pattern of the fugu *ORX* gene driven by its own regulatory region. The 43-kb cosmid (33B9) directed fugu *ORX* mRNA specifically in the mouse neurons in the LHA (Figures 20B, D and F). Indeed all 3 transgenic founder mice generated showed this LHA staining in specific neurons with no ectopic signal detected. To determine if the fugu cosmid directed *ORX* expression to the same LHA neurons as the mouse *ORX*, I did a double in situ hybridization in which mouse *ORX* neurons were stained light brown (Figure 20G) and fugu *ORX* neurons were stained red (Figure 20H). Both transcripts were found to be colocalized in the same neurons, as evident from the reddish brown staining of the neurons (Figure 20I). Thus, this experiment demonstrated that the fugu cosmid contained all the regulatory elements needed for directing fugu *ORX* to the mouse LHA neurons that express the endogenous mouse *ORX*, and that fugu *ORX* is therefore regulated the same way as mouse *ORX*. The question then was where the regulatory elements are located and whether their sequences in mouse and fugu are different.


Endogenous
Transgene
A
3V
LHA
LHA
B
LHA
3V
LHA
C
3V
LHA
D
3V
LHA
E
F

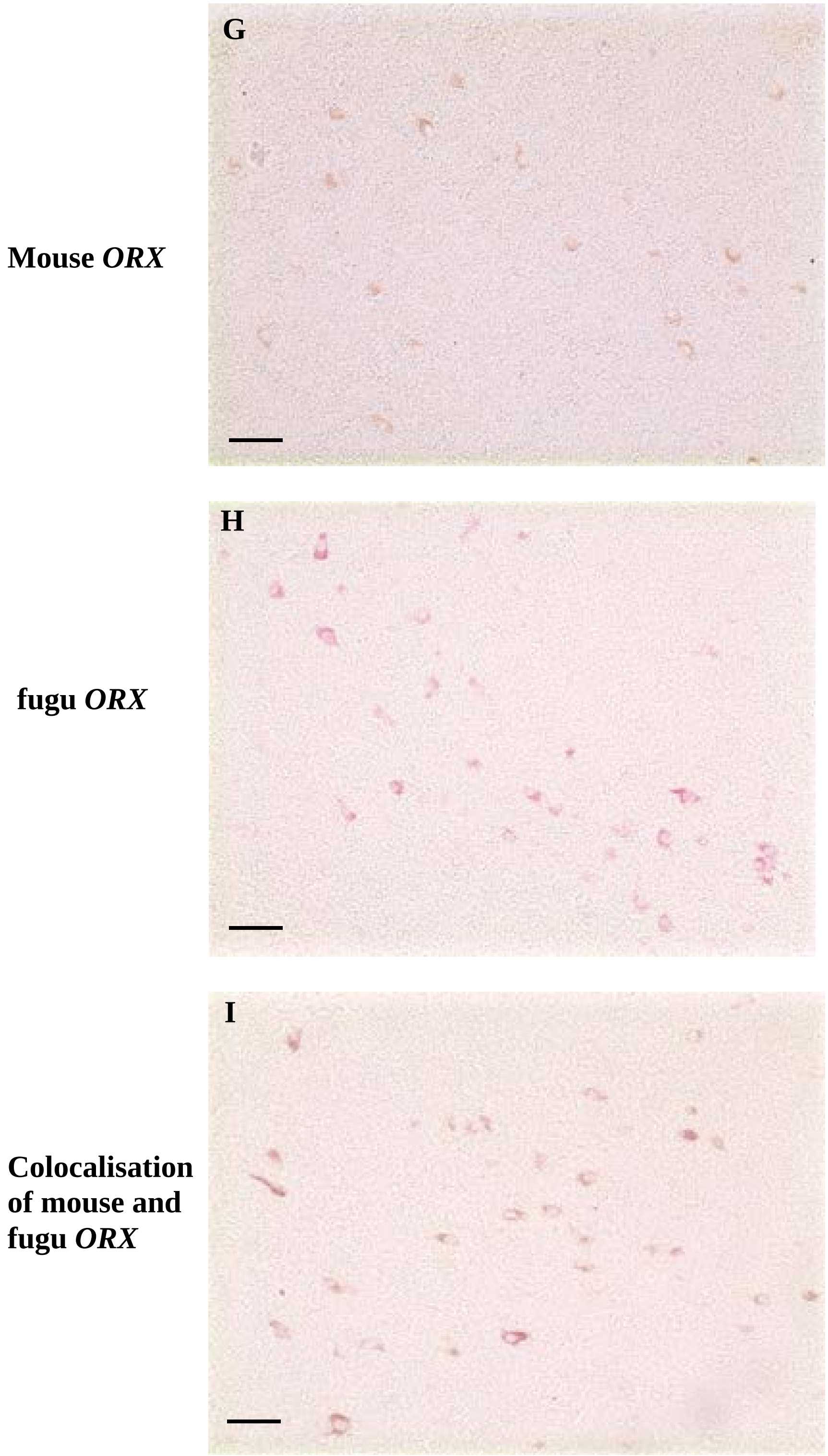


Figure 20: **Expression of fugu *ORX* gene in transgenic mice compared with the expression of the endogenous mouse *ORX* expression**. In situ hybridization of coronal sections of the hypothalamus of transgenic mice expressing a fugu *ORX* cosmid was carried out to detect mouse *ORX* expression (A, C, E, G), fugu *ORX* expression (B, D, F,

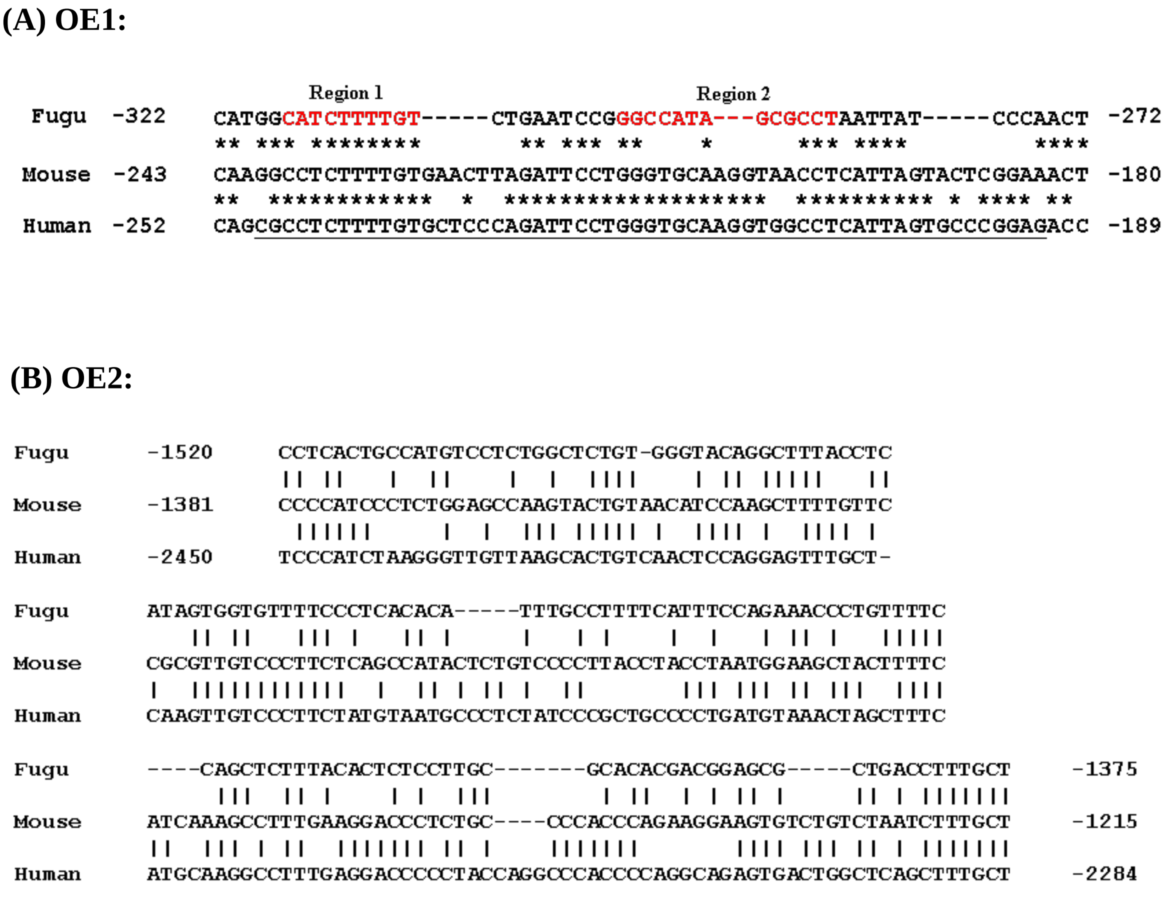
H) or the colocalization of both signals in the same LHA neurons (I). LHA: lateral hypothalamus; 3V: third ventricle. Scale bars: 200 µm in A and B; 50 µm in C and D; 10 µm in E-I.

