Dissertation

submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by:

Éva Kalmár
born in: Budapest, Hungary
Oral-examination: 05.03.2009.

Analysis of the cis-regulatory structure of developmentally regulated genes in zebrafish embryo

Referees: Prof. Dr.Uwe Strähle

. Prof. Dr.Jochen Wittbrodt

Abstract

Transcription regulation during vertebrate embryonic development is tightly regulated by cis-regulatory elements and respective transcription factor complexes, which bind to them. The interaction of these elements, followed by the recruitment of the RNA polymerase II machinery, leads to transcription initiation, which is one of the major regulatory steps in gene expression regulation. In this thesis I study three aspects of cis regulatory function in the zebrafish embryo:

1. Non-coding genomic sequences, in some cases with extreme evolutionary conservation, were shown to harbour enhancer function. After the completion of several mammalian and vertebrate genomes, phylogenetic footprinting became frequently used methods for cis-regulatory element identification. I present the identification of conserved noncoding sequences in the *pax2* locus and their in vivo test for enhancer activity in transient transgenic zebrafish embryos.

2. Conserved non-protein coding sequences working as enhancers were significantly enriched in and or around developmental regulators and/or transcription factor genes. In the second part of this thesis I present the application of a combined global and local alignment tool, which could identify higher number of conserved noncoding elements with enhancer activity, then any of the previous methods. Two thirds of the identified elements were shuffled during evolution. Although the majority of these shuffled conserved elements were still assigned to gene classes of transcription factors and developmental regulators, there were high enrichment in genes belonging to the extracellular regions and behavioural Gene Ontology classes.

3. The assignment of identified enhancers to their target gene promoters is often problematic, because of the potentially very large sequence distances separating them. Furthermore, based on recent results, promoters show an unexpected diversity. As promoter-enhancer interaction is mediated through multiprotein complexes, the composition of these complexes is likely dependent on the properties of the cis-regulatory elements involved and may result in interaction specificities. To investigate whether the DNA sequence of core promoters and enhancers define the specificity of their interaction, we have performed a high throughout screen, where 20 core promoters and 13 enhancers were used to generate 260 combinations. Data analysis after the automated image acquisition and processing revealed that enhancer function is clearly promoter-specific.

Table of content

Abbreviations

ACH	active chromatin hub
AP-1	activation protein 1
BRE	TFIIB recognition elements
CAGE	cap analysis of gene expression
CBF	CAAAT-box binding factor
CFP	cyan fluorescent protein
CNE	conserved non-coding element
CNS	central nervous system
CRM	cis-regulatory module
CTCF	CCCTC-binding factor
CTF	CAAAT-binding transcriptional factor
DCE	downstream core element
DPE	downstream promoter element
ER	estrogen receptor
EST	expressed sequence tag
ETS	erythroblastosis virus E26 oncogene homolog 1
GFP	green fluorescent protein
GO	gene ontology
GR	glucocorticoid receptor
GRB	genomic regulatory block
GTF	general transcription factor
HMG1	high mobility group 1 protein
hpf	hours post- fertilisation
Inr	initiator
ISH	in situ hybridisation
LacZ	beta-galactosidase
LCR	locus control region
MCS	multispecies constrained sequence
MHB	midbrain-hindbrain boundary
MTE	motif ten element
NF-κB	nuclear factor kappa B

PIC	pre-initiation complex
PolII	RNA polymerase II
pTRR	putative transcriptional regulatory regions
PTS	promoter targeting sequence
RACE	rapid amplification of cDNA ends
rCNE	regionally conserved element
SCE	shuffled conserved element
SCP	super core promoter
SINE	short interspersed element
SNP	single nucleotide polymorphism
TAF	TBP associated factor
TBP	TATA binding protein
TF	transcription factor
TFBS	transcription factor binding site
TLF	TBP-like factor
TRF1	TBP-related factor 1
TSS	transcription starts site
UCE	ultra-conserved element
UCR	ultra-conserved non-coding region
UTR	untranslated region
XCPE1	X core promoter element 1
YFP	yellow fluorescent protein
YSL	yolk syntitial layer

1) Introduction

Due to the completion of several vertebrate genomes, it became clear that mammalians encode a remarkably consistent set of genes. Moreover, vertebrate embryonic development is regulated by proteins that have orthologs with more or less sequence conservation in humans, rodents, and even in fish. Understanding the mechanisms of gene regulation during development and how gene expression regulation contributes to morphological differences among organisms expressing almost the same sets of similar proteins is the new challenge of the post-genomic era.

1.1 Gene expression regulation in eukaryotes

From the several thousands of genes of a eukaryotic cell, only a small proportion are expressed at a given time point. The proportion and composition of transcribed genes vary in different cell- or life cycle stages, in different sexes, among cell types, and in response to changes in the physiological and environmental conditions (White et al. 1999; Arbeitman et al. 2002). During metazoan embryonic development terminally differentiated cells of the adult organism are specified from the pluripotent zygote through different successive stages by sequential coordinated expression of genes. While this developmental program can be modified by epigenetic and environmental factors, in principle it is driven by genetic regulatory networks set up at the beginning of embryogenesis. These networks receive inputs from intercellular signals and the output instructions regulate expression of specific genes (Halfon et al. 2002). Eukaryotes utilize different mechanisms to regulate gene expression, including transcriptional (chromatin condensation and modification, DNA methylation, transcription initiation), post-transcriptional (silencing by RNA interference or microRNAs, alternative splicing, mRNA stability), translational and several forms of post-translational controls (covalent post-translational modifications, intracellular trafficking and protein degradation) (Alberts 2002; Levine et al. 2003). For virtually every eukaryotic gene for which relevant information exists, transcriptional initiation appears to be one of the most important determinants of the overall gene expression profile.

1.2 Cis-regulatory elements

Every gene is surrounded by sequences in cis that regulate the timing, spacing and the level of its expression under given environmental conditions. Cis-regulatory elements are stretches of DNA located in and around genes, affecting the transcript synthesis or stability in an allele-specific manner (Figure 1). Two major types can be distinguished by their position: promoters and distal regulatory elements. These regulatory DNA sequences contain binding sites for trans-regulating factors that activate, enhance, repress or keep transcription silenced.

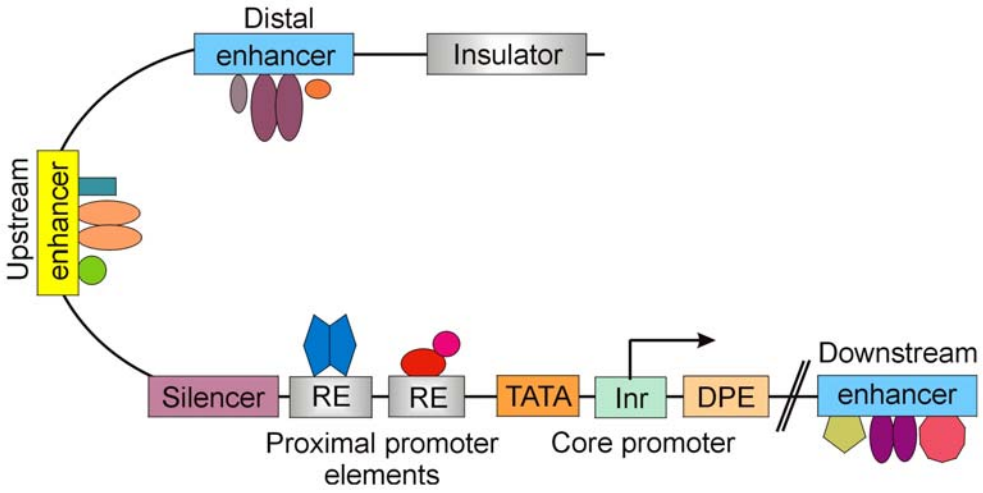


Figure 1: A scheme of eukaryotic cis-regulatory elements

A typical metazoan cis-regulatory module consists of multiple enhancers in combination with silencer(s) and insulators. INR and DPE represent initiator and downstream promoter elements. Redrawn from (Levine et al. 2003)

1.2.1 Promoters

Promoters are cis-regulatory elements where the RNA polymerase II holoenzyme assembles. A typical eukaryotic promoter, spanning a few hundred base pairs around the transcription start site (TSS), consists of a core promoter and a proximal promoter region.

Core promoters

The core promoter is defined as the minimal DNA region required to direct low levels of accurate RNA PolII transcription initiation in the absence of activators in vitro (Gross et al. 2006). Core promoters typically encompass the transcription start site and extend either upstream or downstream for an additional 35-40 nucleotides

(Butler et al. 2002). Core promoters consist of functional motifs, termed *core promoter elements*, examples include:

- **TATA-box** (Breathnach et al. 1981), usually located about 29-31 base pairs upstream (5`) of the transcription start site (Ponjavic et al. 2006);
- initiator (**Inr**), a conserved pyrimidine-rich sequence encompassing the TSS, functions to direct accurate transcription initiation either by itself or together with TATA-box or DPE elements (Smale et al. 1989);
- the downstream promoter element (**DPE**), which is located at +18 to +32 bp upstream of the start site in vertebrates (Burke et al. 1996);
- motif ten element (**MTE**), found between +18 and +29 bp position upstream of the TSS, normally functions in conjunction with the Inr, but it can substitute for the loss of the TATA-box or DPE, or work synergistically with them in an Inr-dependent manner to strengthen the promoter activity (Lim et al. 2004);
- the downstream core element **(DCE)** contains three discontinuous sub-elements, spanning from position +6 to +34 (Lee et al. 2005);
- the upstream TFIIB recognition elements (**BRE^{u})**, (Lagrange et al. 1998);
- the downstream TFIIB recognition elements (**BRE^{d}),** (Deng et al. 2005);
- X core promoter element 1 (**XCPE1**), (Tokusumi et al. 2007);
- **CpG island**, a genomic sequence overrepresented by unmethylated CG dinucleotides (Bourbon et al. 1988);

When the first protein-coding genes were isolated, virtually every gene contained a TATA-box (Breathnach et al. 1981), and further studies showed, that mutations of this element reduced transcription initiation and prevented the proper positioning of the TSS (Grosschedl et al. 1981; Takihara et al. 1986; Peltoketo et al. 1994). Based upon these observations, it was expected that a similar core promoter structure would be found in every PolII-transcribed cellular gene. But later bioinformatic analysis of promoter regions of the Drosophilal and yeast genomes revealed that only few percentages of genes contain TATA-boxes. Several studies have been performed on human promoters to determine the percentage of the TATA-containing promoters, leading to contradictory results (Trinklein et al. 2003; FitzGerald et al. 2004; Gershenzon et al. 2005; Kimura et al. 2006). The different results could arise from the usage of different databases and the experimental TSS mapping techniques. For

example Gershenzon et al. used the EPD database, which was relatively small and appeared “enriched” in TATA-containing core promoters. Indeed, analyses of larger databases, including the database of transcriptional start sites, the dbTSS (Suzuki et al. 2001), obtained by aligning the 5’end of full-length cDNAs to the human genome sequence, revealed a more restricted number of TATA-containing genes. Based on the latest study that performed genome-scale computational analyses of human core promoters present in the UCSC GoldenPath (15,685 genes) and dbTSS (10,271 genes) databases revealed that 24% of the human genes contain TATA-like elements, and only 10% of these TATA-containing promoters (2,4% of the total genes) contain the canonical TATA-box (Yang et al. 2007a).

Proximal promoters

The proximal promoter is defined as a region up to few hundred base pairs upstream from the core promoter, and typically consists of multiple transcription factor binding sites, like Sp1(Kingsley et al. 1992), CAAAT-binding transcriptional factor (CTF) (Santoro et al. 1988), and CAAAT-box binding factor (CBF) (Sakata-Takatani et al. 2004). The regulatory sequences of different inducible genes, like the metal- (Stuart et al. 1985), xenobiotics- (Fujisawa-Sehara et al. 1987), hormone- (Beato 1987) responsive and heat shock elements (Wu 1984), are usually located in the proximal promoter region.

Promoter diversity

Promoters show much higher degree of complexity as thought before, and there is a growing list of evidence of the differential usage of distinct promoters. The first level of diversity arises from the **core promoter element composition**. Different core promoter elements were shown to correlate with gene function – promoters with TATA-box were associated with highly regulated genes, while TATA-less promoters tend to be associated with housekeeping genes in yeast (Basehoar et al. 2004).

CAGE (cap analysis of gene expression), a method used to identify TSSs and to measure their expression levels, was applied in the FANTOM3 (functional annotation of mouse 3) project to sequence more than 7 million mouse and human sequences from more than 20 tissues. Using these FANTOM3 results Carninci et al. found that transcription initiation occurred at multiple nucleotide positions. They could classify four distinct categories: core promoters showing the **TSS distribution** of a.) a single

dominant peak, b.) a general broad distribution, c.) a broad distribution with a dominant peak, and d.) a bi-or multimodal distribution (Figure 2).

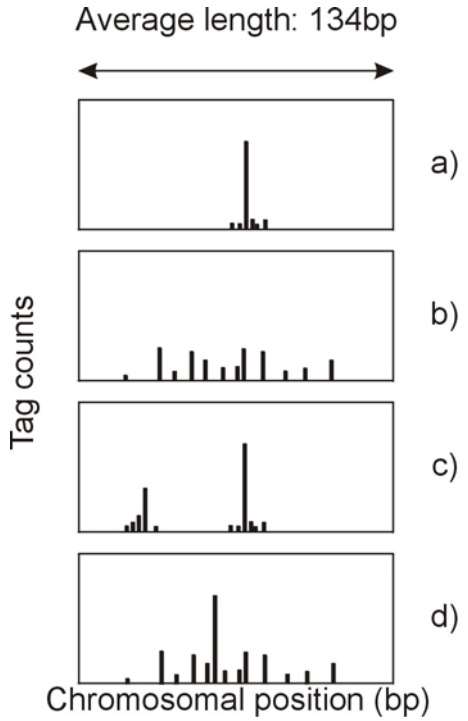


Figure 2: The four promoter categories based on their TSS distribution
a.) single peak, b.) broad, c.) bi-or multimodal, d.) broad with a dominant peak. Redrawn after (Kawaji et al. 2006)

These "promoter-shapes" were shown to be generally very similar between human and mouse orthologous promoter regions. TATA-boxes were strongly overrepresented in promoters showing sharp TSSs, while broad TSS regions were strongly associated with CpG islands (Carninci et al. 2006). Kawaji et al. could demonstrate, that there were distinct, tissue-specific modes of start site selection within core promoters for at least half of the tag clusters they investigated. Some of these tissue-specific TSSs were regulated via DNA methylation and/or subsequent chromatin remodelling (Kawaji et al. 2006).

Additional diversity in gene regulation is achieved by the use of **multiple (alternative) promoters** for a single gene. In alternative promoters core promoters are separated by clear genomic space, while broad or multimodal TSS distribution of a promoter represents an array of closely located initiation sites (Kawaji et al. 2006). Recent large-scale studies that identified promoters by ChIP-on-chip analysis (Kim et al. 2005), or analysing full-length cDNAs (Zavolan et al. 2002; Landry et al. 2003;

Trinklein et al. 2003; Sharov et al. 2005; Carninci et al. 2006) suggested that 14 - 58% of the human genes were subject to regulation by alternative promoters. Seventeen percent of the alternative promoter-containing loci showed tissue-specific use of these promoters (Kimura et al. 2006). The alternative promoter-containing regulatory regions were shown to be enriched in genes coding signal transduction-related proteins. In those genes, which had multiple alternative promoters, the frequency of the CpG island core promoter element was lower compared to those ones, which had only one promoter (Baek et al. 2007). Some of the mammalian genes with alternative promoters produce distinct mRNA isoforms with a heterogeneous 5` UTR, but coding identical proteins. The 5` UTR can affect the mRNA stability and the translational efficiency. In other cases distinct protein isoforms (with potentially different function) are produced from the alternative promoters (Davuluri et al. 2008).

Gene pairs that are arranged in a head-to-head orientation on opposite strands with less than 1000 bp separating their TSSs are termed **bidirectional**. In some cases it has been shown that a bidirectional promoter regulates the transcription of a gene pair whose levels need to be co-ordinately expressed, e.g. bidirectional promoters provide the stoichiometric relationship of histone genes, others regulate the co-expression of genes that function in the same biological pathway, or provide coordinated responses to signals, like heat shock. Genome-wide analysis of gene organization in the human genome identified a large class of bidirectional genes representing more than 10% of all human genes (Trinklein et al. 2004). The shared cis-regulatory elements located in the bidirectional promoters were necessary for full promoter activity in both directions. Although neighbouring genes had a correlation for coordinated regulation higher than random, the correlation for the bidirectional gene pairs was even higher. In functional tests, half of all tested human promoters did not exhibit strong directionality in transcript initiation, and the majority (90%) of the tested bidirectional promoters showed activity in both directions. Some gene categories were overrepresented in the bidirectional gene pairs, like DNA-repair, chaperone, mitochondrial and a special class of RNA-helicase genes. Sequence analysis of the these promoters revealed enrichment of CpG island core promoter element in this group (Trinklein et al. 2004).