6.4 Comparative analyses and validation of *ORX* regulatory elements common in human, mouse and fugu

A 3.2 kb 5'flanking region of human *ORX* has been shown to be sufficient to direct expression to orexinergic neurons in transgenic mice (Moriguchi et al., 2002). Further analysis of this promoter region identified two elements (OE1: 214 bp located 287 bp upstream and OE2: 217 bp located 2.5 kb upstream of the transcription initiation site of human *ORX*) that are conserved in humans and mouse, and essential for the expression in the LHA and for the repression in medial regions of the hypothalamus, with a core 57 bp being critical for the regulatory function of OE1 (Moriguchi et al., 2002). In addition, another study has shown that a 1 kb 5' flanking region of zebrafish *ORX* is sufficient for driving cell-specific expression in the LHA in transgenic zebrafish (Faraco et al., 2006). Further analysis of this promoter region and comparison of motifs with other teleost fishes including fugu identified a critical 250 bp element containing a core 13 bp essential for *ORX* expression (Faraco et al., 2006). Interestingly there was no overlap between both the 57 bp OE1 core and the 13 bp zebrafish core elements, and the human 3.2 kb *ORX* promoter region did not specify any reporter expression in transgenic zebrafish, indicating the regulation of *ORX* in mammals is likely to be different from that in fish (Faraco et al., 2006).

The fugu *ORX* 5’ flanking region spans about 2.2 kb. Alignment of this region with the 5’ flanking regions in human and mouse was carried out using the ClustalW algorithm and the alignment was inspected by eye to determine if there was orthologs for OE1 and OE2 elements in fugu. Comparison of fugu *ORX* 5’ flanking region with the 3.2 kb human and mouse *ORX* regulatory regions identified a 50 bp element in the fugu promoter that showed 44% similarity to the 57-bp core element in the mammalian OE1 (Figure 21A). The fugu element contains three deletions compared to the mammalian element. In the same way a 145 bp element was identified about 1 kb upstream of the fugu OE1 that showed 45% identity with mammalian OE2 element (Figure 21B). Although the sequence similarity suggests that the fugu elements are analogous to the mammalian OE1 and OE2 elements, the function of the fugu elements would need to be confirmed in transgenic experiments.

In the zebrafish study, Faraco et al. (2006) could identify conservation of regulatory motifs in the flanking regions of *ORX* from zebrafish, Tetraodon, fugu, medaka and stickleback in the 500 bp upstream of the TATA boxes, and the deletion of the region from -500 to -250 in the 1 kb zebrafish promoter construct resulted in a complete loss of reporter expression in transgenic zebrafish. In this crucial 250-bp region, there were 4 regions containing clusters of identical residues conserved between zebrafish and Tetraodon, and these were subjected to site-directed mutagenesis. Mutations in 3 of these regions (1, 3 and 4) reduced the efficiency of the zebrafish *ORX* promoter moderately, but complete deletion of region 2 (13 bp) totally abolished the activity of the 1 kb promoter construct (Faraco et al., 2006). Since this crucial 250 bp region would overlap with the location of fugu OE1, I checked manually to see if fugu OE1 would match any of the 4 conserved motifs listed in the zebrafish study (Faraco et al., 2006). Indeed there is a good match for region 1 in fugu OE1 in which 6 out of 10 residues are conserved between fugu and zebrafish; as well as for region 2 in fugu OE1 in which 6 out of 13 residues are conserved between fugu and zebrafish (residues in red, Figure 21A).



(A) OE1:

```
                     Region 1                      Region 2
Fugu   -322   CATGGCATCTTTTGT-----CTGAATCCGGGCCATA---GCGCCTAATTAT-----CCCAACT  -272
              ** *** ********       ** *** **    *      *** ****         ****
Mouse  -243   CAAGGCCTCTTTTGTGAACTTAGATTCCTGGGTGCAAGGTAACCTCATTAGTACTCGGAAACT  -180
              **  ************  *  *******************  ********** * **** **
Human  -252   CAGCGCCTCTTTTGTGCTCCCAGATTCCTGGGTGCAAGGTGGCCTCATTAGTGCCCGGAGACC  -189
```

(B) OE2:

```
Fugu      -1520     CCTCACTGCCATGTCCTCTGGCTCTGT-GGGTACAGGCTTTACCTC
                    || ||   |  ||    |  |  ||||    | || |||||   ||
Mouse     -1381     CCCCATCCCTCTGGAGCCAAGTACTGTAACATCCAAGCTTTTGTTC
                     ||||||     |  |  ||| ||||| |  |||| |  |||| |
Human     -2450     TCCCATCTAAGGGTTGTTAAGCACTGTCAACTCCAGGAGTTTGCT-

Fugu      ATAGTGGTGTTTTCCCTCACACA-----TTTGCCTTTTCATTTCCAGAAACCCTGTTTTC
             || ||   ||| |   || |     |   | |    |  |   | || |   |||||
Mouse     CGCGTTGTCCCTTCTCAGCCATACTCTGTCCCCTTACCTACCTAATGGAAGCTACTTTTC
          |  ||||||||||||  |  || | || |  ||       ||| ||| || |||  ||||
Human     CAAGTTGTCCCTTCTATGTAATGCCCTCTATCCCGCTGCCCCTGATGTAAACTAGCTTTC

Fugu      ----CAGCTCTTTACACTCTCCTTGC-------GCACACGACGGAGCG-----CTGACCTTTGCT     -1375
               |||  || |    | |  |||        | ||  | | || |     || | |||||||
Mouse     ATCAAAGCCTTTGAAGGACCCTCTGC----CCCACCCAGAAGGAAGTGTCTGTCTAATCTTTGCT     -1215
          ||  ||| | ||  ||||||| || |    |||||||       |||| ||| || | |||||||
Human     ATGCAAGGCCTTTGAGGACCCCCTACCAGGCCCACCCCAGGCAGAGTGACTGGCTCAGCTTTGCT     -2284
```

Figure 21: **Poorly conserved mouse and human regulatory elements in the fugu *ORX* locus**. Conserved bases of OE1 (A) and OE2 (B) are shown by asterisks and dashes. The 57-bp core element of OE1 in the human *ORX* promoter characterized by Moriguchi et al. (2002) is underlined. Two of the four conserved motifs investigated by Faraco et al. (2006) that could be identified in fugu OE1 are indicated in red. The numbers flanking the sequences are nucleotide positions in relation to transcription start site of *ORX*.

To identify the importance of conserved OE1 and OE2 elements in fugu, I generated a construct spanning the fugu *ORX* locus and containing the putative OE1 and OE2 elements (N-f*ORX*-K) by making deletions of cosmid 33B9 using the restriction enzymes *NheI* and *Kpn1*. This construct is 2.5 kb long and includes about 2 kb of the 5' flanking sequence containing OE1 and OE2, as well as the coding sequences of the gene, with the 3' flanking sequence removed (Figure 22). Three of seven N-f*ORX*-K transgenic founder mice showed fugu *ORX* mRNA labeling in specific neurons in the LHA, similar to the expression of the fugu cosmid but with lower intensity (Figure 23A). There was no ectopic expression observed. Double *in situ* hybridization indicated the endogenous mouse *ORX* mRNA stained as light brown neuronal cells (Figure 23B) and fugu *ORX* mRNA from N-f*ORX*-K stained as blue neuronal cells (Figure 23C) colocalized in the same LHA neurons (Figure 23D). To further delineate the contributions of fugu OE2 and OE1 elements to the expression in LHA neurons, I cloned fugu OE2 (220 bp) and the fugu *ORX* basal promoter (545 bp) containing OE1 upstream of a β-galactosidase reporter (construct fOE2OE1-lac; Figure 22) and generated transgenic mice carrying this construct. Remarkably double in situ hybridization using *lacZ*-specific RNA probe showed that the transgene expression (Figure 23E) is localized to the same LHA neurons as the mouse *ORX* (Figure 23F), indicating the crucial regulatory elements directing LHA-specific neuronal expression was likely to be contained in OE1 and OE2 in fugu. However there was some 'leaky' expression of the N-f*ORX*-K and fOE2OE1-lac transgenes since not all the neurons expressing the transgene in a blue signal overlapped with the brown endogenous signal (Figures 23D and F). To determine if fugu OE1 alone was sufficient to direct expression in the LHA, I generated a construct containing only

OE1 in the basal promoter cloned upstream of the β-galactosidase reporter (construct fOE1-lac; Figure 22). The minimal construct fOE1-lac containing OE1 within the fugu basal promoter failed to direct *lacZ* mRNA expression to neurons in the LHA of all five transgenic founder mice. This indicated that fugu *ORX* expression in the LHA require the cooperation of both OE1 and OE2 elements. The results of the characterization of the fugu *ORX* locus sequences in transgenic mice are summarized in Figure 22.

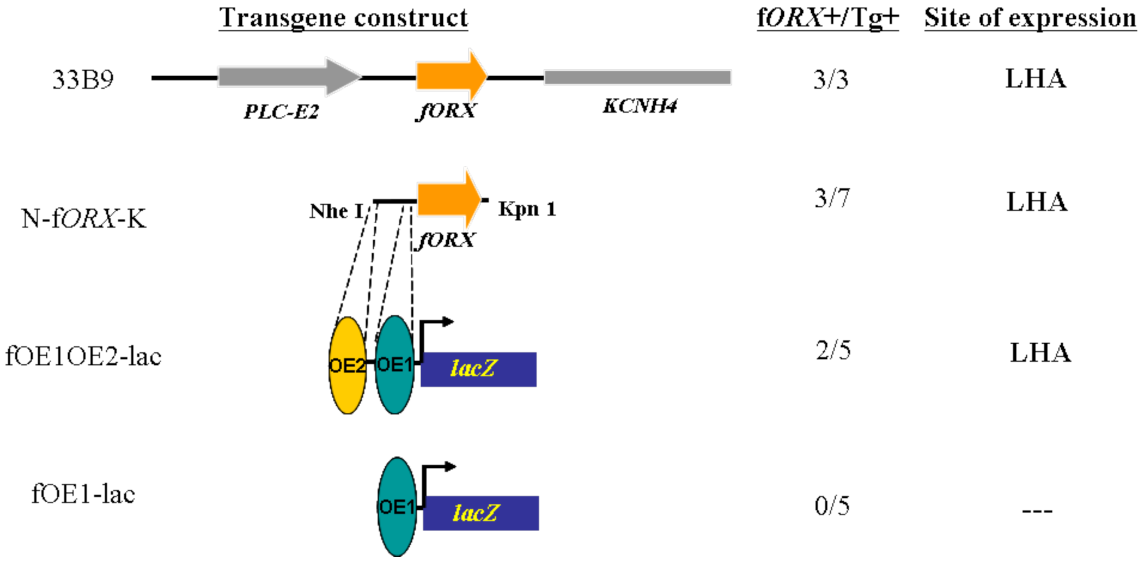


Figure 22: **Analysis of the regulatory region of fugu *ORX* in transgenic mice.** The four transgene constructs tested prepared using cosmid 33B9 are depicted on the left and the number of founder mice showing LHA expression (fORX+) compared to the total number of transgenic founder mice generated (Tg+) are shown on the right. LHA: Lateral hypothalamus.

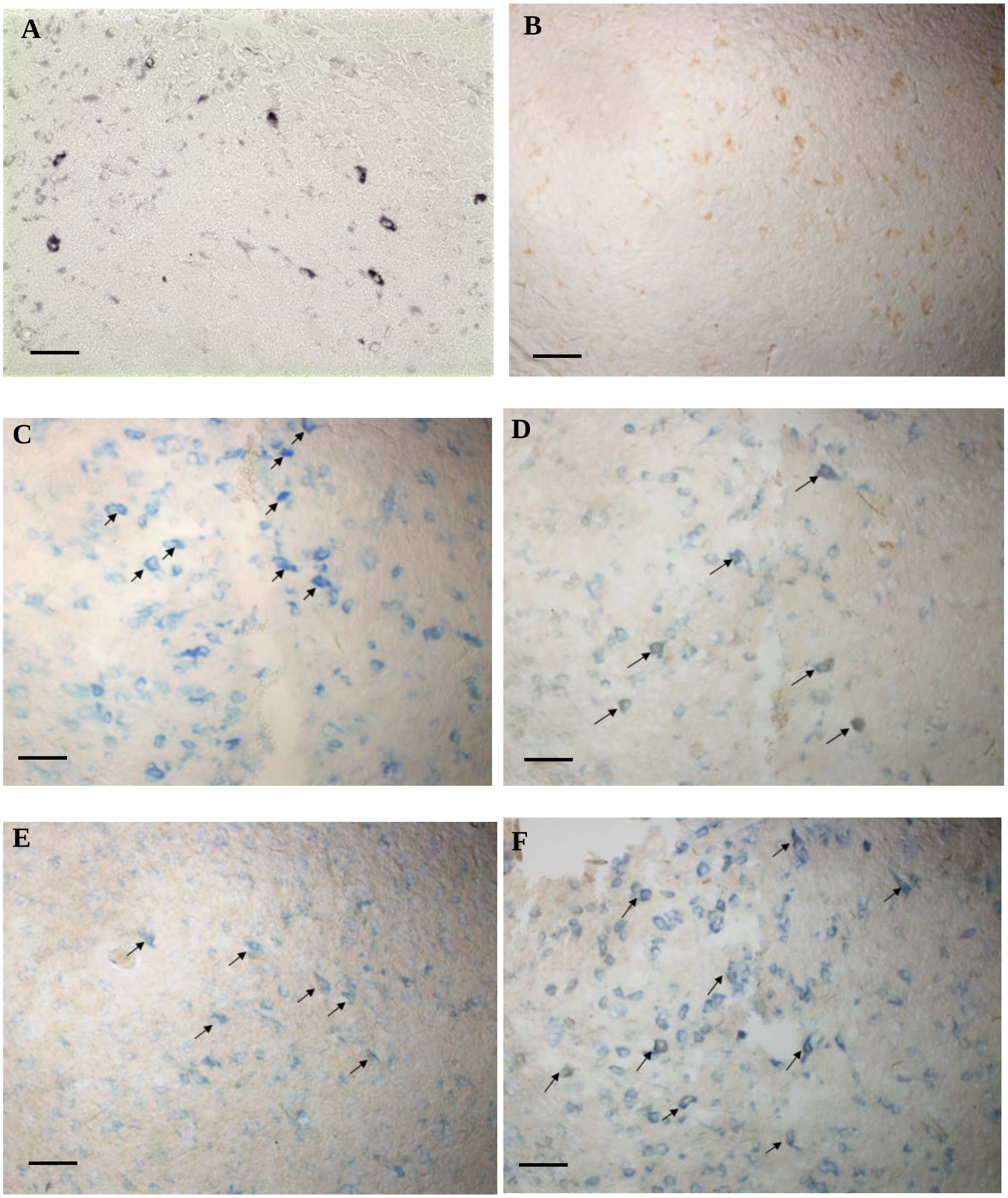


Figure 23: **Expression of fugu *ORX* in transgenic mice compared with the expression of the endogenous mouse *ORX*.** *In situ* hybridization of coronal sections of the hypothalamus of transgenic mice expressing N-f*ORX*-K (A-D) and fOE2OE1-lac (E-F) was carried out to detect fugu *ORX* expression (A, C, E), mouse *ORX* expression (B) and the colocalization of both signals in the same LHA neurons (D, F). Scale bar = 10 µm.

6.5 Discussion

Comparison of non-coding sequences across evolutionarily distant vertebrates is a powerful strategy for identifying conserved regulatory elements that are common to all vertebrates. The fugu is a model genome to characterize genes and gene regulatory regions because of its compact genome size and short intergenic regions that contain very little repetitive sequences (Aparicio et al., 2002). In this study, I analyzed the regulatory mechanisms governing fugu *ORX* gene expression using a transgenic mouse assay in order to shed more light on the regulation of *ORX* in vertebrates and what specifies its precise and exclusive expression in neurons of the hypothalamus. I analyzed 131 kb from the fugu *ORX* locus and showed that the synteny of genes in this locus is conserved in the human and mouse *ORX* loci. However alignment of the *ORX* loci in human, mouse and fugu did not turn up any CNEs in the regulatory regions. To determine if fugu *ORX* was regulated the same way as the mammalian *ORX*, I generated transgenic founder mice expressing a 43 kb fugu cosmid (33B9) containing the complete *ORX* locus with flanking sequences and genes. Remarkably, the fugu *ORX* transgene was expressed in specific neurons of the LHA in which the endogenous mouse *ORX* gene is expressed. This result shows that the mouse transcriptional machinery was able to interact with the enhancers and the basal promoter of the fugu ORX gene located in the fugu cosmid and direct tissue-specific expression to the same neurons as the endogenous mouse gene, despite the apparent lack of sequence conservation in the regulatory regions of fugu and mouse *ORX*.

Previously work done to analyze the 3.2 kb 5’ regulatory region of human *ORX* by comparison with the mouse locus uncovered two patches of conserved sequences designated OE1 and OE2 elements. OE1 was shown to be required to activate *ORX* expression in the LHA and repress it in the medial regions of the hypothalamus in tight cooperation with OE2 to regulate *ORX* expression specifically in the LHA (Moriguchi et al., 2002). Further characterization of OE1 in transgenic mice showed that a 57-bp core region present within OE1 is critical for neuronal expression in the LHA and is likely to be made up of multiple *cis*-regulatory modules (Moriguchi et al., 2002). A closer analysis of the fugu *ORX* 5' flanking sequence for conserved motifs helped me to uncover a 50 bp element orthologous to the 57 bp core of OE1 and located 322 bp upstream of fugu *ORX*. This fugu OE1 core element and the mammalian OE1 core only shared a 44% identity and hence it was not picked up in the 60% identity threshold I used for MLAGAN. In the same way, a fugu ortholog of OE2 was also detected a further 1 kb upstream of OE1 at 45% identity with the mammalian OE2 element.

The presence of sequences with similarity to mouse OE1 and OE2 elements in the fugu ORX locus despite their low conservation, suggest that they could be responsible for directing expression of fugu *ORX* in LHA-specific neurons in the same way as in human and mouse. This is supported by the observation that when transgenic constructs containing the fugu OE1 and OE2 elements (N-f*ORX*-K and fOE2OE1-lac) directed transgene expression to neurons in the LHA in which colocalization of the endogenous mouse gene expression was also observed. While both transgene constructs were able to drive expression to the LHA, the expression levels were significantly lower than that observed with the entire cosmid 33B9, and there was leaky expression of the transgene in some neurons in the LHA that do not express the mouse *ORX*.. These results suggest that the fugu cosmid contained additional enhancer elements outside the sequences of the two shorter constructs used. At the same time they also indicate the fugu cosmid contained silencers that helped to restrict the expression to only the LHA neurons. These additional enhancer and silencer elements could be present in the sequences upstream of the *ORX* intergenic sequence used in construct N-f*ORX*-K, in the introns or 3’ flanking region of fugu *ORX* that were not investigated in this study.