1.2.2 Distal cis-regulatory elements

Enhancers were originally defined as DNA sequences capable of elevating the transcription of a gene containing only a promoter (Banerji et al. 1981; Atchison 1988). They typically regulate transcription in a spatial- or temporal-specific manner, and this function is independent from the distance and the orientation relative to the promoter (Atchison 1988). Enhancers are modular: different enhancers can work independently of one another to direct composite patterns of gene expression when linked within a common cis-regulatory region. Enhancers function in an autonomous fashion, sequence-specific activators or repressors bound to one element do not interfere with the activity of the others (Levine et al. 2003). Enhancers not only regulate gene expression in distinct tissues or cell types, but provide precise timing as well (Zakany et al. 1997). Enhancers consist of groups of clustered TFBSs. The identity, precise order and distance of these binding sites from one another within an enhancer cluster are often highly conserved between species, suggesting a critical role for protein-protein interactions between bound transcription factors in the proper function of the enhancer. This conveys that the distance and orientation independence is only valid for the cluster as a whole. Nice example for this is the *even-skipped stripe 2* element. The *even-skipped stripe 2* expression is conserved in Drosophilal species, but sequence of the enhancer has been diverged. The chimeric enhancer generated by gluing together the 5' and the 3' halves of the original enhancer elements from two species no longer function as an enhancer (Ludwig et al. 2000). The complex structure and the high degree of evolutionary conservation hints that enhancers have largely evolved in parallel with the coding sequences they control (Mackenzie et al. 2004). Tissue-specific enhancers can work over distances of 100kb or even more (Lettice et al. 2003; Vavouri et al. 2006). This type of long-range regulation is not observed in yeast and might be a common feature of genes that play role in morphogenesis (Levine et al. 2003).

There are two mechanisms proposed how enhancers affect gene expression. The "stochastic" model suggests that genes have two transcriptional states, and enhancers shift the balance from "off" to "on" state (Sutherland et al. 1997; Blackwood et al. 1998). The other, "rheorastic" model says that instead of the on/off switch, enhancers regulate the expression in a continuous spectrum, depending on the amount and the nature of bound factors (Rossi et al. 2000).

Silencers are cis-regulatory sequences with similar properties as enhancers but with a negative effect on transcription. They are originally defined as sequence elements capable of repressing promoter activity in an orientation- and position independent fashion (Brand et al. 1985). The negative regulatory element of the *human thyrotropin-β* (*hTSHβ*) gene (Kim et al. 1996), the NRE within the chicken *ovalbumin* 5' promoter (Haecker et al. 1995) or the NRE from the *platelet-derived growth factor A -chain* promoter (Liu et al. 1996) are examples for classical silencers. A significant number of negative regulators of transcription however are position-dependent. These passive or position dependent silencers physically inhibit the interaction of transcription factors with their specific binding sites, or interfere with signals which control splicing sites, 5' polyadenylation signals, translational start sites or by affecting transcriptional elongation (Ogbourne et al. 1998).

Insulators are DNA sequences that usually contain clustered binding sites for large zink finger proteins, such as Su(Hw) and CTCF. They selectively block the interaction of a distal enhancer with its target promoter when positioned between the two (enhancer-blocking insulators), or block the spreading of the heterochromatin (barrier insulators) (Gaszner et al. 2006). Insulators function in a position-dependent, but orientation-independent manner. They were first identified at gene boundaries, but have been also found within complex genetic loci, like the *igf-2* locus in mice (Levine et al. 2003). Although different DNA binding sequences and their associated proteins are involved in enhancer blocking in vertebrates and invertebrates, it seems that similar mechanisms have been developed. Enhancer-blocking elements can interact with each other or tether the DNA to structural elements within the nucleus to establish chromatin loops. These loops can block the direct interaction of promoters and enhancers (a mechanism compatible with the looping model of enhancer action) or block the signal travelling from the enhancer to the promoter (a mechanism compatible with the tracking model) (Gaszner et al. 2006).

Locus control regions (LCRs) are groups of regulatory elements (enhancers, silencers, insulators and matrix or chromosome scaffold attachment regions) involved in regulating an entire locus or gene cluster (Li et al. 2002). Their collective activity defines the LCR and confers proper special and temporal gene expression. Based on the regulatory element composition, LCRs not only positively or negatively regulate the transcription, but also possess all the properties necessary for opening a

chromosome domain and preventing heterochromatin formation at ectopic sites. The first identified and the best-studied one is the mammalian *β-globin* LCR (Grosveld et al. 1987), but LCRs have been found in several other mammalian loci as well (Aronow et al. 1992; Neznanov et al. 1993; Diaz et al. 1994; Dang et al. 1995; Jones et al. 1995; Kamat et al. 1999).

1.3 Transcription factors

Transcription factors are sequence-specific DNA-binding proteins involved in the regulation of transcription initiation or subsequent steps, like elongation, re-initiation (Lee et al. 2000) or in the activation of the RNA PolII complex already assembled on promoters (Kininis et al. 2007). Many of these factors belong to multiprotein families, like the nuclear receptors (Aranda et al. 2001; Kininis et al. 2008), AP-1(Curran et al. 1988), CTF/NF-I (Santoro et al. 1988), NF-κB (Baldwin 1996), p53 (Yang et al. 2002), and Sp families (Kingsley et al. 1992). Transcription factors are modular (Brent et al. 1985): a typical TF has a DNA-binding domain linked to one or more activation or repression modules, potentially contains a multimerization and a regulatory module. There are many distinct DNA binding domains, like the homeodomain, zink finger, leucine zipper, helix-loop-helix, forkhead, ETS, POU or HMG1 domains and others (Pabo et al. 1992). Each TF has a variety of sequences they bind to, summarized as a consensus sequence or a position-specific score matrix (Stormo 2000). Binding of a given TF to its binding site depends on several factors:

- the sequence of the binding site determines the strength of the interaction, the structure and the methylation state of the DNA,
- the methylation, acethylation and phosphorylation state of the neighbouring histones and the presence of other proteins (other TFs or remodelling factors) influence the availability of the site,
- and other proteins such as co-activators or co-repressors can influence or inhibit the DNA-protein interaction.

A TF may bind to a site on the DNA without having effect on the transcription (non-functional binding) (Tabach et al. 2007). As the sequence-specific protein-DNA interactions rarely extend more than 5 base pairs (in the case of zinc finger TFs it is only 3 bp), the extent of this physical interaction is not sufficient to provide much

sequence specificity, other structural features have to increase the number of nucleotides required for efficient binding. Some TFs contain multiple DNA binding domains (like the members of the Pax family have a paired-box and a homeodomain), additional structural features can bind nearby nucleotides through minor groove contacts (like in many homeodomain and GATA TFs), and homo- or heterodimerization of the TFs can be required prior to DNA binding (e.g. for most nuclear receptors) (Wray GA et al. 2003).

Transcription cofactors or coactivators by definition lack DNA-binding domains, but function the same ways as transcription factors. They typically contain domains that mediate a specific protein-protein association with a TF and directly or indirectly with effector complexes (either the transcription machinery or chromatin remodelling complexes) (Meier 1996).

1.3.1 Core promoter binding factors

The general RNA PolII transcription machinery has been biochemically defined as a set of factors essential for accurate transcription initiation at TATA-containing promoters in vitro, and consists of the general transcription factors (GTFs) TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH, and the RNA polymerase II. Transcription initiation requires ordered assembly of RNA PolII and GTFs into a pre-initiation complex (PIC) at the core promoter (Gross et al. 2006). These factors are considered general, as they have been proposed to be present in all multiprotein complexes formed on promoters, although recent results showed that different PICs can contain different GTFs (Muller et al. 2004). The assembly of the PIC on the core promoter is sufficient to drive basal levels of transcription; this basal activity is greatly stimulated by transcription factors, also called as activators (Ptashne et al. 1997).

TFIID is a multiprotein complex playing important role in promoter recognition, consists of TBP (TATA-binding protein) (Horikoshi et al. 1989), which mediates the interaction with the promoter DNA, and TBP-associated factors (TAFs) (Tora 2002) that stabilize the TBP-promoter interaction. TBP is the predominant TATA-box binding protein, but there are several TBP-related factors with partial homology to TBP. TRF1, only present is Drosophilal, was shown to be able to bind to non-canonical TATA-box motifs and to TC box sequences (Crowley et al. 1993). TRF2/TLF, first discovered in Drosophila, but later found in vertebrates as well, does not appear to bind TATA-box, but has been shown to be required for expression of a

specific set of genes, or in specific developmental stages (Dantonel et al. 2000; Veenstra et al. 2000; Muller et al. 2001). TBP2, isolated from vertebrates, binds TATA-box sequences, interacts with TFIIA and B, and is expressed in the gonads and during embryonic development (Bartfai et al. 2004). TFIID binds cooperatively to other core promoter sequences as well, for example it interacts with the Initiator and the DPE elements (Kaufmann et al. 1994; Burke et al. 1996), and this interaction is mediated through TAF1 and 2 in the case of the Inr (Chalkley et al. 1999), while TAF6 and TAF9 interact with the DPE sequence (Burke et al. 1997). DCE is also recognized by TFIID via the TAF1 subunit (Lee et al. 2005). **TFII-I** and **YY1** interact with the Inr (Roy et al. 1991; Weis et al. 1997). **SP1** and related transcription factors bind to GC boxes, sequences found in CpG islands (Butler et al. 2002). **TFIIB** interacts with the upstream (BRE^u) (Lagrange et al. 1998) and the downstream TFIIB recognition elements (BRE^d) (Deng et al. 2005) via different consensus sequences.

The major step for the pre-initiation complex formation in TATA-box containing promoters is the binding of the TBP to the TATA-box sequences present at ~30 base pairs upstream from the TSS (Hahn et al. 1989). The binding of TBP to various TATA sequences induces a dramatic DNA bend (Patikoglou et al. 1999), and is stabilized by cooperative interactions with TFIIB, TFIIA and with TAFs, which interact with the INR and other downstream core promoter elements (Hahn 2004). Transcription initiation from promoters lacking TATA-box elements are mediated by alternative PICs, like the TBP-free TAFII-containing complex (Brand et al. 1999; Hardy et al. 2002).

1.3.2 Enhancer/silencer-binding factors

Studies using non-purified chromatin templates have shown that transcription initiation is massively influenced by distal cis-regulatory sequences. Transcription factor binding of an enhancer results in changes in the nucleosome-structure and in recruitment of histone-modifying enzymes - this step is important to generate protein-accessible chromatin around the promoter region. Co-factor-containing mediator complexes bound to the transcription factors present on the enhancer then mediate protein-protein interactions with the basal transcription machinery that is targeted to the core promoter. The formation of this multiprotein complex (bringing together the promoter and enhancer elements) results in the transcription initiation (Cosma 2002) (Figure 3).

It has been shown for the steroid receptors that they could bind to (or near to) proximal promoter regions as well as sequences located even at several hundred kilobase distance, and different regulatory role has been shown for the distal and the proximal cis-regulatory elements (Kininis et al. 2008). Genome-wide studies showed that not only transcription factors, but also GTFs or the RNA polymerase II itself were bound to the enhancer regions (Shang et al. 2002; Spicuglia et al. 2002; Carroll et al. 2006; Kininis et al. 2007; Kwon et al. 2007). These results suggest transcription of enhancer elements and are consistent with the findings of global transcriptome analysis, which provided evidence that a large proportion of the genome is transcribed (Katayama et al. 2005). One possible answer why PolII or GTFs are present at distal cis-regulatory regions could be that they regulate correct timing of gene activation in different cell types during development (Szutorisz et al. 2005).

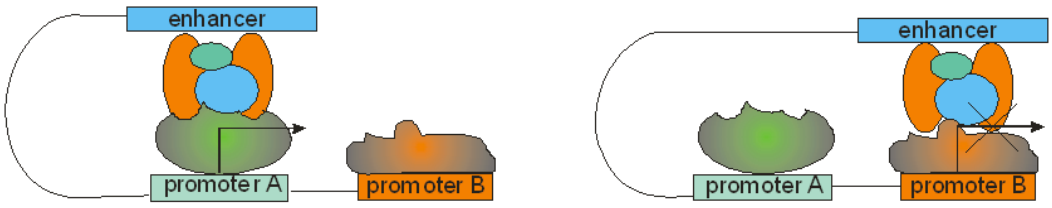


Figure 3: Transcription regulation mediated by differential transcription factor-containing multiprotein complexes formed on cis-regulatory elements

Silencers are binding sites for negative transcription regulators, called repressors. Repressor function can require the recruitment of co-repressors, or TFs can switch to repressors by differential co-factor-binding.

1.4 Genomic organisation of cis-regulatory elements

Scattering of cis-regulatory elements is a general feature of many genes particularly of developmentally regulated genes (Plessy et al. 2005; Kikuta et al. 2007). Because of their unpredictable distance from the target promoter and the potential interdigitate position, the annotation of cis-regulatory elements to their target promoter is difficult.

Introns were thought to be remnants of early assembly of genes, subjects to minimal pressure for their removal (Gilbert et al. 1986), or selfish DNA with no function, the result of the increased capacity of multicellular organisms to accumulate

cellular debris from transposons and other sources (Cavalier-Smith 1985). But there is a growing list of evidence of the functionality of introns: introns were shown to improve transcriptional and translational yield (Juneau et al. 2006), they contain conserved sequences with yet uncovered function, they code all the small nucleolar RNAs (Liu et al. 1990a) and a large fraction of microRNAs (Eddy 1999), and several enhancer elements are located in introns (Brooks et al. 1994; Howell et al. 1997; Muller et al. 1999; Sivak et al. 1999; Hural et al. 2000; Flodby et al. 2007; Khandekar et al. 2007; Camp et al. 2008). The distribution of intronic sequences is probably non-random based upon the results of Taft et al. (2007). They have found correlation between the total intronic sequences within annotated protein-coding genes and their functions. Large introns were overrepresented in genes expressed in the nervous system, uterus and in genes under-expressed in immunologic, embryonic stem and cancer cells; in genes that require precise transcriptional regulation. Small introns were enriched in genes highly expressed in heart, bone narrow, lung and pancreas (Taft et al. 2007). Distal cis-regulatory elements can be embedded in an intron of another gene, with a potentially different function and/or expression pattern. The gene, the enhancer is functionally linked to, is the target gene, and the gene, in which the interdigitate regulatory element is located, is the bystander gene (Kleinjan et al. 2005).

Approximately 25% of the human genome consists of **gene deserts** – long genomic regions containing no protein-coding genes and with no obvious biological function (Venter et al. 2001). Some of these gene deserts were shown to contain conserved elements with enhancer function (Nobrega et al. 2003; Kimura-Yoshida et al. 2004; Uchikawa et al. 2004), while deletion of other gene deserts resulted in no severe effects on survival of mouse embryos (Russell et al. 1982; Rinchik et al. 1990; Nobrega et al. 2004). Based on comparisons of human and chicken genomes, and analyzing the genomic structure, conservation patterns and evolutionary relationships of the gene deserts present in these species, Ovcharenko et al. (2005) could classify them into two functionally different groups: stable and variable gene desert. Stable gene deserts are more conserved between chicken and human, and between fugu and human, than variable ones. Stable gene deserts are flanked with genes functioning as transcription factors, developmental regulator and DNA binding proteins. Stable gene deserts are functionally linked to at least one of the flanking genes, forming large syntenic regions, and the already described conserved enhancers are located in the

stable group. These properties of the two groups hint that stable gene deserts are the ones that contain functional elements, while variable gene deserts are probably more "disposable" (Ovcharenko et al. 2005).

1.5 Evolutionary aspects of cis-regulation

The morphological and behavioural complexity of higher organisms is not reflected in expanded gene numbers (Hahn et al. 2002), so other mechanisms should be responsible for the increase of complexity. These mechanisms involve the redeployment of developmental genes in novel tissues and pathways, multifaceted use of the genes (alternative splicing of transcripts and the usage of different alternative promoters) and alterations in cis-regulation. There are emerging data from the field of evolutional biology showing the importance of the evolution of gene regulatory networks in divergent developmental pathways.

There are many factors contributing to the importance of cis-regulatory DNA in evolution. First, individual cis-regulatory elements can act and evolve independently of others. A good example is the typical organisation of the cis-regulatory regions of developmental genes, composed of many independent elements. The products of most of the genes involved in morphology patterning have pleiotropic function, like influencing multiple phenotypic traits or regulating the expression of many different genes. Mutations affecting protein function may cause disturbance in much more developmental and physiological processes, therefore less tolerable in the evolution. Second, there is a higher degree of freedom in cis-regulatory sequences, which allows greater varieties of mutations. Regulatory elements do not need to maintain any reading frame, they can function at widely varying distances and in either orientation to the transcription units they control. This evolvability of regulatory DNA sequence means that it is a rich source of genetic and, potentially, phenotypic variation. Finally, most elements are controlled by TFs whose DNA binding specificity is sufficiently relaxed that the affinity and number of sites for each factor can evolve at a significant rate, even in functionally conserved elements (Carroll 2000).

When human and chimpanzee homologous proteins were sequenced, and found to be nearly identical, the role of changes in cis-regulatory elements in the variation of gene expression has been hypothesised (King et al. 1975). Since then, mutations of several regulatory elements have been shown to modify specific aspects of patterns and/or levels of gene expression during development, leading to changes in

organogenesis, resulting in morphological and physiological modifications. Several reports showed cases for altered cis-regulatory elements causing different phenotypic effects in metazoans (Stern 1998; Sucena et al. 2000; Wittkopp et al. 2003; Shapiro et al. 2004; Wang et al. 2004; Gompel et al. 2005; Prud'homme et al. 2006), but there are limited data from higher vertebrates. Cretekos et al. (2008) investigated the limb-specific enhancer of the *prx1* gene in different mammalian species. Nevertheless the shocking morphological differences of their forelimbs, the initial limb bud formation in the mouse (*Mus musculus*) and in the short-tailed fruit bat (*Carollia perspicillata*) is identical, the differences only appear in later stages of limb formation. Replacement of the limb-enhancer containing genomic region upstream from the mouse *prx1* gene to the orthologous bat sequence resulted in higher levels of *prx1* transcript and elongated forelimbs in transgenic mouse embryos. Interestingly, deletion of the mouse enhancer did not cause any detectable phenotype, suggesting the presence of additional regulatory elements with redundant function (Cretekos et al. 2008). A conserved noncoding sequence (called as HACNS1) that evolved extremely rapidly in humans worked as an enhancer in the forelimb and some other parts of the body, notably the pharyngeal arches, eye and ear when tested in transgenic mice, while the orthologous elements from chimpanzee and rhesus macaque did not show any enhancer activity. In vivo analyses with synthetic enhancers, in which human-specific substitutions were introduced into the chimpanzee enhancer sequence indicated that 13 substitutions in the otherwise highly constrained element were sufficient to confer human specific limb expression domain (Prabhakar et al. 2008).

New cis-regulatory elements can arise by several mechanisms, including random sequence mutation, genomic insertions (these can bring functionally active sequences with regulatory capacity novel to the host gene), gene duplication followed by divergence in the regulatory modules. Gene duplication is often seen after aberrant recombination or replication, or chromosome and genomwide duplications (Ohno et al. 1968). Transposon-derived sequences, often referred as repetitive sequences or “junk DNA”, were shown to harbour regulatory functions as well (Peaston et al. 2004; Bejerano et al. 2006; Nishihara et al. 2006; Xie et al. 2006).

Gene duplication is thought to be one of the major sources of cis-regulatory element evolution, as it provides material for novel gene functions and expression patterns to arise from (Cooke et al. 1997; Lynch et al. 2000; Gompel et al. 2005; Jeong et al. 2006; Prud'homme et al. 2006). The most common fate of a duplicated

gene pair is the non-functionalisation of one of the genes (one copy collects deleterious mutations, and thus degenerates to a pseudogene) (Nowak et al. 1997). Advantageous mutations can also occur in one of the duplicated genes, of course less commonly, thus one copy evolves new function. The third possible mechanism is the subfunctionalisation, when both of the duplicated paralogs are retained in the genome (Prince et al. 2002). The retention of duplicated paralogs during evolution by subfunctionalisation is the basis of the duplication-degeneration-complementation (DDC) model (Force et al. 1999). This model suggests that each duplicated gene can fulfil only a subset of complementing functions of the ancestral gene Several studies implicated specific mutations in enhancers of paralogous gene copies to be the likely source of subfunctionalisation in duplicated *engrailed2* (Postlethwait et al. 2004), *hoxb2* (Scemama et al. 2002), *hoxb3a* and *hoxb4a* (Hadrys et al. 2004; Hadrys et al. 2006), *fign, pax2* and *unc4.1* (Woolfe et al. 2007a) enhancers in fish.

1.6 Medical aspects of cis-regulation

The proper execution of biological processes such as development, proliferation, apoptosis, aging and differentiation requires a precise regulation of the spatial and temporal expression of genes. Alterations in the properties of the interaction between promoters and other cis-regulatory elements (either by mutation or by physical dissociations) can cause defects in the transcriptional control.

Disease	Mutation (relative to the TSS)	Affected gene	Reference
β-thalassemia	TATA/box CACCC box, DCE	β-globin	(Antonarakis et al. 1984) (Kulozik et al. 1991) (Lewis et al. 2000)
δ-thalassemia	GATA1 (77 bp 5`)	δ-globin	(Matsuda et al. 1992)
Bernanrd-Soulier Syndrome	GATA1 (133 5`)	GpIbβ	(Ludlow et al. 1996)
Charcot-Marie-Tooth disease	(215 5`)	connexin-32	(Wang et al. 2000)
Congenital erythropoietic porphyria	GATA1 (70 5`) CP2 (90 5`)	uroporphyrinogen III synthase	(Solis et al. 2001)
Familian hypercholesterolemia	Sp1 (43 5`)	LDL receptor	(Koivisto et al. 1994)
Familial combined hyperlipidemia	Oct1 (39 5`)	lipoprotein lipase	(Yang et al. 1995)
Haemophilia	CCAAT box	factor IX	(Crossley et al. 1990)
Progressive myoclonus epilepsy	Expansion ~70bp 5`	cystatin B	(Lalioti et al. 1997)
Pyruvate kinase deficient anaemia	GATA1 (72 5`)	PKLR	(Manco et al. 2000)
Treacher Collins syndrome	YY1 (346 5`)	TCOF1	(Masotti et al. 2005)

Table 1: Examples of diseases caused by mutations in the promoter region

Germline chromosomal rearrangements were identified in some human diseases in which the phenotype-associated breakpoints or mutations were found outside of the coding sequences. In these syndromes the mutations were shown to present in the core or the proximal promoter regions (Table 1), or single enhancer, silencer or insulator elements or whole locus control regions were affected (Table 2) (Kleinjan et al. 2005; Maston et al. 2006). One example for these mutations is affecting the limb-specific *ZRS* enhancer of the *sonic hedgehog (shh)* gene. This element is located in one megabase distance from the *shh* locus in human, in an intron of the *limb deformity region 1 (lmbr1)* gene. Genetic lesions affecting this element cause preaxial polydactily in human patients and in mutant mouse strains (Lettice et al. 2002), while complete elimination of this regulatory region causes severe limb truncations in mice (Sagai et al. 2005). A single point mutation in the enhancer element can be responsible for the polydactyly (Lettice et al. 2003).

Disease	Gene	Distance of the cis-reg element	Reference
Aniridia	Pax6	125 kb	(Fantes et al. 1995; Kleinjan et al. 2001)
Saethre-Chotzen Syndrome	Twist	260 kb	(Cai et al. 2003)
X-linked deafness	POU3F4	900 kb	(de Kok et al. 1996)
Reiger syndrome type I	Pitx2	90 kb	(Flomen et al. 1998)
Greig cephalopolysyndactyly syndrome	Gli3	10 kb	(Wild et al. 1997)
Anomalies in cataract and ocular development	MAF	1 Mb	(Paige et al. 2000)
Iridogoniodysgenesis type I	FOXC1	1,2 Mb	(Davies et al. 1999)
Lymphedema distichiasis	FOXC2	120 kb	(Fang et al. 2000)
Blephalrophimosis-posis-epicantus inversus s.	FOXL2	170 kb	(De Baere et al. 2001)
Campomelic Dysplasia	Sox9	850 kb	(Pfeifer et al. 1999)
Holoprosencephaly	Six3	200 kb	(Wallis et al. 1999)
Holoprosencephaly	Shh	265 kb	(Belloni et al. 1996)
Preaxial polydactily	Shh	1 Mb	(Lettice et al. 2002)
Split-hand/split-foot malformation type I	dlx5/6	450 kb	(Scherer et al. 1994)
α-thalassemia	HBA2	18 kb	(Tufarelli et al. 2003)
Limb deformality	gremlin		(Zuniga et al. 2004)

Table 2: Examples of diseases caused by mutations in distal cis-regulatory regions

Improper regulatory function due to mutations in general transcription factors and chromatin remodelling proteins can lead to severe diseases as well. Mutations in TFIIH have been shown to cause xenoderma pigmentosum (Lehmann 2001), while BRG1 and BRM, the mammalian homologs of the SWI/SNF factors, are mutated in

several cancer cell lines, and the mutant proteins participate in the altered regulation of cell proliferation and metastasis (Banine et al. 2005). The best-characterized causes of malignant transformation are the chromosomal rearrangements leading to chimeric DNA sequences containing genes with improper regulatory regions. This type of rearrangement between the regulatory regions of the immunoglobulin or T-cell receptor genes and the cMYC oncogene causes the inadequate activation of the cMYC protein, leading to Burkitt`s lymphoma or acute T-cell leukaemia (Popescu et al. 2002). When *c-myc* is translocated to an immunoglobulin locus, an extra step of alteration occurs in the expression regulation, a shift in the alternative promoter usage of the *c-myc* gene (Marcu et al. 1992). Aberrant activation or repression of genes from alternative promoters is often associated with cancer initiation and progression. CYP19A1 is overexpressed in several estrogen-dependent breast cancers, and this overexpression in often caused by aberrant activation of one of the eight promoters distributed over a 93 kb region. (Bulun et al. 2007)

Disruption of the expression regulation of developmentally regulated genes is implicated in neuropsychiatric disorders, including Parkinson's disease, schizophrenia, bipolar disorder and autism. Most of these genes produce distinct protein isoforms in different brain regions and developmental or differentiation stages via differential expression regulation from alternative promoters of dopamine receptors (Anney et al. 2002), serotonin receptors (Parsons et al. 2004), and brain-derived neurotrophic factor (Liu et al. 2005).

As it was shown for the ZRS enhancer, mutation of one nucleotide can disrupt enhancer activity (Lettice et al. 2003), several laboratories started studying the impact of the single-nucleotide polymorphism (SNP) of the non-coding sequences on gene expression. 30-60% of human promoters contain functional regulatory SNPs, which tend to cluster in an approximately 100 base pair-range region around the TSS, suggesting a high impact of promoter-mutations in diseases (Buckland et al. 2005; Pastinen et al. 2006).

1.7 Mechanism of interaction between cis-regulatory elements

Genes maintain their functional identity in the complex and diverse genomic organisation. Ideas about how independent expression profiles of individual genes are managed originally came from electron microscopy observations showing that lamp brush chromosomes were structurally organized in large loops of varying sizes. (Gall 1956) This observation led to the assumption that loops are **structural domains** that represent functional domains of specific gene expression. The existence of insulator and boundary elements further strengthen the structural domain view, by assuming that chromosomes are subdivided into physically distinct expression domains containing a gene or a gene cluster and all its cis-regulatory elements. This model suggests that functional independence of genes is due to their structural autonomy; they are physically separated from neighbouring domains by specific boundary or insulator sequences, which would block the spread of heterochromatin from one domain to the next and/or counteract the effects of neighbouring enhancers (Dillon et al. 2000). However the findings that independently regulated loci can partially or completely overlap and their cis-regulatory elements can be found within or beyond neighbouring unrelated genes questions the generality of the this structural domain model. Nevertheless, insulator activities appear to co-localize frequently with other transcriptional activities and vice versa. Scs`, a prototypic insulator in Drosophilal, harbours the promoter of the Aurora gene (Glover et al. 1995), the Drosophilal enhancer-blocker gypsy can act as a promoter-specific transcriptional stimulator (Wei et al. 2001), and CTCF, the only mammalian insulator protein knows so far, was originally isolated as a transcriptional enhancer and silencer (Klenova et al. 1993; Filippova et al. 1996; Bell et al. 1999). Single DNA elements can harbour multiple regulatory activities and TFs can exert different effects depending on the DNA context where their binding sites are present. Taking these results into account, instead of structural entities, genes are better characterized as "**functional expression modules**" that encompass both the transcribed regions and their cis-regulatory sequences. These modules function appropriately in different cell types within the context of the local chromatin architecture (de Laat et al. 2003).

Since the demonstration of the existence of distant enhancers, the question arise how these long-range elements interact with their cognate promoters over hundreds of kilobases of intervening DNA. There are several models to interpret the mechanisms

underlining cis-regulatory element interactions. In the **random collision model** both the enhancer and the promoter move around randomly until they encounter each other, and when the contact is established, transcription is activated (Park et al. 1982). The **tracking or scanning model** states that the initially formed enhancer-bound complex scans along the DNA in search of a promoter (Heuchel et al. 1989; Tuan et al. 1992). Combination of the tracking and looping models exist as well (**facilitated tracking**), suggesting a mechanism where the complex tracking along the DNA remain attached to the enhancer, dragging it along to create a loop (Blackwood et al. 1998).

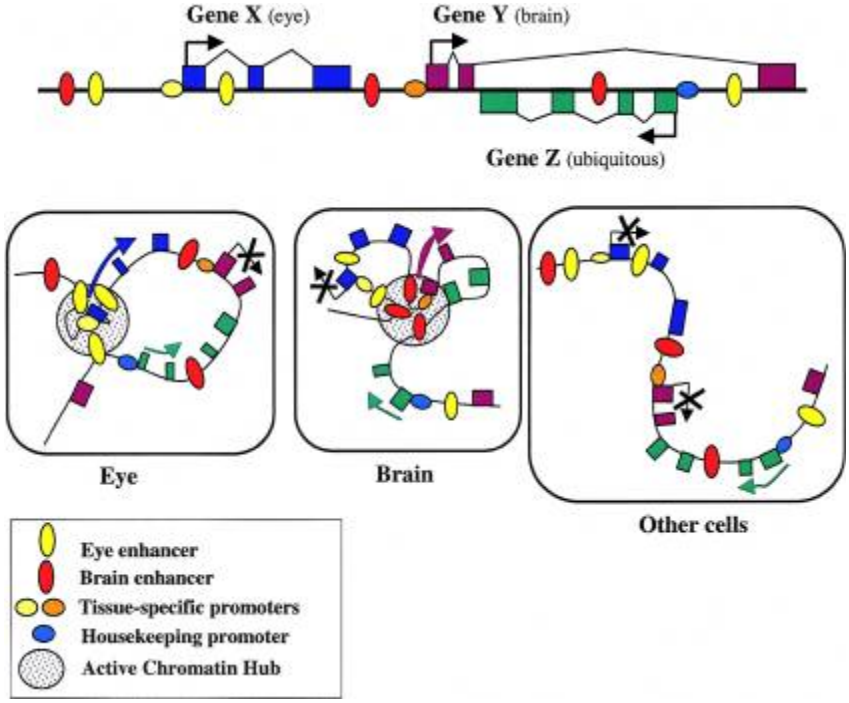


Figure 4: The looping model of cis-regulatory element interaction

The interaction of cis-regulatory elements by loop-formation, thus the expression of tissue-specific genes is cell type –specific due to the availability of activators and coactivators. From (Kleinjan et al. 2005)

Together with the tracking model, the **looping model** (Figure 4) is the most commonly encountered one. In this model, transcription factors bound at the enhancer make direct contact with the promoter and/or with factors bound to the enhancer, while the intervening DNA loops out (Wang et al. 1988). Biochemical analyses of DNA structure suggested that looping is a mechanism that can be used to increase specificity and affinity simultaneously and, at the same time, to control the intrinsic stochasticity of cellular processes (Vilar et al. 2005). Several reports provide strong

experimental support for a mechanism of long-range interaction that involves close contact between the enhancer and the promoter as in the looping model. By using the 3C (chromosome conformation capture) technique de Laat et al. (2003) described the spatial clustering of regulatory elements and active promoters as a formation of an active chromatin hub (ACH). The result of ACH formation is a high-density clustering of regulatory elements, their cognate binding factors, associated coactivators and chromatin modifiers, which sets up a suitable local environment to generate precisely the required expression level, counteracting even heterochromatic surrounding (Figure 4.). Genes are expressed when the hubs make contact with the RNA PolII molecules, which are distributed as multimolecular aggregates within the nucleus that form transcription factories (Osborne et al. 2004). In recent studies, interactions have been detected at these factories within and between chromosomes (Osborne et al. 2004; Spilianakis et al. 2005). One possible mechanism how the promoters to find the hubs or transcription factories is the transcription of the intergenic regions, which would bring together enhancers and promoters by the RNA PolII itself (West et al. 2005).

1.8 Promoter-enhancer interaction specificity

As cis-regulatory elements can be located in large distances from the promoter of the regulated gene, enhancers are potentially able to influence transcription of more than one gene, but in vivo - in their original genomic context – an enhancer generally has only one target gene.

Distance between key regulator elements and promoters is one important parameter in defining the outcome of the competition of promoters for a particular enhancer. Like in the *hoxD* cluster, where genes compete for an upstream enhancer, with proximal genes being favoured over distal ones (Kmita et al. 2002). Distance is also expected to be relevant in terms of spacing between cis-regulatory elements. Structural studies show, that the flexibility and conformation of the chromatin template will restrict the distance between two elements forming a loop (Rippe 2001). In addition to distance, promoter affinity is also important in gene competition. Promoter competition ensures the activation of a specific gene by a given enhancer, enhancer competition or enhancer interference could lead to specific ways of controlling one gene by the selected enhancer (Lin et al. 2007). Since affinity is

dependent on transcription factors bound to the cis-regulatory elements, it can be modulated in time and space (Ohtsuki et al. 1998).

Several studies have shown that the **core promoter sequence context** can significantly influence the responsiveness of a given gene to a gene-specific DNA-binding activator and repressor. The earliest studies of different TATA-box elements revealed that different TATA-box sequences respond differentially to activators. For example the human *hsp70* promoter becomes unresponsive to E1A when its natural TATA-box is substituted by the SV40 TATA element (Simon et al. 1988). Later studies investigated how the presence or absence of different core promoter elements affects activator functions. For example c-FOS preferentially activates transcription from TATA-containing core promoters (Metz et al. 1994) , while ETS family member ELF-1 exhibits a preference for Inr-containing ones (Ernst et al. 1996). Core promoter selectivity is also observed in transcription repression. For example p53 has been reported to repress transcription from promoters containing a consensus TATA motif, whereas promoters containing Inr elements instead of a TATA-box were resistant to p53-dependent repression (Mack et al. 1993). Studies in Drosophila have provided evidence that core promoter structure plays an important role in selectivity of enhancers for their target genes (Li et al. 1994; Ohtsuki et al. 1998). A later study using FLP/Cre excision and enhancer-trapping techniques could demonstrate the existence of promoter type – specific enhancers. Three out of 18 characterized trapped enhancers turned to be DPE- specific, while one was TATA-box-specific, enhancing the transcription from only one specific promoter type (Butler et al. 2001). In vertebrates a cell-type specific enhancer element of the rat carbamyl phosphate synthetase was described to be gene specific, as it requires a proximal GAG for the interaction with the promoter. The activation of the heterologous thymidine kinase promoter by the enhancer was possible when a GAG element was introduced (Goping et al. 1995).

A **promoter targeting sequence (PTS)** was described in Drosophilal in the context of the *bithorax* gene complex. This element has an anti-insulator activity; it allows an enhancer to activate its promoter despite an intervening insulator and facilitates long-distance enhancer-promoter interactions, plus selectively activates a single promoter when more than one is included in the same transgene (Zhou et al. 1999). A later study showed that this *abd-B* locus contains multiple PTSs, all of them can overcome multiple insulators and function from a number of positions relative to

the enhancer and the insulator (Chen et al. 2005). This promoter targeting sequence was found to play a role not only in promoter competition, when multiple promoters are available for a single enhancer, but also in enhancer interference (when several enhancers are competing for one promoter) as well (Lin et al. 2007). Until now, no information is available about PTSs present in other genomes than Drosophilal.