The fugu OE1 and OE2 elements are likely to contain regulatory motifs crucial for LHA-specific expression similar to their human orthologs (Moriguchi et al., 2002) since they have some functional features in common with the human OE1 and OE2 elements. Firstly, fugu OE1 and OE2 worked cooperatively both in the context of the fugu *ORX* locus (N-f*ORX*-K) as well as independently of the *ORX* locus (fOE2OE1-lac) to direct specific expression to LHA neurons (Figures 23D and 23F). Secondly, fugu OE1 in concert with the basal promoter (fOE1-lac) was not sufficient to reproduce endogenous gene expression in the LHA. However no ectopic expression was observed, unlike what was observed in the deletion analysis of Moriguchi et al (2002). Thus the fugu *ORX* basal promoter containing OE1 was unable to specify any expression in the hypothalamus, and required OE2 as an enhancer to interact with the promoter to specify LHA-specific expression. Deletion analysis of the fugu OE1 should be able to shed more light on the function of the regulatory motifs conserved between fugu and mammalian OE1, as well as between fugu and zebrafish (Faraco et al., 2006) and will be the subject for follow up work of the present study. In any case, my experiments have shown that OE1 required for LHA-specific expression did not arise only in mammals as previously hypothesized by

Moriguchi et al. (2002). The presence of elements orthologous to OE1 and OE2 in fugu has clearly shown that these elements are indeed ancient and were present in the common ancestor of mammals and fishes. These elements are therefore likely to be conserved in all bony vertebrates.

The zebrafish study by Faraco et al. (2006) concluded that *ORX* regulation in mammals was different from that in zebrafish. This study found that the 1 kb zebrafish *ORX* 5'flanking region could direct expression to the LHA in transgenic zebrafish, with a 250 bp segment within this region containing a 13 bp core critical for expression. However there was no striking homology between this 13 bp core element and the 57 bp core of mammalian OE1; the 3.2 kb human *ORX* regulatory region could not specify any LHA expression in transgenic zebrafish; and deletion analysis of the zebrafish *ORX* regulatory region decreased efficiency but not specificity in directing LHA-specific expression in zebrafish (Faraco et al., 2006). My analysis of the fugu *ORX* regulatory region for conserved motifs has allowed me to locate the 57-bp core element in OE1, as well as two of the four regulatory motifs conserved between zebrafish and Tetraodon in the 250 bp segment crucial for LHA expression in zebrafish including the critical 13 bp core. It is striking that both the critical mammalian-like 57-bp OE1 core element and zebrafish 13-bp core elements could be found in the fugu OE1 element albeit at low conservation. Since ectopic expression in the medial regions of the hypothalamus was not observed in the zebrafish study as well as in my transgenic analysis of the fugu *ORX* regulatory region, perhaps what has newly evolved in the mammalian system are enhancers that direct expression to those areas, and OE1 and OE2 in the mammalian *ORX* locus have taken on the dual roles of enhancing LHA-specific expression as part of their ancestral function, as well as silencing medial hypothalamic expression as part of their novel function that is not present in teleost fishes. Since both enhancer and silencer functions are intricately linked, the mammalian OE1 and OE2 elements probably recruit a slightly different set of transcription factors compared to the fish OE1 and OE2 elements, and not all these factors might be present in the zebrafish hypothalamic system to sufficiently activate the mammalian OE1 and OE2 elements to give rise to transgene expression.

In summary my work has demonstrated that in the absence of high sequence similarity, mammalian and fish species share the *cis*-regulatory information necessary for LHA-specific expression of *ORX* gene. Similar functional conservation of enhancers across distant phylogenetic groups in the absence of apparent sequence conservation has been recently reported for the *RET* receptor tyrosine kinase-encoding gene locus (Fisher et al., 2006) and the propiomelanocortin (*POMC*) gene locus (Bumaschny et al., 2007). Like the *ORX* locus, this conservation has been attributed to short functional regulatory motifs (4-20 bp) that are undetectable by the criterion (70% identity across >100 bp) used by Fisher et al. (2006) for identifying CNEs in distant vertebrate genomes. The enhancers identified using the CNE criterion seem to be found mainly in transcription factor genes involved in development as demonstrated for *Six3* and *Foxb1*. Thus vertebrate genes seems to contain two distinct types of enhancers: one that is highly conserved over long stretches of DNA and associated with developmental regulators and another one associated with downstream genes in the gene-regulatory networks such as those encoding hormones in which only short elements are conserved.

Chapter 7

General Discussion

7.1 Summary

The identification of evolutionarily constrained sequences is frequently used as part of a battery of approaches to identify and characterize functional sequences like *cis*-regulatory elements in the human genome. This is because functional sequences are likely to be under selective constraint and therefore tend to evolve slowly compared to the nonfunctional sequences flanking them. The objective of comparative genomics is to identify such constrained sequences by comparing genomes that are phylogenetically related. The longer the phylogenetic distance, the higher are the chances that the neutrally evolving regions would have diverged completely leaving behind footprints of highly conserved sequences representing functional elements. For identifying regulatory elements in the human genome, comparison with fishes is particularly attractive because fishes diverged about 420 million years ago from the mammalian lineage and thus represent the most distantly related bony vertebrates. With this in mind, I used a multiple alignment algorithm MLAGAN to compare 50 human forebrain-genes with their orthologs in mouse and fugu. My analysis identified 206 conserved noncoding elements (CNEs) that are longer than 50 bp and exhibit more than 60% identity. These CNEs are associated with 29 of the forebrain-genes analyzed. From these 206 CNEs, I validated 13 CNEs associated with two developmental genes (*Six3* and *Foxb1*) and found that all of them functioned as *cis*-regulatory elements, either as enhancers or silencers, directing spatial and temporal-specific expression of the genes associated with them. The 100% success rate of the functional assay of the 13 CNEs suggests that the remaining CNEs are also most likely to be *cis*-regulatory elements. This work has therefore demonstrated that comparisons between distant vertebrates like mammals and fish is a reliable approach for identifying functional *cis*-regulatory elements in the human genome. My analysis did not identify CNEs in 21 of the 50 genes. However, functional analysis of the regulatory regions of one such gene (*ORX*) showed that although the regulatory elements in this gene locus are not highly conserved to qualify for being identified as a CNE, the regulatory elements are indeed conserved between mammals and fish at levels lower than the criteria used for defining CNEs. This indicates that lack of CNE should not be construed as an indication of nonfunctional sequences or that the mammalian and fish genes are regulated using different mechanisms. Indeed, in-depth functional annotation of 1% of the human genome by the ENCODE project has revealed that while a large number of experimentally determined functional noncoding elements are under evolutionary constraint, many are unconstrained across mammals (Birney et al., 2007; Margulies et al., 2007).

7.2 High-success rate in identifying functional *cis*-regulatory elements

In my study all the 13 CNEs assayed in transgenic mice were found to function as *cis*-regulatory elements. In typical CNE assays in transgenic mice, CNEs are cloned upstream of a β-galactosidase reporter linked to a minimal promoter from the mouse *hsp68* gene and the function is assayed at one (E11.5) or two developmental stages. The success rate in such assays has been found to be about 29% for human-fugu conserved elements and 61% for human-fugu ultraconserved elements that acted as tissue-specific enhancers at E11.5 (Pennacchio et al., 2006). The high success rate in my assay can be attributed to the following: instead of the basal promoter from *hsp68* gene, I used the basal promoter from the same gene, which should give high-specificity for interaction between the enhancer and the basal promoter. The basal promoter has been shown to play an important and specific role in mediating the functions of enhancers (Smale, 2001). Therefore, the expression pattern of an enhancer driven by a heterologous basal promoter may not always reflect the physiological level or the actual expression pattern of the enhancers *in vivo*. My work has also clearly demonstrated that the basal promoters (about 400-800 bp around the transcription start site) for *Six3* and *Foxb1* contain *cis*-regulatory elements directing tissue-specific expression. The CNEs dispersed in the intergenic regions interact with the basal promoter to drive tissue- and developmental-stage specific expression of the target gene. The high success rate in my assay might be due to the homologous basal promoters used in my study. One downside of this approach is that the expression levels of the reporter gene in my assays were generally lower than that observed with *hsp68* promoter and as a result I had to use the more sensitive technique of *in situ* hybridization to detect the mRNA of the *lacZ* reporter gene, instead of β-galactosidase staining used in studies with *hsp68* promoter to detect expression of the reporter gene. The two homologous promoters used here seem to drive ß-galactosidase expression at lower efficiency during embryonic development than the *hsp68* promoter. They are too weak a promoter to produce significant observable ß-galactosidase expression but can nevertheless drive detectable levels of RNA expression in a tissue-specific and a developmental stage-specific way as shown in this study.

Another reason for the high success rate in my assays could be that I checked the expression level at several stages of development from E9.5 to E15.5. Since many of the CNEs showed regulatory function in a temporal-specific manner, this approach was useful to identify their function at different stages of development. A third reason for the high success rate might be due to assay of functions of some CNEs in clusters instead of testing them individually. However the analysis of a cluster of CNEs on basal promoter activity may not always recapitulate their physiological roles in the context of the intact gene. A complementary approach of deleting individual elements from the gene locus through specific targeting of CNEs by homologous recombination would be useful follow-up work to ascertain the precise physiological contribution of each enhancer to the tissue and temporal-specific expression of the gene.