Although most enhancers directly influence the expression of just one gene, many exceptions are known. In the case of bidirectional promoters, cis-regulatory element located between the two promoters can regulate transcription of paralogous loci that lie on opposite strands of DNA (Trinklein et al. 2004). Regulatory element sharing or cross-regulation is also a known phenomenon in paralogs that are transcribed convergently or in parallel, like the *hoxB* cluster (Sharpe et al. 1998). Cross-regulation may be one reason for the long-term physical linkage of genes in the *hox* complexes of animals.

1.9 Identification of novel cis-regulatory elements

The functional and sequence code organization of the cis-regulatory elements is much less understood than that of the protein coding sequences. Automated search for regulatory sequences is thus quite difficult, as there are no sequence features that provide a consistent and general relationship to promoter, enhancer or insulator function. There are numerous experimental and computational methods to predict sequences with potential cis-regulatory activity. The success rate of predicting or detecting cis-regulatory elements depends greatly on the quality of the genome assembly, as the correct choice of the genomic region around the target gene is a crucial step for assigning functional elements into this region. The prediction or experimental identification of TSSs is crucial for the proper definition of promoters.

1.9.1 Transcription factor binding site analysis

The common feature in the cis-regulatory elements is that they contain multiple transcription factor binding sites (TFBSs) forming cis-regulatory modules (CRMs). Because of the enrichment of cis-regulatory element in TFBSs, techniques used for the identification of cis-regulatory elements are usually combined with transcription factor binding site analysis. The average TFBS spans 5-8 bp, most of them tolerate at least one, and often more, specific nucleotide substitution without losing functionality. The full range of sequences that can bind to a particular TF is often

displayed in position-specific score matrices (Stormo 2000). The consensus sequence of a particular TFBS refers to the single best variant of the binding site matrix or to a degenerate sequence that captures most of the binding sites. Given that there are many transcription factors with different binding matrices and that binding sites are short and imprecise, every kilobase of genomic DNA contains dozens of potential TFBSs. Based upon biochemical tests, many of these consensus matches do not bind protein in vivo and have no influence on transcription (Wasserman et al. 2004; Vavouri et al. 2005). Less false positive outcome is gained with those methods, which use extra criteria, such as conservation of sites across species, clustering of binding sites in regulatory regions, or association with existing information about the expression pattern of the gene (Bailey et al. 1995).

1.9.2 Promoter-predicting tools

As discussed in a previous chapter, promoters can contain a large variety of core promoter elements in different combinations, so simply searching for the co-occurrence of known core promoter motifs has only limited success (Fickett et al. 1997). The more powerful promoter prediction programs are based on the analysis of training data set of already described promoters and scan the genomic sequences for a common sequence contexts (Knudsen 1999; Scherf et al. 2000; Davuluri et al. 2001). The newest algorithms that predict promoters and TSSs use data sets containing information about promoters, exons and introns as well (Knudsen 1999; Davuluri et al. 2001; Bajic et al. 2002; Bajic et al. 2004; Lu et al. 2008). Still, the prediction potential of these programs is limited due the training sets they use predetermine the search.

1.9.3 Experimental identification of TSSs

The transcriptional start site can be identified as the first nucleotide copied at the 5` end of the nascent mRNA by using different methods like nuclease protection assays, primer extension or 5` RACE. Known TSSs are used to define core promoters and aid in searching for further cis-regulatory elements (Sandelin et al. 2007). The construction of full-length cDNA libraries containing the cap associated 5` ends allowed the determination of the exact position of the TSSs and the adjacent putative promoters from the human genomic sequences in a high-throughput manner (Suzuki et al. 2001). Information about eukaryotic promoters of which the TSS is

experimentally defined are gathered in the EDP (Schmid et al. 2006), dbTSS (Suzuki et al. 2002) or PromSer (Halees et al. 2003) databases. Results from these large-scale studies have revealed a surprisingly large number of novel intergenic transcripts, containing transcribed distal enhancers or non-coding RNA products that function in imprinting or as transcriptional co-activators (Sandelin et al. 2007). Recent results from the detailed analysis of 1% of the human genome by the ENCODE project consortium have found that over 90% of the regions tested were transcribed into primary transcripts (King et al. 2007), suggesting that the genome is transcriptionally more active than thought before, or our categories and definitions of functional elements are out of date (Elgar et al. 2008).

1.9.4 Experimental methods to identify functional elements in the genome

Regions in the genomic DNA in which the chromatin state is perturbed can be detected with **DNaseI hypersensitive site mapping**. This method was developed for high-throughput genome-wide detection of transcriptionally active regions (Crawford et al. 2004).

A technique called chromatin immuneprecipitation-coupled DNA microarray analysis (**ChIP-on-chip**) can be used to investigate whole genomes for sequences that are able to bind a specific transcription factor. These DNA sequences can contain enhancers, silencers or active promoters. With this technique Kim et al. (2005) could determine around 10.000 TFIID-binding DNA regions in the human genome, which were in close proximity to the 5` end of known transcripts, and enriched in core promoter elements like CpG islands, Inr and DPE, so these were considered as promoters. This list of in vivo TFIID-binding elements contained roughly 4200 new promoters for at least 2500 known genes, and 1200 putative promoters that correspond to previously un-annotated transcription units (Kim et al. 2005).

Transposon-based vectors are generally used to detect regulatory sequences by **gene trap or enhancer trap** experiments. In a promoter trap system, a reporter gene is cloned into the terminal repeats of the original transposons, which is only expressed, when the insertion occurs near to a functional promoter. In the enhancer trap system, a minimal attenuated promoter is cloned in front of the fluorescent reporter gene, which is switched on only when the construct can “sense” an enhancer. As the sites of the transposition events are easy to detect by PCR performed with transposon-specific primers, the neighbouring sequences (containing the regulatory

elements driving the expression of the reporter into distinct tissues) can be identified. So far the Sleeping Beauty (Ivics et al. 1997), the Tol2 (Kawakami et al. 1998) and the Ac/Ds (Mc Clintock 1951) transposons have been used in zebrafish and medaka (*Oryzias latipes*) for identifying cis-regulatory regions (Davidson et al. 2003; Kawakami 2004; Parinov et al. 2004; Emelyanov et al. 2006; Fisher et al. 2006b). A similar, but retrovirus-based technique was used to generate enhancer trap lines in zebrafish, using a modified murine leukaemia virus containing the 1kb *gata2* promoter followed by the *yfp* gene (Ellingsen et al. 2005).

1.9.5 Phylogenetic footprinting

Pair wise or multiple sequence comparisons between evolutionary diverged species can highlight functional **conserved regions** (orthologous DNA sequences with high similarity), based upon a hypothesis that functionally important sequences evolve more slowly than the non-functional sequences in the neighbourhood (Wasserman et al. 2000). This strategy is called “phylogenetic footprinting” and is used for identification of conserved non-coding regions. Initially, this method included cloning and sequencing of orthologous non-coding sequences from two or more organisms. Later, when the whole human and mouse genomes were available, global sequence comparisons between genomes became the most commonly employed approach in comparative studies (Ahituv et al. 2004). In many genomic regions the evolutionary divergence between mammals is not sufficient to select neutrally evolving sequences from functionally constrained ones. Multiple genome comparison of species of comparable evolutionary divergence or the use of evolutionary distant species for pair-wise comparisons can better highlight those non-coding elements, which are most likely functional, as the increase of the total phylogenetic branch length enables the removal of similarities between neutrally evolving sequences (Venkatesh et al. 2006). The initial observation of the compactness of the fugu genome (7.5 times smaller than the human) led to the suggestion that genes and non-coding sequences conserved between these species would represent the minimal set of genes and regulatory elements required to construct a vertebrate organism (Brenner et al. 1993; Aparicio et al. 1995). On the basis of the first reports showing functional conserved regulatory elements, a conventional threshold was created for the identification of human-fish non-coding elements, requiring 70% identity over a minimum size of 100 bp (Ahituv et al. 2004).

Phylogenetic footprinting, using either mouse-human or fish-human comparisons, has been useful to select candidate regions, which were functionally tested by transgenic assays (Muller et al. 2002; Nobrega et al. 2003; de la Calle-Mustienes et al. 2005; Goode et al. 2005; Poulin et al. 2005; Woolfe et al. 2005). Enhancer databases, like the Vista Enhancer Browser (Visel et al. 2007b), the Condor (Woolfe et al. 2007b) or the Ancora (Engstrom et al. 2008) databases collected evolutionary conserved vertebrate non-coding sequences with enhancer activity.

The level of conservation of non-coding sequences in some cases is extremely high. Comparative analysis of the human, mouse and rat genome revealed 481 genomic regions termed as **ultraconserved elements (UCEs)**, which shared 100% homology with no insertions or deletions over 200 bp. They are widely distributed in the genome and often found in clusters. These elements show extremely high sequence conservation with orthologous regions in chicken and fugu as well, and in the human population these ultraconserved elements exhibit extremely low level of natural variation (SNPs). Two third of these elements are non-exonic (256) or possibly exonic (probably non-coding), showing a tendency of congregating in clusters near transcription factors and developmental genes, or located in gene deserts (Bejerano et al. 2004). Several other studies investigated the abundance and the function of sequences with such a high degree of conservation. Changing the search criteria (decreasing the length of the conserved fragments to 50 bp) resulted in higher number (roughly 3500) of ultraconserved non-coding regions (UCR) between human, mouse and pufferfish. This study also found these elements clustering near to genes that act as master regulators during vertebrate development (Sandelin et al. 2004). Plessy et al. (2005) performed systematic analysis of experimentally verified mouse enhancers, and could show that genes with enhancers conserved between mouse and zebrafish were significantly enriched in developmental regulators (Plessy et al. 2005). Another studies focused on highly conserved non-coding elements (CNEs) found between human and fugu (with a minimum length of 100 bp), related to transcriptional regulator or developmental genes. Functional analysis of a small portion of these CNEs was performed, and the majority of the tested elements showed enhancer-like activity in transient expression assays in zebrafish, compared to non-conserved non-genic regions (Woolfe et al. 2005; McEwen et al. 2006). Ultraconserved elements were also subject to in vivo functional tests. Paparidis et al. (2007) reported that an ultraconserved non-coding element from the second intron of

gli3 gene was a transcriptional enhancer (Paparidis et al. 2007). Pennacchio et al. (2006) tested 167 UCEs, and could demonstrate that 45% of these sequences could work as enhancers in transgenic mouse assays (Pennacchio et al. 2006) Besides the enhancer activity, ultraconserved non-protein coding sequences can function as splicing regulators (Lareau et al. 2007; Ni et al. 2007), factors of epigenetic modifications (Bernstein et al. 2006; Lee et al. 2006), transcriptional co-activators (Feng et al. 2006) or encode a particular set of noncoding RNA (ncRNA) (Calin et al. 2007). The ultraconserved element located between the *dlx5* and *dlx6* genes codes a noncoding RNA (Evf-2), which is able to increase the transcriptional activity of Dlx2 on the *dlx5/dlx6* locus, by forming a stable complex with the transcription factor. This particular example shows that a subset of vertebrate ultraconserved regions may function at both the DNA and RNA level to control key developmental regulators, and may explain why ultraconserved sequences exhibit 90% or more conservation even after 450 million years of vertebrate evolution (Feng et al. 2006).

Target gene		Bystander gene
shh		*lmbr1*
gremlin		*formin*
pax6		*elp4*
nkx2.8 – pax9		*scl25a21*
foxL2		*mrps22*
cd79b-hgh		*scn4a*
otp		*ap3p1*
fgf8		*fbxw4*
barhl1		*ddx31*
mir9		*mef2*

Table 3: Examples of gene interdigitation

From Kikuta et al. (2007)

Conserved synteny blocks are stretches of chromosome similarities where orthologous protein coding sequences are located on the same chromosome and in the same linear order in more than one species (Barbazuk et al. 2000). Chromosomal rearrangement events within genomes are not completely random, a significant portion occurs within similar parts of the genome. MacKenzie et al. (2004) hypothesise that long-range gene interdigitation and the ability of individual cis-regulatory elements directly affect the expression of many genes at a distance and thus contribute to the persistence of conserved synteny blocks in higher vertebrates. They claim that these sites represent the border of areas permissive to translocation events through evolution, as translocations do not disrupt any functional linkage, such

as a cis-regulatory element - target gene. Kikuta et al. (2007) showed (by using fish-human genome comparisons) that target genes of long-range cis-regulatory elements and their phylogenetically and functionally unrelated bystander genes, in which the regulatory elements reside, form regions of conserved synteny, confirming the hypothesis of MacKenzie et al. (Table 3 shows examples of interdigitation). Single copies of these genomic regulatory blocks (GRBs) are protected from chromosomal breakage, while in cases of teleost duplication of GRBs, bystander genes have often lost by neutral evolution. They claim that combination of human-teleost synteny, enhancer detection and GRB duplication analysis allows recognition of target versus bystander genes and permits annotation of highly conserved elements to target genes within a syntenic chromosomal segment. Based upon their analysis, genes encoding developmental transcriptional regulators tend to be surrounded by larger regions of synteny than other functional categories of genes (Kikuta et al. 2007).

Some conserved non-coding sequences function as enhancers in gain-of-function assays, but in contrast, some apparently **constrained non-coding DNA** sequences have **no obvious function**, and some functional cis-regulatory elements do not show any conservation (Fisher et al. 2006a). Because many genes show different expression patterns even between human and mouse, there is no reason to expect that all cis-regulatory elements to be under the same level of constraint. Based on the ENCODE protein occupancy and chromatin modification data gained from the 1% of the human genome, King et al. (2007) defined a set of putative transcriptional regulatory regions (pTRRs) and used the promoters and DNaseI hypersensitive sites analysed by the ENCODE consortium, and tested these sequences for conservation. They have found that while most classes of non-coding functional elements (pTRRs, promoters and DNaseI hypersensitive sites) are enriched for multispecies constrained sequences (MCS), many of the functional non-coding elements are not constrained. They suggest that these MCSs select for only a very highly constrained subset of regulatory elements and miss many other regions that are under constraints. The genes nearest to the conserved pTRRs were checked for gene ontology, and they have found that different classes of elements tend to be constrained over different phylogenetic spans (King et al. 2007). Visel et al. (2008) tested for embryonic enhancer activity 231 non-coding ultraconserved human genomic regions out of the total 256 existing, and 206 extremely conserved regions lacking ultraconservation in transgenic mice.

They found no differences between these two categories in the number of sequences working as enhancers, equally half of the sequences from both groups drove expression of the reporter gene in various tissues in the developing mouse, and they could not find any tissue or anatomical region where the ultraconserved sequences selectively activated transcription (Visel et al. 2008). Nevertheless ultraconserved elements have remained frozen during mammalian evolution; a relatively small number of them may more likely to be functional due to their higher level of conservation.

A recent study demonstrated that knock out mice, in which conserved non-coding sequences with enhancer function were deleted, showed no severe detectable phenotype during the development, and the deletions did not affect the viability of the mice (Ahituv et al. 2007). Based upon their results Ahituv et al. questioned the relevance of the in vivo function of these constrained sequences, not taken to account that only one aspect of the complex endogenous mRNA expression is affected when only a single enhancer is deleted, while expression of the gene in other tissues or at other stages is maintained. Moreover, deletion of functional enhancers were previously shown to cause little or no phenotypic changes for the *engrailed2* (Li Song et al. 2000), *fgf4* (Guyot et al. 2004), *gata1* (Guyot et al. 2004) or *myoD* (Chen et al. 2004) genes. Deletion of a single *hoxd11* enhancer in mice does not cause severe defects, just delays the expression of *hoxd10* and *hoxd11*, in later stages the normal expression is restored by complementary regulatory elements (Zakany et al. 1997). Functional redundancy can also give an explanation to this phenomenon, which has been shown for the *sgs-4* developmental gene in Drosophilals (Jongens et al. 1988), for the *tcr-gamma* locus (Xiong et al. 2002), or for the *shh* gene (Jeong et al. 2006): several enhancers can be responsible for the expression in a given tissue, and the deletions of single enhancers not necessarily cause major changes, but deletion of all enhancers results in a severe reduction of the given gene.

Investigation of the *Latimeria menadoensis* (coelacanth) genome led to the identification of an ancient SINE (short interspersed elements, 75-500 bp long retrotransposons that contain internal promoters for RNA PolIII)) family, the members of which are related to SINEs present in mammals, birds and in fish species. These retrotransposon-derived sequences are not only present in these species, but more than 100 human copies are highly conserved among mammalian orthologs (Nishihara et al. 2006). One member of this lungfish-SINE family from the human

genome, located 500kb from the *isl1* gene, showed *isl1*-specific enhancer activity (Bejerano et al. 2006). Recent research demonstrated high abundance of transposable element-derived sequences in mammalian genomes (Mikkelsen et al. 2007), and based on Bejerano et al. (2006), 5,5% of all the conserved non-coding sequences originated from transposons – the high level of conservation should be revisited. Inter-lineage transfer and intraspecies proliferation of transposable elements can cause high levels of sequence similarities between element copies in different lineages, as the time of divergence of these "junk" elements is different from the one of the host genome.

1.10 Experimental approaches to verify cis-regulatory elements

Biochemical analysis methods, like DNase hypersensitivity assay, electromobility shift assay and chromatin immuneprecipitation can be used to determine whether a given sequence is bound by transcription factors, but these do not provide information about the in vivo relevance of the TF-binding. **Reporter-gene assay** is a generally used method to identify and analyse transcriptional regulatory activity in vivo of given DNA sequences: the piece of DNA of interest is cloned in front of a reporter gene – e.g. *chloramphenicol acetyl transferase (cat), β-galactosidase, luciferase* or a *fluorescent protein* (like *gfp*) gene. If a putative core promoter is assayed, no additional sequence is added, but in the case of analysing distal regulatory elements, a weak promoter is attached in front of the reporter gene. Then the construct is transformed into cell culture, and the activity of the reporter is compared to a control construct. After the given genomic region is identified as a regulatory region, serial deletions, linker scanning mutagenesis, or site-directed mutagenesis can be applied for more precise analysis. The in vivo activity of a reporter gene may fail to recapitulate the endogenous gene activity even with the full set of its cis-regulatory elements due to different chromatin context. Furthermore, it is possible that a given element is only used in limited context such as in a specific tissue, developmental stage or physiological response. To overcome these issues, one can use **model organisms** such as mouse, frog or zebrafish. After microinjection of the constructs into embryos of these animals, the expression of the reporter can be detected throughout the development. By generating stable transgenic lines, the expression of the reporter gene can be followed in different tissues and in different conditions as well, but large-scale screens are not easily manageable in this way. Loss

of function studies, like deletion of a cis-regulatory element allows analysis of requirement of the regulatory architecture of a locus (Yanagisawa et al. 2003). Tissue and time-specific knockouts can be generated as well with the use of the CRE/lox system (Gu et al. 1994) to check the requirement for function by studying the direct effect of the loss of a regulatory sequence (Vong et al. 2005).

1.11 Zebrafish as a model organism

Zebrafish (*Danio rerio*) are easy to keep and breed under laboratory conditions. Females produce large number of eggs that develop externally and easy to manipulate. They not only have short generation cycle (approximately three months), but the embryos are transparent and develop rapidly. In 48 hours after fertilization at 28^0C a free-swimming larva develops from a fertilized egg. The genome assembly and the annotation of zebrafish genes are close to the finish at the time of writing.

Large-scale identification of zebrafish mutations affecting early embryogenesis (Driever et al. 1996; van Eeden et al. 1999; Burkhart 2000) led to the identification of several genes as key players in vertebrate gastrulation, brain development and midline signalling (Feldman et al. 1998; Griffin et al. 1998; Karlstrom et al. 1999). These findings not only founded studies of early developmental mechanisms, but also laid the ground to establish zebrafish as a model for human diseases. In the last decade zebrafish models have been established to elucidate the molecular mechanisms of human diseases like cardiovascular defects, muscle- and neural disorders, haematopoiesis and cancer (Zon 1999; Dooley et al. 2000; Amsterdam 2006).

The transparency of the developing embryo gives a unique quality to whole-mount in situ (Fjose et al. 1992) and antibody staining (Wilson et al. 1990). The genetic analysis in zebrafish have been furthermore facilitated by the completion of the zebrafish genome sequencing and by the improvements in the assembly quality and gene annotations, which makes it a suitable model for comparative genomic studies as well. The fast ex utero development in water and the easily detectable phenotype changes due to the transparency of the embryos were the advantages why toxicological studies started to use zebrafish for environmental and chemical toxicity tests (Van Leeuwen et al. 1990).

These approaches, in combination with the zebrafish sequence and genome-wide gene expression studies (Stickney et al. 2002; Lo et al. 2003), provide the

possibility to understand the vertebrate development, disease susceptibility and evolution in more detail.

1.11.1 Transcription regulation analysis in zebrafish

In the past several decades, it has become clear that the expression and function of a variety of regulatory genes guides developmental processes, thus studying the properties of differentially regulated gene expression during embryogenesis became a highly effective way to investigate developmental mechanisms.

Zebrafish mutant screens characterized genes at a molecular level, and several of these genes have been found to code transcription factors e.g. (Schulte-Merker et al. 1994; Talbot et al. 1995; Brand et al. 1996). Gene expression analyses in mutants revealed transcriptional pathways by the identification of direct downstream target genes for example in axis formation (Strahle et al. 1993; Chang et al. 1997), somatic muscle development (Weinberg et al. 1996; Griffin et al. 1998; Yamamoto et al. 1998), hindbrain patterning (Moens et al. 1996; Prince et al. 1998), neural crest development (Henion et al. 1996), neuronal phenotype (Guo et al. 1999) and heart development (Alexander et al. 1998).

The development of transgenesis in zebrafish by microinjection of linearized plasmid DNA in the late eighties (Stuart et al. 1988) and the short generation cycle of the fish provided the possibility for transcription regulation studies e.g. (Long et al. 1997; Meng et al. 1997; Meng et al. 1999) in zebrafish, Japanese medaka fish (*Oryzias latipes*) and Xiphophorus (Winkler et al. 1992) by injecting a promoter followed by a reporter gene . The most generally used reporters are fluorescent proteins, which provide the investigation of the transgene in living animals (Amsterdam et al. 1995). The usage of bacterial artificial chromosomes provided the possibility of injecting large fragments of DNA (Jessen et al. 1999). Generating transgenic zebrafish is although laborious. First, generating the desired expression constructs by conventional subcloning can require multistep cloning strategies, because the choice of restriction enzymes is often limited for long genomic or cDNA fragments. Long-range PCR methods can circumvent some of these problems, but require resequencing of coding sequences. Second, rates of germline transgenesis are low with plasmid-based transgenesis, requiring the injection, raising, and screening of scores to hundreds of potential founders to ensure recovery of a stable line. Injection of supercoiled or linear DNA yields 1-10% germline transgenic founders (Stuart et al.

1988; Stuart et al. 1990), while linearization with I-SceI meganuclease yields 20-30% germline transgenic founders (Thermes et al. 2002). Retroviral and transposon-based insertions have dramatically increased the transgenesis rate to 30% with Sleeping Beauty (Davidson et al. 2003) or 50% with Tol2 (Kawakami 2004) and with pseudotyped retrovirus (Laplante et al. 2006). The generation of enhancer detection or enhancer trap lines, where in each line the expression of a reporter gene is under the transcriptional control of tissue-specific enhancers, provided transgenic fish with differentially marked cells or tissues. These lines are particularly useful for studying development of distinct organs, to analyze the effects of other genes or toxic compounds on the marked cells/tissues, or simply to collect information about the in vivo expression pattern of genes during development (Amsterdam et al. 2005). GAL4-UAS bitransgenic zebrafish lines have been developed for efficient tissue-specific and temporally controlled transgene expression to mark cell types or ectopically express proteins (Scheer et al. 1999; Koster et al. 2001; Scheer et al. 2002; Thummel et al. 2005).

Injection of multiple different DNA sequences, such as activating sequences and gene fragments with a reporter construct, is also possible (Muller et al. 1997). This co-injection approach exploits the rapid concatamerisation of injected DNA in fish embryos (Stuart et al. 1988; Winkler et al. 1991) and by-passes the need to generate multiple expression constructs. To exclude the generation of stable transgenic lines, the mosaic transient transgene fish can be monitored for expression as well. The high degree of mosaicism observed in the injected fish is due to the fact, that cytoplasmically injected foreign DNA is compartmentalised into a subset of cells in the cleaving embryos (Westerfield et al. 1992), and persist mainly extrachromosomally. Despite the mosaic expression, cell-type-specific gene expression can be analyzed by generating a large number of transgenic animals and summing up their expression (Muller et al. 1997).

Although gene knock out is not jet possible in zebrafish, microinjections of mRNAs or morpholino oligonucleotides can result in specific inactivation of genes (Nasevicius et al. 2000). For example DNA microarray analysis was performed in morpholino knock down embryos to determine the generality and function of TBP (Ferg et al. 2007).

Genomic microarray coupled with chromatin immuneprecipitation (ChIP-Chip) can be used in zebrafish as well to determine the genomic binding locations of DNA

interacting proteins during development and investigate the assembly of the genetic networks that regulate embryogenesis (Wardle et al. 2006).

1.11.2 Large-scale and high throughput screening methods using zebrafish

High throughput screens (HTS) provide the possibility to quickly perform large-scale biochemical, genetic or pharmacological tests by automated sample handling and/or programmed data collection and processing. Zebrafish are highly reproductive and the embryos develop *ex vivo*, thus an ideal model for whole-organism gene expression studies.

The first large-scale screens performed with zebrafish were systematic genome-wide mutagenesis screens, which led to the identification of thousands of mutations in genes affecting early zebrafish development. These screens used either chemical mutagens like N-ethyl-N-nitrosourea (ENU) (Driever et al. 1996; Haffter et al. 1996) or mouse retroviral vectors (Amsterdam et al. 1999; Chen et al. 2002). TILLING (Targeting Induced Local Lesions in Genomes), a traditional chemical mutagenesis followed by high-throughput screening for point mutations (Wienholds et al. 2002) further provided large number of mutant lines (Henikoff et al. 2004). Transgenic fish expressing fluorescent proteins provide real-time readouts of phenotype. Transposon- or retrovirus-based gene trap or enhancer trap experiments not only yielded in large-scale stable transgenic lines, where specific tissues and cells are labelled with fluorescent proteins, but also provided insights of transcription regulation. Behavioural outcomes can also be screened as a phenotype (Bang et al. 2002; Gahtan et al. 2004). As zebrafish has been found to be a useful tool for toxicological analysis, high-throughput assays were developed for testing bioactive compounds, including drugs, pesticides and industrial by-products either using developing embryos or adults, even in a microtiter plate (Parng et al. 2002; Milan et al. 2003; Pichler et al. 2003; Behra et al. 2004; Kokel et al. 2008; Lam et al. 2008).

Expression patterns of genes playing role in regulation of development were determined by using high throughput in situ hybridization in several large-scale screens (Kudoh et al. 2001; Pollet et al. 2001; Thisse 2001; Wienholds et al. 2005; Visel et al. 2007a; Thisse et al. 2008). The sequencing of the zebrafish genome and the extensive collection of expressed sequence tags have led to the development of many commercial or self-designed microarrays for defining the set of genes expressed. The method based on hybridization of the transcripts to immobilized

cDNA accelerated the molecular analysis of zebrafish mutants (Stickney et al. 2002), but more importantly unravelled defined developmental processes and biochemical pathways (Ton et al. 2002; Lo et al. 2003; Hedlund et al. 2004; Mathavan et al. 2005; Sumanas et al. 2005; Giraldez et al. 2006; Xu et al. 2006). The outcome of the different treatments can be assayed on the transcriptome level as well (Yang et al. 2007b; Kily et al. 2008; van Boxtel et al. 2008) – even the effect of chronic tuberculosis on gene expression of non-treated adults was investigated this way (Meijer et al. 2005).

Functional tests, like dissection of the functions of particular isoform combinations of large multiprotein complexes (using in situ hybridization studies and antisense-based reverse genetic knockdowns) (Cheng et al. 2003), analysis of metabolic pathways (Ho et al. 2003) or measurement of circadian gene expression in vivo (Kaneko et al. 2005) were also adapted to analyze large number of zebrafish embryos.

As the micro-manipulation and handling of a large number of fish embryos is time-consuming and laborious, methods like automated micro-injection (Wang et al. 2007) or embryo-handling (Furlong et al. 2001) have been developed to provide the opportunity of automated manipulation and sorting embryo samples in standard conditions. To increase the facility and throughput of scoring phenotypic traits in zebrafish, automated fluorescence microscopy of transgenic embryos expressing GFP were developed in a microtiter plate format (Burns et al. 2005). Fully automated fluorescence stereomicroscopes were utilized for time-lapse imaging of transgenic embryos (Distel et al. 2006). Using three-dimensional image recording, spatial reconstruction of expression patterns was possible, moreover, by combining three-dimensional image recording over time with subsequent deconvolution analysis, subcellular dynamics could be resolved (Distel et al. 2006). Although automated microscopic picture taking can speed up screens, the tremendous number of digital images generated from large numbers of embryos frequently leads to a bottleneck in data analysis and interpretation. The development of algorithms recognizing tissues or specific cell types and changes of reporter signals within these have been reporter in the last two years (Li et al. 2007; Tran et al. 2007; Zanella et al. 2007; Liu et al. 2008). Unfortunately these are specified to distinct cells, tissues, developmental stages or microscopes to be used generally.

2) Objectives

Our laboratory is interested in cis-regulatory mechanisms regulate vertebrate embryonic development. The following topics raised questions, and to answer these questions, different projects were designed.

1. When conserved non-coding sequences (CSTs) around a particular gene of interest are tested for enhancer activity, generally the endogenous promoter is used for the assay. In contrast, larger scale studies usually prefer to use the basal promoter of a ubiquitously expressed gene to avoid cloning of each and every endogenous promoter. But do these conserved non-coding sequences show the same results, when tested with different promoters? Enhancer trap experiments performed with different promoters could fish out different sets of enhancers, giving a hint that the promoter specificity of enhancers observed in Drosophilal is a valid phenomenon in vertebrates as well.

To test whether predicted cis-regulatory elements show preference toward their endogenous promoters, I aimed to test CSTs determined by phylogenetic footprinting from the *pax2* locus for enhancer activity with an endogenous and a heterologue promoter.

2. Phylogenetic footprinting relies on computer programs that compare large pieces of genomic DNA from multiple species. Different alignment methods give slightly different results when the same genomic regions are used as templates. Local alignment approaches compare relatively short intervals of genomic sequences with each other and return the best match between two genomes for each subregion. Because they do not take into account the regions surrounding these matches, they can result in false hits. Global alignment tools align entire syntenic regions, and return less false positive matches, but not sensitive to rearrangements. Knowing these, we wanted to know in what extent does the algorithm of the phylogenetic footprinting prejudice the outcome of the search, and whether these algorithms can be further developed to better predict cis-regulatory elements.

For this project, a collaborator partner designed a new sequence comparison method, which resulted in large number of conserved non-protein coding sequences. My aim was to test a subset of these elements for enhancer activity.

3. The partially overlapping results gained with the CSTs from the *pax2* locus with two different promoters raised further questions: What is the level of the cis-regulatory element interaction specificity? Do the different properties of core promoters determine which enhancers can interact with them? Or does the chromatin structure or other DNA elements located in the original genomic context needed for the interaction specificity?

To answer these questions, I aimed to perform a high throughput analysis: cloning of a set of enhancers in combination with a set of promoters, injection of these constructs into zebrafish embryos. Our aim was to develop computational algorithms for automated picture acquisition and quantification to handle the enormous amount of data.

3) Materials and methods

Chemicals, if not mentioned, were purchased from Sigma-Aldrich.

3.1 Standard molecular cloning

3.1.1 Isolation of zebrafish genomic DNA

The genomic DNA isolation was performed with Qiagen DNeasy Tissue Kit.

To 100 embryos in a 1.5 ml microcentrifuge tube 180 µl ATL buffer and 20 µl Protease K were added, and the tubes were incubated at 55^0C for at least 3 hours. After vortexing the samples, 200 µl AL buffer was added, and the vortexed samples were incubated for 10 minutes at 70^0C. Then 200 µl 100% ethanol was added, and after vortexing, the solutions were pipetted into DNeasy spin column, centrifuged for 1 minute at 13.000 rpm at room temperature. After discarding the flow-through, the collection tubes were replaced. 500 µl AW1 buffer was added onto the columns, which were centrifuged at 13.000 rpm for 1 minute at room temperature. After discarding the flow-through, the collection tubes were replaced. 500 µl AW2 buffer was added onto the columns, which were centrifuged at 13.000 rpm for 3 minutes at room temperature. The columns were placed into new microcentrifuge tubes, 200 µl AE buffer was added into the membranes, and he columns were centrifuged for 1 minute at 13.000 rpm at room temperature. The concentrations of the DNA solutions were measured on the Nanodrop machine.

3.1.2 Amplification of DNA sequences from zebrafish genomic DNA

The amplification was carried out with Eppendorf Triple Master and dNTPs with sequence specific primers. (See primer sequences under chapter 5.1.16 and 4.2.6) The composition of the PCR reaction:

Master Mix1	
Genomic DNA	0.5µl
Primer1	0.25µl
Primer2	0.25µl
Nuclease free water	9µl
Total:	10µl

Master Mix2	
High Fidelity buffer	5µl
dNTP 10mM	1µl
enzyme	0.5µl
nuclease free water	33.5µl
total:	40µl

The two Master Mix solutions were mixed only in the PCR tubes, just prior the cycles started.

The program used: TM01

1. 93^{0}C	1.00 min	5. GOTO 2	29 times
2. 93^{0}C	0.15 min	6. 68^{0}C	10.00 min
3. 56^{0}C	0.30 min	7. 4^{0}C	forever
4. 68^{0}C	2.00 min	8. END	

3.1.3 Restriction digest

The restriction enzymes were purchased from Promega. All enzymatic reaction were performed according to the following basic protocol: 3 Units of enzyme was used to each mg of DNA, the reaction mix contained the enzyme-specific buffer and 1% BSA. The reaction was incubated at 37^{0}C for at least 3 hours, and then the fragments were gel-purified.

3.1.4 Gel-purification of PCR products and restriction fragments

The Promega SV Gel and PCR Clean-Up System was used.

The DNA was run on a 1% agarose gel containing 0,1v/m% Ethidium bromide. The desired fragments were cut out from the gel under UV light, and the gel slices containing the DNA fragments were put into microcentrifuge tubes and measured. 10µl Membrane Binding Solution was added to each 10mg of gel slice; the tubes were mixed and incubated at 65^{0}C until the gel slices were completely dissolved. The dissolved gel mixtures was transferred into the SV Minicoulmns inserted into collection tubes, incubated for 1 minute at room temperature, then centrifuged at 13.000 rpm for 1 minute. The flow through was discarded and the Minicolumns was reinserted into the Collection Tubes. 700µl Membrane Wash Solution was added into the columns, they were centrifuged at 13.000 rpm for 1 minute. The flowthrough was discarded and the Minicolumns were reinserted into the Collection Tubes. The washing step was repeated with 500µl Membrane Wash Solution, but after discarding the flowthrough, the tubes were centrifuged once more at 13.000 rpm for 1 minute. Then the columns were transferred into clean microcentrifuge tubes, 50µl nuclease-free water was added into the columns, and after 1-minute room temperature incubation, they were centrifuged at 13.000 rpm for 1 minute. The Minicoulmns were discarded and the DNA samples were kept at -20^{0}C.