7.3 Cooperativity and redundancy in *cis*-regulatory elements

Analysis of multiple CNEs from the *Six3* and *Foxb1* loci presented an opportunity to understand the interactions between CNEs. For example, I found that the basal promoters of both genes contain *cis*-regulatory elements and contribute significantly to the spatio-temporal expression of the genes. Another interesting finding is that the *cis*-regulatory elements exhibit a high degree of cooperativity in their function. For example, CNE5/6/7 and CNE12 cooperate to inhibit lens expression directed by the basal promoter at E15.5 to help maintain a physiologically appropriate level of *Six3* expression in the eye and suppress ectopic expression. Another interesting aspect that emerged from my study is the high level of redundancy among the multiple enhancers associated with a gene. Several enhancers were found to enhance expression or suppress expression in the same tissue at a particular developmental stage. For example, I discovered at least five silencers (CNE1, CNE2/3/4, CNE5/6/7, CNE8/9 and CNE14) that suppressed optic vesicle expression directed by the mouse *Six3* basal promoter during E10.5-E11.5. This indicates that the regulatory codes have a built-in redundancy to ensure that the genes are tightly regulated to obtain the correct expression patterns. The redundancy in the regulatory code also allows mutations and selections to act on redundant enhancers to acquire novel expression patterns. Evolutionary changes to *cis*-regulatory elements have been shown to have a high potential for morphological innovations and adaptive evolution (Wray, 2007). The modular nature and redundancy of *cis*-regulatory elements make them attractive template for adaptive evolution.

7.4 Conserved function of *cis*-regulatory elements in mammals and fish without apparent sequence conservation

My study has shown that the functions of regulatory elements could be conserved in distant vertebrates even though the sequences do not exhibit apparent conservation. I failed to identify any CNEs in the orexin (*ORX*) gene locus but functional analysis of the regulatory region of the fugu *ORX* showed that the function of the enhancer is highly conserved between fish and mammals. This means that functional information is conserved in these vertebrate sequences at levels below the metric used for identifying CNEs. Indeed a recent study comprehensively analyzed both conserved and nonconserved regions around the zebrafish paired-like homeobox gene *phox2b* for enhancer activity and found that many regulatory sequences (42-51%) are not detectable using standard methods for detecting evolutionary constraint (McGaughey et al., 2008). In addition, nonaligned sequences in the *phox2b* locus were shown to contain conserved transcription factor binding sites that would discriminate them from nonfunctional sequences, but these are distributed at a low density that makes them hard to detect by alignment alone (McGaughey et al., 2008). It is likely that orthologous *cis*-regulatory elements control the expression of these genes but these elements have evolved beyond recognition through small changes in transcription factor binding sites, rearrangement of these binding sites or multiple coevolved changes that give rise to compensatory mutations along the enhancer as a result of a stabilizing selection process (Fisher et al., 2006; Ludwig et al., 2000). Such weak constraint on functional sequence could be attributed to sequence degeneracy of binding sites or redundancy of individual functional elements or the need for secondary structure that is indirectly related to primary sequence (Cooper and Brown, 2008). The *cis*-regulatory elements of *ORX* gene locus belongs to one of these categories of enhancers.

Enhancers have been classified into two models: the first is called the “enhanceosome model”, which describes enhancers as highly structured with a precise arrangement of transcription factor binding sites. The enhanceosome features a high degree of cooperativity between enhancer-bound proteins, such that alterations in individual binding sites can have drastic effects on enhancer output (Arnosti and Kulkarni, 2005). The high degree of conservation seen in the *cis*-regulatory elements of genes that encode transcription factors and involved in development like *Six3* and *Foxb1* indicate that they follow the enhanceosome model. The second model is the “billboard model” which describes enhancers as unstructured and representing loose assemblies of transcription factor binding sites that can vary in orientation and spacing. A billboard enhancer displays potential transcriptional information that is interpreted and deciphered by interaction with the basal transcriptional machinery, and exact positioning of bound transcription factors is less critical than with an enhanceosome (Arnosti and Kulkarni, 2005). Billboard enhancers are more evolutionarily pliable than enhanceosomes and can include extreme sequence and binding site divergence between functionally equivalent enhancers (Hare et al., 2008). This type of enhancers are likely to be present in the *ORX* regulatory regions characterized in this study, as well as in many other gene loci in which I could not detect any CNEs.

In summary, no single metric of conservation can satisfactorily identify all the *cis*-regulatory elements in the human genome. Although sequence conservation is a useful sign for identifying functional *cis*-regulatory elements, lack of conservation does not imply such noncoding sequences do not have a function. Functional *cis*-regulatory sequences in such regions can be more efficiently identified by comparing closely and distantly related vertebrates and by looking for conserved transcription factor-binding sites. Thus, a combination of several strategies is required for the identification of all the *cis*-regulatory elements in the human genome.

References

Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L.A., and Rubin, E.M. (2007). Deletion of ultraconserved elements yields viable mice. PLoS Biol *5*, e234.

Alvarez-Bolado, G., Zhou, X., Cecconi, F., and Gruss, P. (2000). Expression of Foxb1 reveals two strategies for the formation of nuclei in the developing ventral diencephalon. Dev Neurosci *22*, 197-206.

Ang, S.L., Wierda, A., Wong, D., Stevens, K.A., Cascio, S., Rossant, J., and Zaret, K.S. (1993). The formation and maintenance of the definitive endoderm lineage in the mouse: involvement of HNF3/forkhead proteins. Development *119*, 1301-1315.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., *et al.* (2002). Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science *297*, 1301-1310.

Aparicio, S., Morrison, A., Gould, A., Gilthorpe, J., Chaudhuri, C., Rigby, P., Krumlauf, R., and Brenner, S. (1995). Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. Proc Natl Acad Sci U S A *92*, 1684-1688.

Archer, Z.A., Findlay, P.A., Rhind, S.M., Mercer, J.G., and Adam, C.L. (2002). Orexin gene expression and regulation by photoperiod in the sheep hypothalamus. Regul Pept *104*, 41-45.

Arnosti, D.N., and Kulkarni, M.M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J Cell Biochem *94*, 890-898.

Bagheri-Fam, S., Barrionuevo, F., Dohrmann, U., Gunther, T., Schule, R., Kemler, R., Mallo, M., Kanzler, B., and Scherer, G. (2006). Long-range upstream and downstream enhancers control distinct subsets of the complex spatiotemporal Sox9 expression pattern. Dev Biol *291*, 382-397.

Bagheri-Fam, S., Ferraz, C., Demaille, J., Scherer, G., and Pfeifer, D. (2001). Comparative genomics of the SOX9 region in human and Fugu rubripes: conservation of short regulatory sequence elements within large intergenic regions. Genomics *78*, 73-82.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321-1325.

Bergman, C.M., and Kreitman, M. (2001). Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. Genome Res *11*, 1335-1345.

Berman, B.P., Pfeiffer, B.D., Laverty, T.R., Salzberg, S.L., Rubin, G.M., Eisen, M.B., and Celniker, S.E. (2004). Computational identification of developmental enhancers:

conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. Genome Biol *5*, R61.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R., *et al.* (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell *120*, 169-181.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., *et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816.

Blanchette, M., Bataille, A.R., Chen, X., Poitras, C., Laganiere, J., Lefebvre, C., Deblois, G., Giguere, V., Ferretti, V., Bergeron, D., *et al.* (2006). Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res *16*, 656-668.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science *299*, 1391-1394.

Bovolenta, P., Mallamaci, A., Puelles, L., and Boncinelli, E. (1998). Expression pattern of cSix3, a member of the Six/sine oculis family of transcription factors. Mech Dev *70*, 201-203.

Bray, N., Dubchak, I., and Pachter, L. (2003). AVID: A global alignment program. Genome Res *13*, 97-102.

Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B., and Aparicio, S. (1993). Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. Nature *366*, 265-268.

Broglio, C., Gomez, A., Duran, E., Ocana, F.M., Jimenez-Moya, F., Rodriguez, F., and Salas, C. (2005). Hallmarks of a common forebrain vertebrate plan: specialized pallial areas for spatial, temporal and emotional memory in actinopterygian fish. Brain Res Bull *66*, 277-281.

Brudno, M., Chapman, M., Gottgens, B., Batzoglou, S., and Morgenstern, B. (2003a). Fast and sensitive multiple alignment of large genomic sequences. BMC Bioinformatics *4*, 66.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. (2003b). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res *13*, 721-731.

Bumaschny, V.F., de Souza, F.S., Lopez Leal, R.A., Santangelo, A.M., Baetscher, M., Levi, D.H., Low, M.J., and Rubinstein, M. (2007). Transcriptional regulation of pituitary

POMC is conserved at the vertebrate extremes despite great promoter sequence divergence. Mol Endocrinol *21*, 2738-2749.

Carl, M., Loosli, F., and Wittbrodt, J. (2002). Six3 inactivation reveals its essential role for the formation and patterning of the vertebrate eye. Development *129*, 4057-4063.

Chemelli, R.M., Willie, J.T., Sinton, C.M., Elmquist, J.K., Scammell, T., Lee, C., Richardson, J.A., Williams, S.C., Xiong, Y., Kisanuki, Y., *et al.* (1999). Narcolepsy in orexin knockout mice: molecular genetics of sleep regulation. Cell *98*, 437-451.

Chen, K., and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet *8*, 93-103.

Christoffels, A., Koh, E.G., Chia, J.M., Brenner, S., Aparicio, S., and Venkatesh, B. (2004). Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol Biol Evol *21*, 1146-1151.

Consortium, I.H.G.S. (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931-945.

Conte, I., and Bovolenta, P. (2007). Comprehensive characterization of the cis-regulatory code responsible for the spatio-temporal expression of olSix3.2 in the developing medaka forebrain. Genome Biol *8*, R137.

Cooper, G.M., and Brown, C.D. (2008). Qualifying the relationship between sequence conservation and molecular function. Genome Res *18*, 201-205.