3.1.5 Ligation

For the ligation of DNA fragments produced by restriction digest, T4 DNA was used from Promega. The vector: insert ratio was 1:3; the reaction contained the ligase-specific buffer. The reaction was incubated at 4^0C overnight, and then transformed into competent bacteria.

3.1.6 TOPO-cloning

The TOPO pCRII was purchased from Invitrogen.
PCR products amplified by TripleMaster enzyme mix were incubated with 1U GoTaq for 10 minutes at 72^0C for adding 3`A overhangs.
4µl PCR product was mixed with 1µl Salt solution and 1µl of TOPO vector. The reaction was incubated at room temperature for 30 minutes, then placed on ice, and transformed into competent bacteria.

3.1.7 Preparation of chemically competent bacteria

A single colony of bacteria was inoculated in 10 ml of LB media for overnight growth at 37^0C. The next day 1 ml of the overnight culture was inoculated into 200 ml of LB media, kept at 37^0C until the OD at 600 nm reached 0.3-0.4. Then the bacteria were kept on ice for 10 minutes, centrifuged at 5000 rpm at 4^0C for 10 minutes. After discarding the supernatant, the bacterial pellet was re-suspended in 40 ml of ice-cold 0.1M $CaCl_2$, and kept on ice for 1 hour. After the incubation, the bacteria were centrifuged at 5000 rpm at 4^0C for 10 minutes, the supernatant was discarded and the bacterial pellet was re-suspended in 20 ml 0.1M $CaCl_2$ supplemented with 15% glycerol, and 50 µl aliquots were fast-frost in liquid nitrogen, and kept at -80^0C.
The bacterial stains used: TOP10, Mach1T1R, ccdB-Survival T1R

3.1.8 Plasmid transformation into chemically competent bacteria

The competent cells were defrosted on ice. 1-10 µl of plasmid DNA was added to each vial of bacteria, and incubated on ice for 30 minutes. Then the bacteria received a short heat-shock (42^0C for 45 seconds), and after a 20-minute incubation on ice, 250µl SOC media was given to each vial, and the bacteria were shifted to 37^0C, and were shaken for 1 hour. Then 200 µl were spread on antibiotic-containing LB-plates, which were kept at 37^0C for overnight.

3.1.9 Preparation of electro-competent bacteria

1 colony on the specific bacteria was inoculated for overnight culture in 20 ml of SOC media. The next day, the 10 ml of the culture was inoculated into 200 ml of pre-warmed SOC media, and the culture was grown at 37^0C until the OD at 600 nm reached 0.3. The cells were chilled on ice for 15 minutes after transferring them into centrifuge tubes. The bacteria were centrifuged for 15 minutes at 4^0C at 3000 rpm, the supernatant was discarded, and the bacterial pellet was gently re-suspended with a pipette in 200 ml of ice-cold water. The cells were centrifuged again with the same parameters, and the pellet was re-suspended in 200 ml in ice-cold water. The centrifuging step was repeated, and the pellet was re-suspended in 100 ml of ice-cold water containing 10% glycerol. The centrifuging step was repeated and the bacterial pellet was re-suspended in 2ml of ice-cold water containing 10% glycerol. 40 µl aliquots were fast-frost in liquid nitrogen, and kept at -80^0C. The bacterial stain used: Mach1T1R

3.1.10 Electroporation of plasmid DNA

1-2µl plasmid DNA was mixed with the competent cells defrosted on ice, the mixture was moved to 1mm electroporation cuvettes, and the electroporation was performed with the following parameters: 2500V. After the electroporation 460µl of LB media was added into the cuvettes, and the mixture was transferred into microcentrifuge tubes. The bacteria was kept at 37^0C for 1 hour in a water bath, then half was spread

3.1.11 Identification of colonies by PCR reaction

Single colonies were picked from the plates with yellow pipette tips, and were rinsed into the PCR mix, then into 3 ml LB media for inoculation the miniprep cultures.

The PCR reaction:	
5x coloured buffer	4µl
dNTP 10mM	0.5µl
primer1 100nM	0.1µl
primer2 100nM	0.1µl
GoTaq	0.5µl
Nuclease free water	15µl
Total:	20µl

The program used:

1. 95^0C 1.30 min
2. 95^0C 0.20 min
3. 54^0C 0.20 min
4. 72^0C 2.00 min
5. GOTO 2 29 times
6. 72^0C 5.00 min
7. 4^0C forever
8. END

3.1.12 Plasmid miniprep

The Qiagen QiaPrep Spin Miniprep Kit was used.

3 ml of antibiotics-containing LB was inoculated with a single colony and incubated at 37^0C for overnight. Next day the bacteria were collected in microcentrifuge tubes by centrifugation at 10.000 rpm for 2 minutes in a table centrifuge. The supernatant was discarded and the bacterial pellet was re-dissolved in 250µl P1 solution. 250µl P2 solution was added, and after gentle mixing and 4 minutes of room temperature incubation 300µl P3 solution was added as well, and after gentle mix the tubes were centrifuged at 13.000 rpm for 15 minutes. The supernatant was transferred into the Qiaprep spin columns, the columns were centrifuged at 13.000 rpm for 1 minute, washed with 750µl Buffer PE, centrifuged again at 13.000 rpm for 1 minute, and after the follow-through was removed, the centrifugation step was repeated once more. Then the columns were transferred into clean microcentrifuge tubes, and the DNA was eluted with 50µl Buffer EB, with 1-minute centrifugation at 13.000 rpm.

3.1.13 Plasmid maxiprep

The Qiagen Plasmid Maxi Kit was used.

100ml antibiotic-containing LB was inoculated with a single colony. and incubated at 37^0C for overnight. Next day the bacteria were centrifuged at 5.000 rpm for 20 minutes at 4^0C. The supernatant was discarded and the bacterial pellet was re-dissolved in 10ml P1 solution. 10ml P2 solution was added, and after gentle mixing and 4 minutes of room temperature incubation 10ml P3 solution was added as well, and after gentle mix the tubes were centrifuged at 10.000 rpm for 30 minutes. The supernatant was transferred into the equilibrated Qiagen maxiprep columns and let to flow through by gravity flow. Then the columns were washed with 2x 30 ml QC buffer, and finally the DNA was eluted with 15 ml QF buffer. The plasmid DNA was precipitated with 10.5ml isopropanol, and was centrifuged at 4^0C for half an hour at 10.000 rpm. The pellet was washed with 70% ethanol, then air-dried and re-dissolved in 200µl nuclease free-water.

3.1.14 Plasmid DNA sequencing

The plasmids were sent to GATC Biotech AG for sequencing.

3.1.15 In vitro transcription of an in situ probe

The Ambion Message Machine Kit was used for this purpose.

2 µl 10x transcription buffer
2 µl DIG RNA labelling kit
1 µl RNase inhibitor
1 µg template
1 µl T3 polymerase or 2 µl T7/ SP6 polymerase
adjusted to 20 µl with nuclease-free water

The reaction was incubated for 2 hours at 37^0C, then the nucleic acids were precipitated with 10 µl 7.5M NH_4Oac and 75 µl 100% ethanol, incubated for 20 min at -80^0C, then centrifuged for 15 minutes with 12000 rpm at 4^0C. The pellet was washed with 80% ethanol, air-dried, and re-suspended in 30 µl HYB-buffer.

3.1.16 Primer sequences

	Forward primer	**Reverse primer**
eng2b promoter	ACTGGAGTGAATTGTTTTTCGTTG	TGAAACTCTCCAAATGTTC
eng2b CXE	CGATACACTTTGATGATACGCATTG	GCTCACATGACATTTCTCATTTTCC
eng2b reg5	TATCTTGTCCCCATTCCAACAGAG	ATGTCAGCCAGAATGGTCAAAAAC
mef2d promoter	CATGTGCTTAAGGGAACGTTAAATAA	ACAGTCAAAACCCTCCATGTACAGAG
dre-mir9-1 promoter	GAGGGTAAATCTGCGGAAAACTAAAGCA	GGCTTGCTCTCACAATAAAATGATGCA
elp4 promoter	CTAGTTCAGAAAGCTGTCGGTTTCA	ACTGAGCTTCAACCCATCGAATAAT

3.2 The Multisite Gateway cloning

3.2.1 The principals of the Multisite Gateway technology

The Gateway cloning is based on site-specific recombination, excluding the need of restriction digests and ligations, which is advantageous when cloning several hundreds of constructs (Hartley et al. 2000; Walhout et al. 2000). The cloning system comprises two distinct recombination reactions that rely on cis elements and excision/integration enzyme complexes from bacteriophage λ. The BP clonase (a mixture of the λ phage integrase and integration host factor) catalyses the BP recombination between the PCR product and the "donor" vector yielding to an "entry" vector. The Gateway system relies on a counter-selection method for high efficiency recovery of entry plasmids. The donor vectors contain a *ccdB* cassette coding a protein that is toxic to standard bacterial strains, plus a chloramphenicol resistance gene, which allows the maintenance of the cassette in *ccdB*-tolerant cells. Therefore, un-recombined donor vectors will not be propagated after the BP recombination, only the entry vectors. The entry vectors, together with the destination vectors, then further used in an LR recombination reaction to generate the final expression vectors. This reaction is catalysed by the LR clonase, a mixture of the integrase, integration host factor and excisionase. In this case, the destination vector contains the *ccdB* – chloramphenicol resistance cassette, so these are constraselected after the LR reaction. The Multisite Gateway system provides the possibility of building up complex expression vectors, containing an ORF of interest, a promoter and a cis-regulatory region by combining different entry clones (Figure 61.).

For cloning roughly 260 different enhancer-promoter combinations we used a modified version (Table 16.) of the commercially available Multisite Gateway system (Roure et al. 2007).

vectors for cloning	**after recombination**
pDONR-221-P1/P2	pENTRY-L1-promoter-L2
pDONR-221-P3/P5	pENTRY-L3-enhancer-L5
pSP72-R3-ccdB/cmR-R5::RfA-venus	pSP72-B3-enhancer-B5::B1-promoter-B2-*venus*

Table 16: The Multisite Gateway vectors used for the high throughput screen

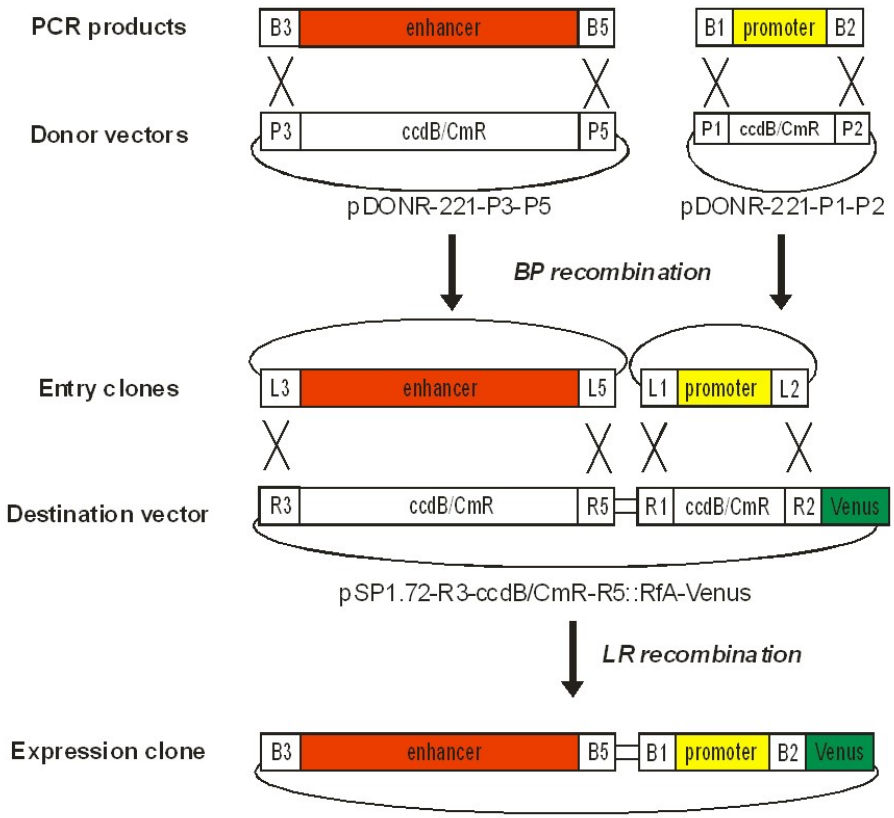


Figure 61: The scheme of the Multisite Gateway cloning

3.2.2 Amplification of DNA sequences

The amplification was performed in two steps, for the first reaction sequence-specific primers were used which contained a short adapter stretch (see standard molecular cloning methods), and then the first reaction was used as a template for the second PCR made with the adaptor primers (primer sequences in Table 17.). The two Master Mix solutions were mixed only in the PCR tubes, just prior the cycles started. The composition of the Gateway second PCR reaction:

Master Mix1	
1st PCR reaction	1µl
Primer1	0.25µl
Primer2	0.25µl
Nuclease free water	9.5µl
Total:	10µl

Master Mix2	
High Fidelity buffer	5µl
dNTP 10mM	1µl
enzyme	0.5µl
nuclease free water	33.5µl
total:	40µl

The program used: TM02

1. 93^{0}C	1.00 min	7. 55^{0}C	0.30 min
2. 93^{0}C	0.15 min	8. 68^{0}C	2.00 min
3. 45^{0}C	0.30 min	9. GOTO 6	19 times
4. 68^{0}C	2.00 min	10. 68^{0}C	10.00 minutes
5. GOTO 2	9 times	11. 4^{0}C	forever
6. 93^{0}C	0.15 min	12. END	

3.2.3 The Gateway BP recombination – generation of entry clones

For the generation of promoter entry clones PCR products containing the amplified minimal promoters with B1 and B2 attachment sites and the donor vector pDONR-221-P1/P2 were used, for the enhancer entry clones PCR products containing the amplified enhancers with B3 and B5 attachment sites and the donor vector pDONR-221-P3/P5 were used. The recombination reaction contained an equal amount of 50 fentomoles of PCR products and donor vectors; the volume was adjusted with TE to 4µl, and finally 1µl of BP clonase II was added to the reaction. After overnight room temperature incubation 0.5µl Proteinase K was added to the reaction, and incubated for 10 minutes at 37^0C. The reaction was then transformed into competent bacteria.

3.2.4 The Gateway LR recombination – generation of expression vectors

Each entry clone and destination vector was used in an equal amount of 10 fentomoles, 1µl of 5x buffer was added to the mix, the volume was adjusted to 4µl with TE buffer, and finally 1µl of LR Clonase Plus was added to the reaction. After 18-20 hours of room temperature incubation 0.5µl Proteinase K was added to the reaction, and incubated for 10 minutes at 37^0C. The reaction was then transformed into competent bacteria.

3.2.5 Testing the colonies after transformation by colony PCR

For checking entry clones, M13 FP (TGTAAAACGACGGCCAGT) and RP (CAGGAAACAGCTATGACC) primers, for checking expression vectors, attB3 (GGGGACAAGTTTGTATAATAAAGTAGGCT) and Venus RP (TAGCTCAGGTAGTGGTTGTC) primers were used.