Crawford, G.E., Holt, I.E., Mullikin, J.C., Tai, D., Blakesley, R., Bouffard, G., Young, A., Masiello, C., Green, E.D., Wolfsberg, T.G., *et al.* (2004). Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc Natl Acad Sci U S A *101*, 992-997.

Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D., Davis, S., Margulies, E.H., Chen, Y., Bernat, J.A., Ginsburg, D., *et al.* (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res *16*, 123-131.

Cutler, D.J., Morris, R., Sheridhar, V., Wattam, T.A., Holmes, S., Patel, S., Arch, J.R., Wilson, S., Buckingham, R.E., Evans, M.L., *et al.* (1999). Differential distribution of orexin-A and orexin-B immunoreactivity in the rat brain and spinal cord. Peptides *20*, 1455-1470.

de la Calle-Mustienes, E., Feijoo, C.G., Manzanares, M., Tena, J.J., Rodriguez-Seguel, E., Letizia, A., Allende, M.L., and Gomez-Skarmeta, J.L. (2005). A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. Genome Res *15*, 1061-1072.

de Lecea, L., Kilduff, T.S., Peyron, C., Gao, X., Foye, P.E., Danielson, P.E., Fukuhara, C., Battenberg, E.L., Gautvik, V.T., Bartlett, F.S., 2nd, *et al.* (1998). The hypocretins: hypothalamus-specific peptides with neuroexcitatory activity. Proc Natl Acad Sci U S A *95*, 322-327.

Dermitzakis, E.T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B.J., Flegel, V., Bucher, P., Jongeneel, C.V., *et al.* (2002). Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature *420*, 578-582.

Dorschner, M.O., Hawrylycz, M., Humbert, R., Wallace, J.C., Shafer, A., Kawamoto, J., Mack, J., Hall, R., Goldy, J., Sabo, P.J., *et al.* (2004). High-throughput localization of functional elements by quantitative chromatin profiling. Nat Methods *1*, 219-225.

Dou, C., Ye, X., Stewart, C., Lai, E., and Li, S.C. (1997). TWH regulates the development of subsets of spinal cord neurons. Neuron *18*, 539-551.

Dyer, C.J., Touchette, K.J., Carroll, J.A., Allee, G.L., and Matteri, R.L. (1999). Cloning of porcine prepro-orexin cDNA and effects of an intramuscular injection of synthetic porcine orexin-B on feed intake in young pigs. Domest Anim Endocrinol *16*, 145-148.

Elnitski, L., Jin, V.X., Farnham, P.J., and Jones, S.J. (2006). Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res *16*, 1455-1464.

Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S., Wiebe, V., Kitano, T., Monaco, A.P., and Paabo, S. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. Nature *418*, 869-872.

Faraco, J.H., Appelbaum, L., Marin, W., Gaus, S.E., Mourrain, P., and Mignot, E. (2006). Regulation of hypocretin (orexin) expression in embryonic zebrafish. J Biol Chem *281*, 29753-29761.

Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L., and McCallion, A.S. (2006). Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science *312*, 276-279.

Frazer, K.A., Elnitski, L., Church, D.M., Dubchak, I., and Hardison, R.C. (2003). Cross-species sequence comparisons: a review of methods and available resources. Genome Res *13*, 1-12.

Ghanem, N., Jarinova, O., Amores, A., Long, Q., Hatch, G., Park, B.K., Rubenstein, J.L., and Ekker, M. (2003). Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters. Genome Res *13*, 533-543.

Granadino, B., Gallardo, M.E., Lopez-Rios, J., Sanz, R., Ramos, C., Ayuso, C., Bovolenta, P., and Rodriguez de Cordoba, S. (1999). Genomic cloning, structure, expression pattern, and chromosomal location of the human SIX3 gene. Genomics *55*, 100-105.

Griffin, C., Kleinjan, D.A., Doe, B., and van Heyningen, V. (2002). New 3' elements control Pax6 expression in the developing pretectum, neural retina and olfactory region. Mech Dev *112*, 89-100.

Grinblat, Y., Gamse, J., Patel, M., and Sive, H. (1998). Determination of the zebrafish forebrain: induction and patterning. Development *125*, 4403-4416.

Hanlon, S.E., and Lieb, J.D. (2004). Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. Curr Opin Genet Dev *14*, 697-705.

Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., *et al.* (2003). Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res *13*, 13-26.

Hare, E.E., Peterson, B.K., Iyer, V.N., Meier, R., and Eisen, M.B. (2008). Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet *4*, e1000106.

Holland, L.Z., and Holland, N.D. (1999). Chordate origins of the vertebrate central nervous system. Curr Opin Neurobiol *9*, 596-602.

Hudson, M.E., and Snyder, M. (2006). High-throughput methods of regulatory element discovery. Biotechniques *41*, 673, 675, 677 passim.

Jeong, Y., and Epstein, D.J. (2003). Distinct regulators of Shh transcription in the floor plate and notochord indicate separate origins for these tissues in the mouse node. Development *130*, 3891-3902.

Johren, O., Neidert, S.J., Kummer, M., Dendorfer, A., and Dominiak, P. (2001). Prepro-orexin and orexin receptor mRNAs are differentially expressed in peripheral tissues of male and female rats. Endocrinology *142*, 3324-3331.

Juan, A.H., and Ruddle, F.H. (2003). Enhancer timing of Hox gene expression: deletion of the endogenous Hoxc8 early enhancer. Development *130*, 4823-4834.

Kaestner, K.H., Schutz, G., and Monaghan, A.P. (1996). Expression of the winged helix genes fkh-4 and fkh-5 defines domains in the central nervous system. Mech Dev *55*, 221-230.

Kammermeier, L., and Reichert, H. (2001). Common developmental genetic mechanisms for patterning invertebrate and vertebrate brains. Brain Res Bull *55*, 675-682.

Kaslin, J., Nystedt, J.M., Ostergard, M., Peitsaro, N., and Panula, P. (2004). The orexin/hypocretin system in zebrafish is connected to the aminergic and cholinergic systems. J Neurosci *24*, 2678-2689.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. Nature *436*, 876-880.

Kimura-Yoshida, C., Kitajima, K., Oda-Ishii, I., Tian, E., Suzuki, M., Yamamoto, M., Suzuki, T., Kobayashi, M., Aizawa, S., and Matsuo, I. (2004). Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. Development *131*, 57-71.

Kleinjan, D.A., Seawright, A., Mella, S., Carr, C.B., Tyas, D.A., Simpson, T.I., Mason, J.O., Price, D.J., and van Heyningen, V. (2006). Long-range downstream enhancers are essential for Pax6 expression. Dev Biol *299*, 563-581.

Kleinjan, D.A., Seawright, A., Schedl, A., Quinlan, R.A., Danes, S., and van Heyningen, V. (2001). Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. Hum Mol Genet *10*, 2049-2059.

Kleinjan, D.A., and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet *76*, 8-32.

Kloetzli, J.M., Fontaine-Glover, I.A., Brown, E.R., Kuo, M., and Labosky, P.A. (2001). The winged helix gene, Foxb1, controls development of mammary glands and regions of the CNS that regulate the milk-ejection reflex. Genesis *29*, 60-71.

Kobayashi, M., Toyama, R., Takeda, H., Dawid, I.B., and Kawakami, K. (1998). Overexpression of the forebrain-specific homeobox gene six3 induces rostral forebrain enlargement in zebrafish. Development *125*, 2973-2982.

Labosky, P.A., Winnier, G.E., Jetton, T.L., Hargett, L., Ryan, A.K., Rosenfeld, M.G., Parlow, A.F., and Hogan, B.L. (1997). The winged helix gene, Mf3, is required for normal development of the diencephalon and midbrain, postnatal growth and the milk-ejection reflex. Development *124*, 1263-1274.

Lagutin, O., Zhu, C.C., Furuta, Y., Rowitch, D.H., McMahon, A.P., and Oliver, G. (2001). Six3 promotes the formation of ectopic optic vesicle-like structures in mouse embryos. Dev Dyn *221*, 342-349.

Lagutin, O.V., Zhu, C.C., Kobayashi, D., Topczewski, J., Shimamura, K., Puelles, L., Russell, H.R., McKinnon, P.J., Solnica-Krezel, L., and Oliver, G. (2003). Six3 repression of Wnt signaling in the anterior neuroectoderm is essential for vertebrate forebrain development. Genes Dev *17*, 368-379.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lehmann, O.J., Sowden, J.C., Carlsson, P., Jordan, T., and Bhattacharya, S.S. (2003). Fox's in development and disease. Trends Genet *19*, 339-344.

Lettice, L.A., Heaney, S.J., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet *12*, 1725-1735.

Lettice, L.A., Horikoshi, T., Heaney, S.J., van Baren, M.J., van der Linde, H.C., Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B.A., Endo, N., *et al.* (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci U S A *99*, 7548-7553.

Lin, L., Faraco, J., Li, R., Kadotani, H., Rogers, W., Lin, X., Qiu, X., de Jong, P.J., Nishino, S., and Mignot, E. (1999). The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. Cell *98*, 365-376.

Liu, W., Lagutin, O.V., Mende, M., Streit, A., and Oliver, G. (2006). Six3 activation of Pax6 expression is essential for mammalian lens induction and specification. Embo J *25*, 5383-5395.

Loosli, F., Koster, R.W., Carl, M., Krone, A., and Wittbrodt, J. (1998). Six3, a medaka homologue of the Drosophila homeobox gene sine oculis is expressed in the anterior embryonic shield and the developing eye. Mech Dev *74*, 159-164.