3.2.6 Primer sequences

Construct name	Forward primer sequence		Reverse Primer sequence
promoters			
ctr	AAGCTTCGTGTATTGTACGG		TATGTGTGTTATTTTTGTATAG
apoeb	TGGGATGACAAAAGACGA		CCCTTCTGTAATAAGAGGATGA
atp6v1g1	CTGTGAGTCTCGTGCAGTC		GCTTTGGTACGGATTTTATTT
gtf2a1	CAGCTGACTGCACGGTAAGA		CTCTTTACGGTCTTATTCACAGTCC
klf4	ACTACATCCCAAGCGTCAT		AGGTGTTTACTCTCATTCAGT
krt4	CAAGTGTGTGTGTGTGTGAGAG		CTGAGAAGGAGGTACGAGAGTG
ndr1	CTGACCATCAAAAGACTGCAAG		TCAAATCAAGGTAATAACCACACG
Pcpb2	CAGTGTGCAGTGTGGAGTACG		GGGGAAGAGGGAAGACACG
rdh10	CATAACAGGCGGACACAC		CCACGAAATCTGCCCAAA
tbp	AGTATGCGAGCCAATAGTGC		CTCCGTCTAGAAACAGTGTTAGATCA
tram1	GCTCTCTCGTCTCCTTGC		GTTCTGGATCACAAACTCATGG
c20orf45	GGGAGATTTTCCATTTAGATTGC		TTTAGAGTTTAACGGGCGACT
ccne	GTGCTTCGTTGTCATTCTAGGAG		AGTCTGTAAGCAGGCAGCAT
shha	GTTTTGTGGGATAACATCAGAAGTG		CGGAGGTTTGCGGCGGGGA
mef2d	CTTCCACACAGCAGTATCCATTCTA		GGTTATTATTTAGCCCGTACAGTCA
dre-mir9-1	TTGATCTAAATACAGTTGACTTTCTAA		GGATTCTTGTTACTTTCGGTTA
elp4	TCTCTTTCTGATTGGCTGAGATTAC		GCTGCGGGTTTTCTTCTGA
hsp70	TTGATTGGTCGAACATGCTG		CAGTCCGCTCGCTGTCTC
eng2b	TGAGAATAAGGCGAGGTTGG		TTCAGAATCAAAGCAGTAGACCTG
enhancers			
ctr	GTGTGTCATCCTCATCCACG		CATTCCATGATGGTGCTCTG
shha arc	AGCTTGACAACGGAGAGCAT		GAAACGCGCACATAAGGAAT
b-actin-i1	GCAGCCCTTCAAGTCTTTCATTT		GACAAAGGAAGTCCCTCTGCATT
pax6-eye	GCTGGCAAACACACTAACTTCACTT		TCATGTTTCTGTGTTTTGTTGCAGT
eng2b-CXE	TATCTTGTCCCCATTCCAACAGAG		ATGTCAGCCAGAATGGTCAAAAAC
eng2b-reg5	CGATACACTTTGATGATACGCATTG		GCTCACATGACATTTCTCATTTTCC
dre-mir9-1	ATTCCTTTCCTTGGCATCAA		GGGACACCGTTGTTCCTCTA
myl7	CCATCCTTTTCATCCCTCAA		AGCTTTGTCTACTCACCATGTTC
isl1	TCCAGCACCATAATTCACCA		CCAGTATCGTGCAGCCCTA
zCREST2			
dlx2b/6a ei	AATCAGAAAGCAAGGCAAAATTAG		TGTCATATAAACACACTGGCTGAA
mnx1-regB	ATGTGGAGGATCGGTGTCAT		CCGGTGACTTGTTGATTTCC
kdrl	CCGCGGTCACCTTCTGCTAGTTAAAACC		GCGGCCGCAATCCAAAGTAATTGATCCCTG
myf5	AAGACATAAAAACAGACATCCGAAG		GTTTGGTGTTGAAGGTTTCTGAGT
Gateway	**primers**		
attB1	GGGGACAAGTTTGTACAAAAAAGCAGGCT	attB2	GGGGACCACTTTGTACAAGAAAGCTGGGT
attB3	GGGGACAAGTTTGTATAATAAAGTAGGCT	attB5	GGGGACCACTTTGTATACAAAAGTTGGGT

3.3 DNA injection into zebrafish embryos

The circular plasmid DNA was injected to one-cell stage zebrafish embryos in a concentration of 5ng/µl, the linear DNA solutions were used in concentration of 25ng/µl, in 10x diluted Phenol red solution. The embryos were either dechorionated with 10mg/ml Pronase before injection, either at 24hpf stage. The fish embryos were kept in fish water containing 0.003% Phenylthiourea (PTU) at 28^0C, until they reached the proper developmental stage.

3.4 Fish husbandry and care

The adult zebrafish stocks were maintained in the fish facility of the ITG, in an aquarium system build by Aquarienbau Schwarz (Göttingen) in conditions referring to The Zebrafish Book. (Westerfield 1993). Approximately 15 pairs were kept in each tank (30 l) under the following water conditions: conductivity 400-500 µS; hardness 5° dH; pH 7,0-7,5 and temperature between 26 and 28°C. The light/dark cycle in the facility was set to 14 hours light and 10 hours dark. The fish were fed two times per day and the ammonium, nitrate, nitrite and phosphate levels are checked once per week to ensure a good water quality. Wild type zebrafish from the AB strain were used for the experiments. The crossing of fishes was performed in one liter crossing cages, filled with system water. One fish pair was put in every cage in the evening. To avoid parental cannibalism the cage contained a sieve, which separated the eggs from the parents after the laying. The laying started the next morning with the switching on of the facility light, which is one of the main breeding stimuli for the fishes. The eggs are collected shortly after using a small net, transferred to a Petri dish and used for experiments.

3.5 Staining methods

3.5.1 Detection of the fluorescent proteins

The Venus YFP and the GFP were detected under an epifluorescent microscope with the proper filter under UV light.

3.5.2 X-Gal staining

The embryos were fixed at room temperature in BT-Fix (4% paraformaldehyde, 4% sucrose, 0.12mM $CaCl_2$, 0.1M NaPi pH 7.4) for 2-4 hours in 24-well plates, then washed 3 times with 0.02% NP40 containing 1x PBS, once with staining buffer (0.15M NaCl, 3mM $K_4Fe_3(CN_6)_6$, 3mM $K_3Fe_4(CN_6)_6$, 0.02M NaPi, pH 7.4) then stained with 1 ml staining buffer containing 5µl 8% X-Gal in DMSO. After the staining was complete, the embryos were washed 3 times with 0.02% NP40-PBS, and finally were fixed with BT-Fix.

3.5.3 In situ hybridization

The embryos were fixed at the appropriate stage in BT-Fix in 24-well plates overnight at 4^0C, then the fixative was changed to 100% methanol, and the plates were kept at -20^0C for at least 2 hours. Then the embryos were rehydrated with descendent alcohol series: a 5-minute wash with 75% methanol in PTW (1xPBS, 0.1% Tween 20), a 5-minute wash with 50% methanol in PTW and finally a 5-minute wash with 25% methanol in PTW, 4 times 5-minute wash with PTW, at room temperature. Then the embryos were treated with Proteinase K (in a 10µl/ml final concentration) for 1 minute, followed by a 20-minute fixation with BT-Fix, and 5 times wash with PTW, at room temperature.

Then the embryos were incubated in hybridisation buffer (HYB: 50% Formamide, 5x SSC, 0.5 mg/ml yeast RNA, 50 µg/ml heparin, 0.1 % Tween 20.9 mM citric acid).at 65^0C for at least 3 hours for pre-hybridization, followed by the hybridization step with the HYB-buffer containing the RNA-probe in a 1:500 dilution, overnight at 65^0C.

The next day the embryos were washed with the SSC-buffers at 65^0C: 2x 30-minute wash with 50% formamide/50% 2xSSC, 0.1% Tween 20; 1x15 min in 2x SSC, 0.1% Tween 20; 2x30 min in 0.2x SSC, 0.1% Tween 20 and 1x 5min blocking buffer(1x PBS, 0.1% Tween 20, 5% sheep serum, 0.2% BSA, 1% DMSO). After the washes the embryos were incubated at room temperature with blocking buffer for at least 2 hours, then the embryos were incubated with the anti- dioxigenin alkaline phosphatase Fab fragments ON at 4^0C in a 1:4000 dilution

The next day the embryos were washed 6 times with PTW for 20 minutes, once with staining buffer (100 mM Tris-HCl pH 7.9, 100mM NaCl, 0.1% Tween 20, 50mM MgCl2) for 5 minutes, at room temperature. The bound antibody was revealed by adding the substrates, NBT and BCIP (0.34mg/ml and 0.175 mg/ml). Reaction was stopped by repeated rinses in PTW followed by BT-Fix.

3.6 High throughput screening

3.6.1 Automated microscopy

The microscopy was performed in the laboratory and with the help of Urban Liebel (ITG, FZK, Karlsruhe, Germany). Imaging of the 96 well plates was done on a "Scan^R" high content screening microscope (Olympus Biosystems, Munich, Germany) with a SWAP plate gripper (Hamilton, Switzerland), a 2x objective (Plan-Apo, Olympus, Germany) and an Olympus Biosystems DB-1 (1300x1024 pixels) camera in bright field and with CFP, YFP filter cubes. Image integration times were fixed (180ms CFP and 1000ms for YFP). Central focal plane of the embryo was detected by an object detection auto-focus algorithm. Each embryo was acquired with four z-slices (55µm) and projection was performed. Light source was an ultra stable MT-20 (Olympus Biosystems, Munich, Germany) with a xenon lamp. Data management, thumbnail gallery generation, data compression was carried out via an assembly of LabView software modules (National Instruments, Germany).

3.6.2 Embryo referencing, overlay of experiments and normalisation

All the algorithms were written and performed by Markus Reischl (IAI, FZK, Karlsruhe, Germany).

Images were processed by an automated computer vision routine using bright field images (showing a high contrast and small structures) and CFP images (to find outlines and to remove dirt objects) to gather embryo morphology and reporter expression information (Detailed description of all algorithms are described in Reischl et al., manuscript in preparation). Errors caused by noise, dirt or malformed embryos were suppressed. The exact outline of the embryo was extracted from the inverted monochrome CFP image using histograms, dynamic thresholds, opening algorithms, binary largest object algorithms, hole filling algorithms and low pass filtering. The alignment of the embryos was identified by a regression routine. To save computation time, the images were cropped to contain only the embryo. The cropped aligned image was checked for errors of detection (e.g. empty wells, embryos out of focus, malformed embryos etc.), and images with gross errors were discarded. Embryos not reaching a minimum level of CFP activity (uninjected) and top 5% of embryos having the highest *venus* expression were excluded from quantification analysis.

A final image was generated from the 4 YFP and CFP aligned images taken at different planes per embryo by an extended depth of field algorithm using a 5-fold wavelet analysis (Daubechies wavelets). Furthermore, background yolk fluorescence and saturated areas were deduced in the YFP images. To handle contrast differences and noise effects a robust mathematical model was fitted to each embryo by a new model based regression analysis. A model function for the ventral and the dorsal side was introduced. Parameters for the model functions were adapted to an outline resulting from a greyscale threshold obtained by a histogram of the CFP image. Similar algorithms were applied to the bright field image to detect embryo domains. Each model function was used to define characteristic landmarks and coordinate systems within the embryo image.

A data mining routine evaluated validity and reproducibility of all experiments to allow merging of repetitions. To gain reliable results, the minimum number of embryos for the analysis of a construct was set to 25, which provided a reliable indicator of domain specificity (effect of number of analysed embryos on the rate of error is described in Reischl et al., in preparation).

3.6.3 Definition of domains of the prim 16 stage zebrafish embryo:

Based on the expression patterns of the chosen enhancers (summarized in Table 2) the following tissues – anatomical regions were chosen for the quantitative analysis: eye, brain, cerebellum, heart, notochord, and spinal chord. After injection of distinct constructs high levels of Venus expression was observed in the yolk, the yolk plug and the skin, so these tissues were chosen for the signal quantification as well, and finally the signal coming from the yolk structures were removed from the overlays. Domains of the embryo were arbitrarily defined to include but not restricted to characteristic features/ tissues of the zebrafish embryo as below. **Domain of the heart**: the region containing the heart was defined as the curved keel shape territory bordered by the ventral brain dorsally and the yolk cell ventrally. The anterior boundary was arbitrarily set as the line between the anterior tip of the brain (olfactory placode) and the ventral joint between the yolk ball and yolk extension. The posterior boundary is defined by the line between the posterior end of the retina and the ventral joint between the yolk ball and yolk extension. **Domain of the yolk**: this domain includes the yolk ball and the yolk extension. **Spinal cord domain**: the anterior boundary of the domain containing the spinal cord was defined at 2 OVL (otic vesicle

length) posterior to the otic vesicle itself. The territory is posterior to this position, above the notochord and beneath the skin domain (also including somites). **Brain domain**: the domain is defined as the region anterior to the spinal cord excluding the notochord, the MHB and the retina. **Notochord domain**: this domain is defined by the notochord and starts anteriorly below the anterior end of the otic vesicle. **Domain of the midbrain hindbrain boundary (MHB)**: this domain contains the MHB bordered by the two prominent vertical furrows anteriorly and posteriorly of the MHB, respectively. Thus, it includes tissue of the posterior part of the tectum and tegmentum, cerebellar tissue and tissue of rhombomere 1. Ventrally the domain extents to the floorplate. **Domain of the eye**: this domain contains the retina region including the lens placode and tissues between the eyes. **Domain of the skin**: this domain contains a stripe of a single cell width at the outline of the embryo (15 µm). The defined region contains mainly skin cells of the midsection and overlaps partially with the median fin fold.

4) Results and discussion

4.1 Evolutionary conserved regions around the pax2 locus show differential enhancer activity with different promoter constructs

Pax2 is a transcription factor involved in the development of the midbrain hindbrain boundary (MHB) organizer and specification of neuronal cell fates in vertebrates (Nornes et al. 1990). It is required for establishment of *eng2* and *eng3* gene expression in the midbrain and MHB primordium during late gastrulation (Song et al. 1996; Lun et al. 1998), and plays role in kidney (Dressler et al. 1990), ear and eye (Nornes et al. 1990) development. In zebrafish and in fugu two orthologues are present, *pax2a* and *b*. In zebrafish they share 93% identity at amino acid sequence level, but the two genes are completely different in their 3' and 5' non-coding sequences (Pfeffer et al. 1998). *pax2a* expression is initiated at 8-9 hours post fertilization (hpf) in zebrafish embryos in two lateral stripes of the anterior neural plate. At 24hpf stage the expression is detected in the ventral retina and optic stalk, in the otic vesicle, in specific neurons of the hindbrain and spinal cord, in the pronephric duct and in the proctodeum (Krauss et al. 1991) (Figure 5.). *pax2b* is expressed in all these domains except in the pronephros, and differs from *pax2a* in the temporal onsets and transcription level at the otic region (Pfeffer et al. 1998; Picker A et al. 2002).

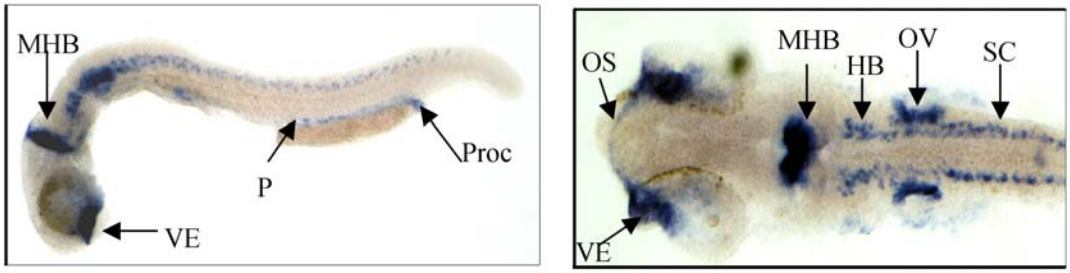


Figure 5: In situ hybridization of wild type zebrafish embryo with *pax2a* probe
The expression is visible in the ventral eye (VE), optic stalk (OS), midbrain-hindbrain boundary (MHB), hindbrain (HB) and spinal cord (SC) neurons, otic vesicle (OV), pronephros (P) and proctodeum (Proc).

Although several *pax2* enhancers have been identified, its cis-regulatory grammar is still not fully understood. For example, the elements that regulate the expression in the hindbrain and the MHB are clearly characterized: a 120 bp early

enhancer (at –3.7 kb position from the TSS of the mouse *pax2*) activates *pax2* in the neural plate of late gastrula embryos, while *pax2* transcription is subsequently maintained at the MHB by a 410 bp late enhancer at –2.8 kb (Pfeffer et al. 2002). Nevertheless, up to date there is only one report mentioning a regulatory region that is responsible for the eye expression. This not further characterized *pax2* optic stalk enhancer is located within a 9kb region upstream to the mouse *pax2* TSS (Schwarz et al. 2000).

4.1.1 Identification of conserved non-coding sequences in the *pax2* locus

Our major aim was to identify those enhancer elements, which drive the expression of *pax2* into the developing vertebrate eye. As the proteins and their expression patterns are highly conserved between mammals and fish, and as several already described enhancers were shown to be conserved between human, mouse and fugu (Pfeffer et al. 2002), Sandro Banfi et al. (TIGEM, Naples, Italy) performed phylogenetic footprinting using the fugu, mouse and human genomic DNA around the orthologous *pax2* regions by using Vista (Mayor et al. 2000). They have chosen those regions, which shared more than 75% homology between the human and fugu sequences. They named the identified conserved non-coding sequences as *CSTs*, and numbered them related to their genomic position (Table 4., Figure 6.).

CST	Associated with		Amplified from	Length	Distance from TSS (in bp)	
	pax2a	*pax2b*		(in bp)	in human	in fugu
1	+	+	*pax2b*	175	135849	59378
2	+	+	*pax2b*	483	134988	58799
3	+	+	*pax2b*	203	133424	58164
4	+	+	*pax2b*	168	126044	51446
5	-	+	*pax2b*	172	102954	39736
6	-	+	*pax2b*	228	99116	35755
7	+	+	*pax2b*	339	94367	31521
8	+	+	*pax2b*	528	93521	30990
10	-	+	*pax2b*	166	92332	30052
11	-	+	*pax2b*	150	85585	23934
12	-	+	*pax2b*	223	80532	22493
13	-	+	*pax2b*	195	78033	20634
14	+	+	*pax2b*	724	61235	16027
17	+	+	*pax2b*	182	7527	1915
18	+	-	*pax2a*	172	120609	40062
19	+	-	*pax2a*	224	89269	24050
20	+	-	*pax2a*	124	36196	7178
21	+	-	*pax2a*	153	35098	5588

Table 4: Properties of the conserved non-coding sequences analysed in the screen

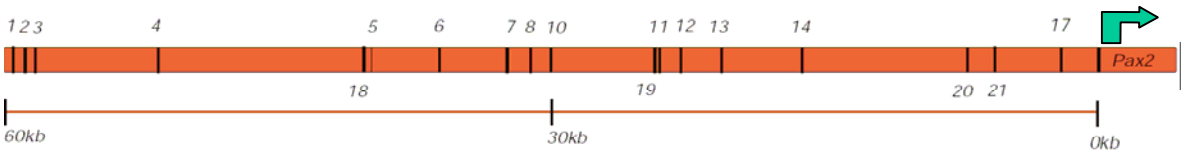


Figure 6: The genomic position of the tested *CSTs* amplified from fugu *pax2b* (numbers on top) and *pax2a* (numbers on the bottom) genomic regions.

They amplified the corresponding fugu sequences, containing the conserved regions plus several tens of base pairs of flanking DNA. I tested these PCR products for enhancer activity with different promoter-reporter constructs in developing zebrafish embryos.