Loosli, F., Winkler, S., and Wittbrodt, J. (1999). Six3 overexpression initiates the formation of ectopic retina. Genes Dev *13*, 649-654.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science *288*, 136-140.

Lopez-Rios, J., Tessmar, K., Loosli, F., Wittbrodt, J., and Bovolenta, P. (2003). Six3 and Six6 activity is modulated by members of the groucho family. Development *130*, 185-195.

Ludwig, M.Z., Bergman, C., Patel, N.H., and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. Nature *403*, 564-567.

Ma, B., Tromp, J., and Li, M. (2002). PatternHunter: faster and more sensitive homology search. Bioinformatics *18*, 440-445.

Margulies, E.H., Blanchette, M., Haussler, D., and Green, E.D. (2003). Identification and characterization of multi-species conserved sequences. Genome Res *13*, 2507-2518.

Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A., Birney, E., Keefe, D., Schwartz, A.S., Hou, M., *et al.* (2007). Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res *17*, 760-774.

Margulies, E.H., Vinson, J.P., Miller, W., Jaffe, D.B., Lindblad-Toh, K., Chang, J.L., Green, E.D., Lander, E.S., Mullikin, J.C., and Clamp, M. (2005). An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc Natl Acad Sci U S A *102*, 4795-4800.

Markstein, M., Zinzen, R., Markstein, P., Yee, K.P., Erives, A., Stathopoulos, A., and Levine, M. (2004). A regulatory code for neurogenic gene expression in the Drosophila embryo. Development *131*, 2387-2394.

Mathis, L., and Nicolas, J.F. (2006). Clonal origin of the mammalian forebrain from widespread oriented mixing of early regionalized neuroepithelium precursors. Dev Biol *293*, 53-63.

McGaughey, D.M., Vinton, R.M., Huynh, J., Al-Saif, A., Beer, M.A., and McCallion, A.S. (2008). Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. Genome Res *18*, 252-260.

Medina, L., Brox, A., Legaz, I., Garcia-Lopez, M., and Puelles, L. (2005). Expression patterns of developmental regulatory genes show comparable divisions in the telencephalon of Xenopus and mouse: insights into the evolution of the forebrain. Brain Res Bull *66*, 297-302.

Mercer, E.H., Hoyle, G.W., Kapur, R.P., Brinster, R.L., and Palmiter, R.D. (1991). The dopamine beta-hydroxylase gene promoter directs expression of E. coli lacZ to sympathetic and other neurons in adult transgenic mice. Neuron *7*, 703-716.

Metin, C., Alvarez, C., Moudoux, D., Vitalis, T., Pieau, C., and Molnar, Z. (2007). Conserved pattern of tangential neuronal migration during forebrain development. Development *134*, 2815-2827.

Mieda, M., Willie, J.T., Hara, J., Sinton, C.M., Sakurai, T., and Yanagisawa, M. (2004). Orexin peptides prevent cataplexy and improve wakefulness in an orexin neuron-ablated model of narcolepsy in mice. Proc Natl Acad Sci U S A *101*, 4649-4654.

Miller, W., Rosenbloom, K., Hardison, R.C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D.C., Baertsch, R., Blankenberg, D., *et al.* (2007). 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res *17*, 1797-1808.

Moriguchi, T., Sakurai, T., Nambu, T., Yanagisawa, M., and Goto, K. (1999). Neurons containing orexin in the lateral hypothalamic area of the adult rat brain are activated by insulin-induced acute hypoglycemia. Neurosci Lett *264*, 101-104.

Moriguchi, T., Sakurai, T., Takahashi, S., Goto, K., and Yamamoto, M. (2002). The human prepro-orexin gene regulatory region that activates gene expression in the lateral region and represses it in the medial regions of the hypothalamus. J Biol Chem *277*, 16985-16992.

Muller, F., Chang, B., Albert, S., Fischer, N., Tora, L., and Strahle, U. (1999). Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord. Development *126*, 2103-2116.

Nambu, T., Sakurai, T., Mizukami, K., Hosoya, Y., Yanagisawa, M., and Goto, K. (1999). Distribution of orexin neurons in the adult rat brain. Brain Res *827*, 243-260.

Nobrega, M.A., Ovcharenko, I., Afzal, V., and Rubin, E.M. (2003). Scanning human gene deserts for long-range enhancers. Science *302*, 413.

Ohkubo, T., Boswell, T., and Lumineau, S. (2002). Molecular cloning of chicken prepro-orexin cDNA and preferential expression in the chicken hypothalamus. Biochim Biophys Acta *1577*, 476-480.

Oliver, G., Mailhos, A., Wehr, R., Copeland, N.G., Jenkins, N.A., and Gruss, P. (1995). Six3, a murine homologue of the sine oculis gene, demarcates the most anterior border of the developing neural plate and is expressed during eye development. Development *121*, 4045-4055.

Ovcharenko, I., Loots, G.G., Giardine, B.M., Hou, M., Ma, J., Hardison, R.C., Stubbs, L., and Miller, W. (2005). Mulan: multiple-sequence local alignment and visualization for studying function and evolution. Genome Res *15*, 184-194.

Ovcharenko, I., Nobrega, M.A., Loots, G.G., and Stubbs, L. (2004). ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. Nucleic Acids Res *32*, W280-286.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., *et al.* (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*, 499-502.

Peyron, C., Tighe, D.K., van den Pol, A.N., de Lecea, L., Heller, H.C., Sutcliffe, J.G., and Kilduff, T.S. (1998). Neurons containing hypocretin (orexin) project to multiple neuronal systems. J Neurosci *18*, 9996-10015.

Pheasant, M., and Mattick, J.S. (2007). Raising the estimate of functional human sequences. Genome Res *17*, 1245-1253.

Pollard, K.S., Salama, S.R., King, B., Kern, A.D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J.S., Bejerano, G., Baertsch, R., *et al.* (2006). Forces shaping the fastest evolving regions in the human genome. PLoS Genet *2*, e168.

Poulin, F., Nobrega, M.A., Plajzer-Frick, I., Holt, A., Afzal, V., Rubin, E.M., and Pennacchio, L.A. (2005). In vivo characterization of a vertebrate ultraconserved enhancer. Genomics *85*, 774-781.

Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E.M., Couronne, O., and Pennacchio, L.A. (2006). Close sequence comparisons are sufficient to identify human cis-regulatory elements. Genome Res *16*, 855-863.

Prakash, A., and Tompa, M. (2005). Discovery of regulatory elements in vertebrates through comparative genomics. Nat Biotechnol *23*, 1249-1256.

Prakash, A., and Tompa, M. (2007). Measuring the accuracy of genome-size multiple alignments. Genome Biol *8*, R124.

Prober, D.A., Rihel, J., Onah, A.A., Sung, R.J., and Schier, A.F. (2006). Hypocretin/orexin overexpression induces an insomnia-like phenotype in zebrafish. J Neurosci *26*, 13400-13410.

Puelles, L., and Rubenstein, J.L. (2003). Forebrain gene expression domains and the evolving prosomeric model. Trends Neurosci *26*, 469-476.

Sackerson, C., Fujioka, M., and Goto, T. (1999). The even-skipped locus is contained in a 16-kb chromatin domain. Dev Biol *211*, 39-52.

Sakurai, T., Amemiya, A., Ishii, M., Matsuzaki, I., Chemelli, R.M., Tanaka, H., Williams, S.C., Richardson, J.A., Kozlowski, G.P., Wilson, S., *et al.* (1998). Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. Cell *92*, 573-585.

Sakurai, T., Moriguchi, T., Furuya, K., Kajiwara, N., Nakamura, T., Yanagisawa, M., and Goto, K. (1999). Structure and function of human prepro-orexin gene. J Biol Chem *274*, 17771-17776.

Sanges, R., Kalmar, E., Claudiani, P., D'Amato, M., Muller, F., and Stupka, E. (2006). Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. Genome Biol *7*, R56.

Santagati, F., Abe, K., Schmidt, V., Schmitt-John, T., Suzuki, M., Yamamura, K., and Imai, K. (2003). Identification of Cis-regulatory elements in the mouse Pax9/Nkx2-9 genomic region: implication for evolutionary conserved synteny. Genetics *165*, 235-242.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. Genome Res *13*, 103-107.

Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. (2000). PipMaker--a web server for aligning two genomic DNA sequences. Genome Res *10*, 577-586.

Seo, H.C., Drivenes, Ellingsen, S., and Fjose, A. (1998). Expression of two zebrafish homologues of the murine Six3 gene demarcates the initial eye primordia. Mech Dev *73*, 45-57.

Shibahara, M., Sakurai, T., Nambu, T., Takenouchi, T., Iwaasa, H., Egashira, S.I., Ihara, M., and Goto, K. (1999). Structure, tissue distribution, and pharmacological characterization of Xenopus orexins. Peptides *20*, 1169-1176.

Shin, J.T., Priest, J.R., Ovcharenko, I., Ronco, A., Moore, R.K., Burns, C.G., and MacRae, C.A. (2005). Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. Nucleic Acids Res *33*, 5437-5445.

Smale, S.T. (2001). Core promoters: active contributors to combinatorial gene regulation. Genes Dev *15*, 2503-2508.

Taylor, J.S., Braasch, I., Frickey, T., Meyer, A., and Van de Peer, Y. (2003). Genome duplication, a trait shared by 22000 species of ray-finned fish. Genome Res *13*, 382-390.

Tessmar-Raible, K., Raible, F., Christodoulou, F., Guy, K., Rembold, M., Hausen, H., and Arendt, D. (2007). Conserved sensory-neurosecretory cell types in annelid and fish forebrain: insights into hypothalamus evolution. Cell *129*, 1389-1400.

Thannickal, T.C., Moore, R.Y., Nienhuis, R., Ramanathan, L., Gulyani, S., Aldrich, M., Cornford, M., and Siegel, J.M. (2000). Reduced number of hypocretin neurons in human narcolepsy. Neuron *27*, 469-474.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., *et al.* (2003). Comparative analyses of multi-species sequences from targeted genomic regions. Nature *424*, 788-793.

Ureta-Vidal, A., Ettwiller, L., and Birney, E. (2003). Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet *4*, 251-262.

van Heyningen, V., and Williamson, K.A. (2002). PAX6 in sensory development. Hum Mol Genet *11*, 1161-1167.

Vandepoele, K., De Vos, W., Taylor, J.S., Meyer, A., and Van de Peer, Y. (2004). Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci U S A *101*, 1638-1643.

Venkatesh, B., Kirkness, E.F., Loh, Y.H., Halpern, A.L., Lee, A.P., Johnson, J., Dandona, N., Viswanathan, L.D., Tay, A., Venter, J.C., *et al.* (2006). Ancient noncoding elements conserved in the human genome. Science *314*, 1892.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., *et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.

Visel, A., Bristow, J., and Pennacchio, L.A. (2007). Enhancer identification through comparative genomics. Semin Cell Dev Biol *18*, 140-152.

Visel, A., Prabhakar, S., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E.M., and Pennacchio, L.A. (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet *40*, 158-160.

Wallis, D.E., Roessler, E., Hehr, U., Nanni, L., Wiltshire, T., Richieri-Costa, A., Gillessen-Kaesbach, G., Zackai, E.H., Rommens, J., and Muenke, M. (1999). Mutations in the homeodomain of the human SIX3 gene cause holoprosencephaly. Nat Genet *22*, 196-198.

Wang, Q.F., Prabhakar, S., Chanan, S., Cheng, J.F., Rubin, E.M., and Boffelli, D. (2007). Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons. Genome Biol *8*, R1.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Wehr, R., Mansouri, A., de Maeyer, T., and Gruss, P. (1997). Fkh5-deficient mice show dysgenesis in the caudal midbrain and hypothalamic mammillary body. Development *124*, 4447-4456.

Wilson, S.W., and Houart, C. (2004). Early steps in the development of the forebrain. Dev Cell *6*, 167-181.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., *et al.* (2005). Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol *3*, e7.

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nat Rev Genet *8*, 206-216.

Wu, J., Smith, L.T., Plass, C., and Huang, T.H. (2006). ChIP-chip comes of age for genome-wide functional analysis. Cancer Res *66*, 6899-6902.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338-345.

Zakany, J., Gerard, M., Favier, B., and Duboule, D. (1997). Deletion of a HoxD enhancer induces transcriptional heterochrony leading to transposition of the sacrum. Embo J *16*, 4393-4402.

Zerucha, T., Stuhmer, T., Hatch, G., Park, B.K., Long, Q., Yu, G., Gambarotta, A., Schultz, J.R., Rubenstein, J.L., and Ekker, M. (2000). A highly conserved enhancer in the Dlx5/Dlx6 intergenic region is the site of cross-regulatory interactions between Dlx genes in the embryonic forebrain. J Neurosci *20*, 709-721.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. J Comput Biol *7*, 203-214.

Zhao, T., Zhou, X., Szabo, N., Leitges, M., and Alvarez-Bolado, G. (2007). Foxb1-driven Cre expression in somites and the neuroepithelium of diencephalon, brainstem, and spinal cord. Genesis *45*, 781-787.

Zhou, X., Hollemann, T., Pieler, T., and Gruss, P. (2000). Cloning and expression of xSix3, the Xenopus homologue of murine Six3. Mech Dev *91*, 327-330.

Annex I: List of 50 genes expressed in the forebrain. The gene IDs of the human and fugu orthologs from Ensembl are indicated.

No	Gene	Symbol	Human Gene Ensembl ID	Fugu Gene Ensembl ID
1	Empty spiracles homeobox 1	EMX1	ENSG00000135638	SINFRUG00000136589
2	Aristaless related homeobox	ARX	ENSG00000004848	SINFRUG00000150852
3	Ventral anterior homeobox 1	VAX1	ENSG00000148704	SINFRUG00000120620
4	Orthodenticle homeobox 1	OTX1	ENSG00000115507	SINFRUG00000156103
5	Retina and anterior neural fold homeobox	RAX	ENSG00000134438	SINFRUG00000147714 SINFRUG00000136200
6	Orthopedia homeobox	OTP	ENSG00000171540	SINFRUG00000129005
7	GS homeobox 1	GSH1	ENSG00000169840	SINFRUG00000149945
8	GS homeobox 2	GSH2	ENSG00000180613	SINFRUG00000126231
9	Paired-like homeodomain 2	PITX2	ENSG0000016409	SINFRUG00000155006
10	Sine oculis-related homeobox homolog 3	SIX3	ENSG00000138083	SINFRUG00000147597
11	Sine oculis-related homeobox homolog 6	SIX6	ENSG00000184302	SINFRUG00000149651
12	Cartilage paired-class homeoprotein 1	CART1	ENSG00000180318	SINFRUG00000145309
13	LIM homeobox 2	LHX2	ENSG00000106689	SINFRUG00000135058
14	LIM homeobox 5	LHX5	ENSG00000089116	SINFRUG00000159859
15	LIM homeobox 6	LHX6	ENSG00000106852	SINFRUG00000147876 SINFRUG00000127105
16	LIM homeobox 8	LHX7/ LHX8	ENSG00000162624	SINFRUG00000136556
17	POU class 3 homeobox 3	BRN1/ POU3f3	ENSG00000198914	SINFRUG00000124122 SINFRUG00000163366
18	POU class 3 homeobox 2	BRN2/ POU3f2	ENSG00000184486	SINFRUG00000149835 SINFRUG00000160476
19	Transducin-like enhancer of split 1	TLE1	ENSG00000196781	SINFRUG00000125941
20	Single-minded homolog 1	SIM1	ENSG00000112246	SINFRUG00000127347
21	T-box brain gene 1	TBR1	ENSG00000136535	SINFRUG00000144384
22	Eomesodermin homolog	TBR2/ EOMES	ENSG00000163508	SINFRUG00000132983
23	cellular nucleic acid binding protein	CNBP1/ ZNF9	ENSG00000169714	SINFRUG00000126211

	isoform 1			
24	FEZ family zinc finger 2	FEZF2/ ZFP312	ENSG00000153266	SINFRUG00000146900
25	Zinc finger protein of the cerebellum 2	ZIC2	ENSG00000043355	SINFRUG00000151780
26	GLI-Kruppel family member isoform 2	GLI2	ENSG00000074047	SINFRUG00000153761 SINFRUG00000149811
27	GLI-Kruppel family member isoform 3	GLI3	ENSG00000106571	SINFRUG00000153715
28	Forkhead box G1	BF1/ FOXG1	ENSG00000176165	SINFRUG00000125793
29	Forkhead box B1	FOXB1/ FKH5	ENSG00000171956	SINFRUG00000139631
30	Forkhead box H1	FOXH1/ FAST1	ENSG00000160973	SINFRUG00000146944
31	Hypocretin (orexin) neuropeptide precursor	HCRT	ENSG00000161610	SINFRUG00000161995
32	Cholecystokinin preproprotein	CCK	ENSG00000187094	SINFRUG00000134679 SINFRUG00000141073
33	Neuropeptide Y	NPY	ENSG00000122585	SINFRUG00000144489
34	Agouti related protein	AGRP	ENSG00000159723	SINFRUG00000164565
35	Thyrotropin-releasing hormone	TRH	ENSG00000170893	SINFRUG00000125121
36	Somatostatin	SST	ENSG00000157005	SINFRUG00000143244
37	Cocaine and amphetamine regulated transcript	CART	ENSG00000164326	SINFRUG00000164538
38	Pro-melanin-concentrating hormone	PMCH	ENSG00000183395	SINFRUG00000145296
39	Calcitonin-related polypeptide alpha	CGRP/ CALCA	ENSG00000110680	SINFRUG00000141111 SINFRUG00000125998
40	Proenkephalin	PENK	ENSG00000181195	SINFRUG00000165185
41	Nerve growth factor (beta polypeptide)	NGFB	ENSG00000134259	SINFRUG00000139732 SINFRUG00000162576
42	Brain-derived neurotrophic factor	BDNF	ENSG00000176697	SINFRUG00000142602
43	Insulin-like growth factor 1	IGF1	ENSG00000017427	SINFRUG00000140885
44	Vasoactive intestinal peptide	VIP	ENSG00000146469	SINFRUG00000122509
45	Cryptochrome 1 (photolyase-like)	CRY1	ENSG00000008405	SINFRUG00000140891
46	Cryptochrome 2 (photolyase-like)	CRY2	ENSG00000121671	SINFRUG00000129038

47	Ring finger protein 111 / Arkadia	RNF111 /ARK	ENSG00000157450	SINFRUG00000134880
48	Noggin	NOG	ENSG00000183691	SINFRUG00000142423
49	Chordin	CHRD	ENSG00000090539	SINFRUG00000121889
50	TGFB-induced factor homeobox 1	TGIF	ENSG00000177426	SINFRUG00000139204