4.1.2 CSTs show enhancer activity when co-injected with the endogenous promoter

First I used an endogenous promoter construct, the 5.3kb zebrafish *pax2a* promoter followed by a *gfp* tag (Picker A et al. 2002) for the enhancer-assays. This promoter drives expression of the reporter to the MHB, otic placode, hindbrain, spinal cord and pronephros (Picker A et al. 2002). To verify the expression gained with the promoter construct, I injected the isolated linearized fragment to one-cell stage zebrafish embryos, and detected the fluorescence in a concentration dependent manner at 24hpf stage: in the forebrain, hindbrain, at the MHB, in the spinal cord, pronephric duct and ventral mesoderm (50ng/µl), or in the hindbrain, forebrain and MHB (10 ng/µl).

I performed co-injections of the PCR fragments using the lower concentration of the promoter construct, fixed the embryos at 24hpf stage, and performed in situ hybridization (ISH) with a *gfp*-specific probe. The staining provided the possibility of detailed expression domain analysis. I counted the *gfp*-expressing cells, and collected the staining patterns into composite expression maps from head preparations, because the optic stalk is only visible from dorsal view. In this set of co-injection experiments *CST4*, *7*, *10*, *14* and *18* resulted in the enhanced appearance of the *gfp* transcript in the ventral retina and/or optic stalk (Figure 8. and 10.A), while *CST2*, *17* and *18* turned be a kidney enhancer, and co-injection of *CST10* and *20* resulted in expression of the reporter in the otic vesicle (Table 5.).

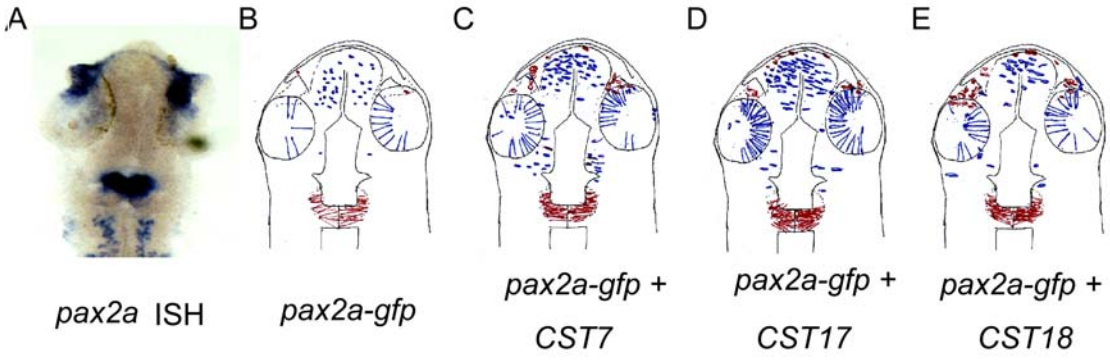


Figure 8: Expression maps from the head preparations

A: A head preparation of an embryo after ISH with *pax2a*-specific probe. B-E: Expression maps collected from roughly 30 embryos. The CSTs were co-injected with the 5.3kb *pax2a-gfp* construct, and the embryos were subject to ISH with a *gfp*-specific probe. B: *pax2a-gfp* promoter only, C: *pax2a-gfp* co-injected with *CST7*, D: *pax2a-gfp* co-injected with *CST17*, E: *pax2a-gfp* co-injected with *CST18*.

4.1.3 Different results gained when *hsp68* minimal promoter was used

To test whether the *pax2a* promoter is required for the activity of the tested conserved non-coding sequences, and to get rid of the basal eye-activity of the promoter construct detected by ISH, the whole set of CSTs were co-injected with a heterogonous promoter as well. The minimal promoter of the mouse heat shock protein 68 (*hsp68)* (Kothary et al. 1989) was chosen, as it has been shown to have weak basal activity, but its expression was enhanced by tissue-specific enhancers in transgenic mice (Tuggle et al. 1990) and in zebrafish as well (Muller et al. 1999). After co-injections of the CST fragments with the linearized *hsp68-lacZ* construct I mildly fixed the 24hpf stage embryos, checked for LacZ activity, and took the expression maps (Figure 10.). As a positive control I used the well characterized notochord enhancer, the *sonic hedgehog activation region C (shh arC*) (Muller et al. 1999).

The *hsp68* promoter itself hardly turned on the reporter gene expression, while *shh arC* gave enhanced expression in the notochord, as expected. Several *CSTs* showed enhancer activity in *pax2*-specific domains: like *CST7*, *8*, *10*, *14* and *18* directed the *lacZ* expression to the eye and the ventral retina (Figure 10.B), *CST8* and *17* showed activity in the developing kidney, while the co-injection of *CST6, 8, 14, 17* and *18* resulted in reporter expression in the spinal cord (Figure 9., Table 5.).

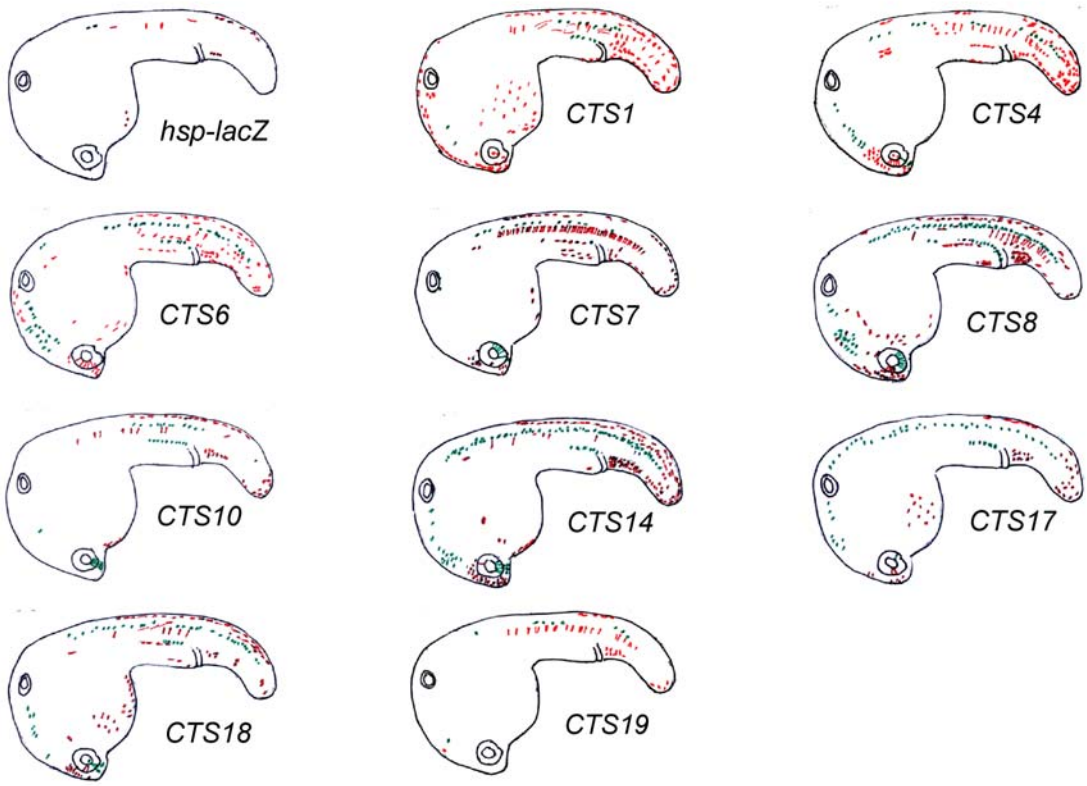


Figure 9: Expression maps of the *CSTs* co-injected with *hsp68* promoter.
The expression maps were collected from approximately 50 embryos after X-Gal staining. The green dots represent *pax2a*-specific expression patterns, the red ones represent ectopic activation of the *lacZ* reporter.

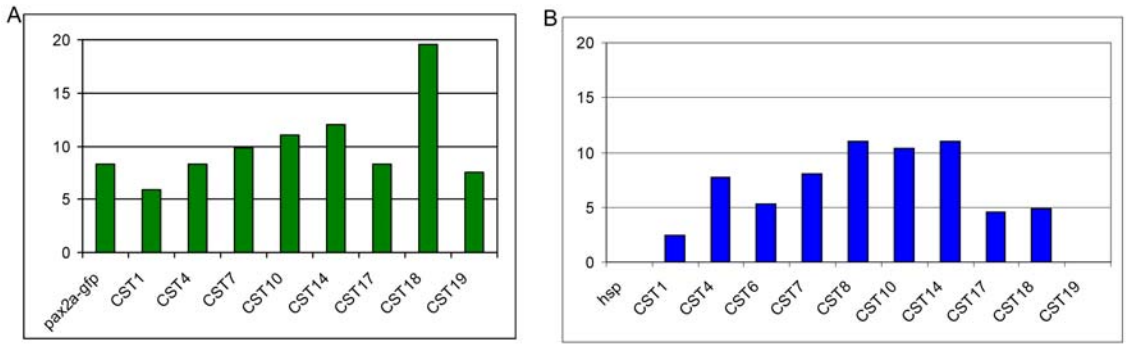


Figure 10: The number of reporter-expressing retina and optic stalk cells in the co-injected embryos, normalized by the total embryo number. A: for the CSTs co-injected with the 5.3kb *pax2a* promoter, B: for the CSTs co-injected with the *hsp68* promoter.

The results gained with the two promoters are partially overlapping. The number of expressing cells normalized by the expressing embryo number in the two co-injection systems is comparable. The major "outlier" is the *CST18*, it activated the expression of the *pax2a* promoter-driven reporter in twice as much cells in the optic stalk and retina, compared to other CSTs and the control, and four times more compared to the *CST18* co-injected with the *hsp68* promoter (Figure 10.). Some elements showed eye and optic stalk expression only when co-injected with one promoter, like the *CST4* gave retina expression with the *pax2a-gfp* construct, but not with the other, or the *CST8* and *CST20* vice versa (Table 5.).

	MHB	HB	SCN	PD	OV	OS, R
pax2	+	+	+	-	-	-
ArC	-	-	-	-	-	-
1	-	-	-	-	-	-
2	-	-	-	+	-	-
3	-	-	-	-	-	-
4	+	+	-	-	-	+
5	-	-	-	-	-	-
6	-	-	-	-	-	-
7	+	-	-	-	-	+
8	-	-	-	-	-	-
10	-	+	+	-	+	+
11	-	-	-	-	-	-
12	-	-	-	-	-	-
13	-	-	-	-	-	-
14	-	+	+	-	-	+
17	-	+	+	+	-	-
18	-	+	+	+	-	+
19	+	+	-	-	-	-
20	+	+	-	-	+	-
21	-	-	-	-	-	-

	MHB	HB	SCN	PD	OV	OS, R
hsp	-	-	-	-	-	-
ArC	-	-	-	-	-	-
1	-	-	-	-	-	-
2	-	-	-	-	-	-
3	-	-	-	-	-	-
4	+	+	-	-	-	-
5	-	-	-	-	-	-
6	+	+	+	-	-	-
7	-	-	-	-	-	+
8	-	-	+	+	-	+
10	-	-	-	-	-	+
11	-	-	-	-	-	-
12	-	-	-	-	-	-
13	-	-	-	-	-	-
14	+	+	+	-	-	+
17	-	+	+	+	-	-
18	-	+	+	-	-	+
19	+	+	-	-	-	-
20	+	+	-	-	-	+
21	-	-	-	-	-	-

Table 5: Summary of the enhancer activities of the CSTs tested with the two promoters

A: Summary of the co-injections performed with the *pax2a-gfp* promoter, B: summary of the co-injections performed with the *hsp-lacZ* promoter. Elements highlighted with light colours showed *pax2*-specific enhancer activity, while the ones highlighted with dark colours enhanced the reporter activity in the optic stalk and/or retina. MHB: midbrain-hindbrain boundary, HB: hindbrain, SCN: spinal cord neurons, PD: pronephric duct, OV: otic vesicle, OS: optic stalk, R: ventral retina.

When comparing other *pax2*-specific expression domains, more differences are visible between the two data series (Table 5.). While *CST4*, *19* and *20* were able to activate both promoters in the midbrain hindbrain boundary and in the hindbrain, *CST7* was MHB-specific with the endogenous promoter, and *CST6* and *14* showed MHB- and hindbrain-expression with the *hsp68-lacZ* construct. The *CST14, 17* and *18* were driving the reporter expression to the hindbrain and spinal cord with both promoters, only the *pax2a* promoter was activated by the *CST10,* while the *hsp68* promoter with the *CST6* and *8* in the spinal cord. When co-injected with the *pax2a* promoter, *CST2, 17* and *18* could activate pronephric expression, while the *hsp68-lacZ* construct was activated by the *CST8* and *17* in the embryonic kidney. No otic vesicle expression could be observed when the CSTs were co-injected with the *hsp68* promoter, while the *CST10* and *CST20* was able to activate the *pax2a* promoter in the developing ear.

The fact that *CST17* showed hindbrain, spinal cord and pronephros enhancer activity with the *hsp68* promoter confirms the co-injection experiments, as the *pax2a* promoter itself, which contains the CST17, has the ability to drive the expression into these domains.

4.1.4 Verification of the co-injection by injecting covalently joint fragments

Co-injection of linear DNA fragments, like cis-regulatory elements and promoters followed by a reporter gene is a fast method for testing putative enhancers. It does not require the generation of expression construct, the amplified sequences can be simply mixed with the linearized promoter construct in a reaction tube (Muller et al. 1997). To verify the specificity of the expression domains gained by co-injection, those fragments, which showed enhancer activity in ventral retina and optic stalk in both reporter systems, namely *CST7, 10, 14* and *18* were cloned in front of the *hsp-lacZ* cassette, and I injected these as circular plasmids into zebrafish embryos. The low-concentration plasmid injections (Uemura et al. 2005) resulted in less background in non-related tissues, such as muscle and notochord, compared to the co-injections, but the specificity of eye-expression was indifferent in all cases (Figure 11.). The major difference observed was that the co-injected fragments turned on the reporter expression in the spinal cord neurons, while the covalently joint fragments did not.

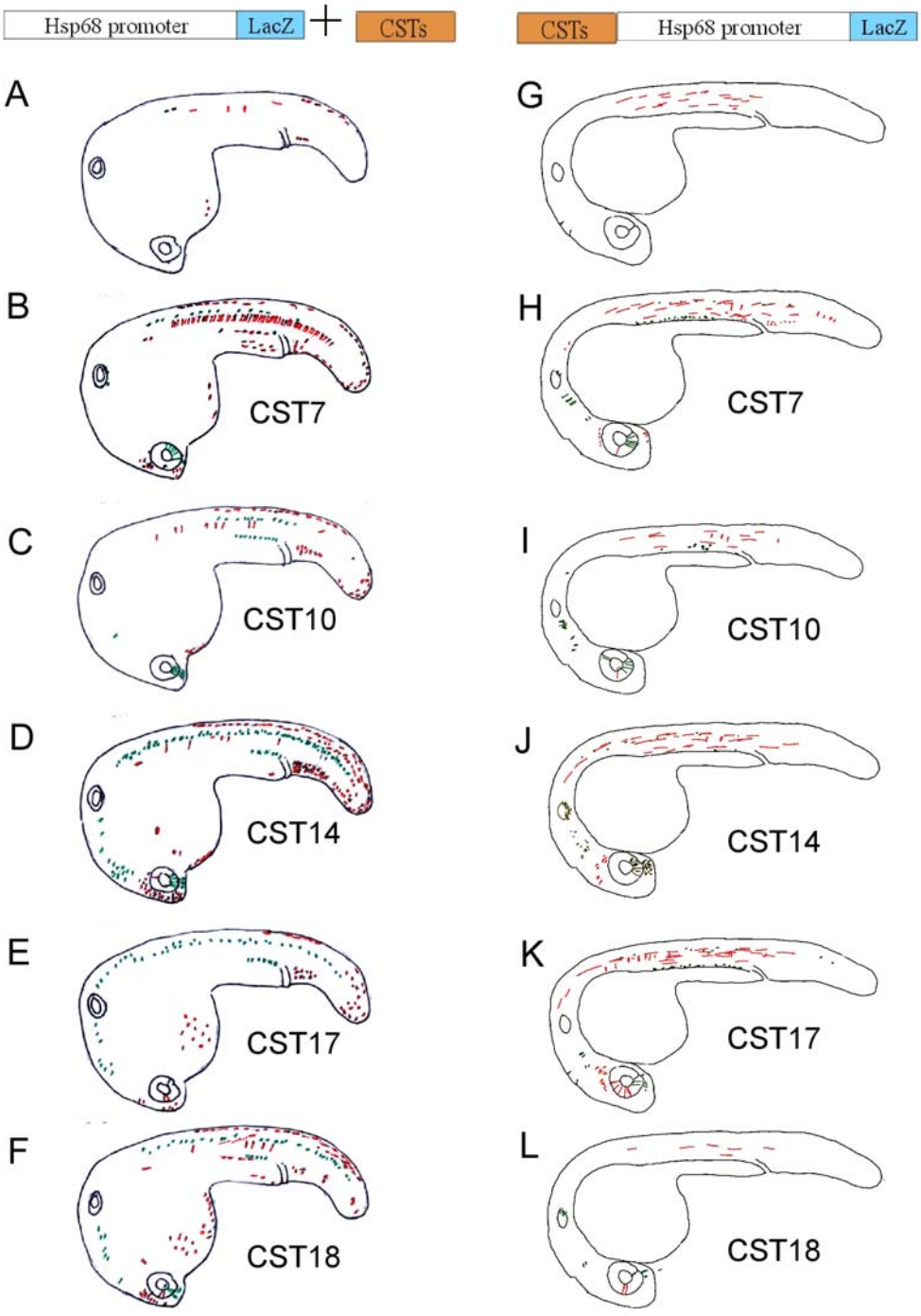


Figure 11: Verification of the co-injection with covalently linked enhancers

Expression maps from roughly 50 embryos. The embryos were co-injected with the *hsp* promoter and the *CSTs* (A-F) or injected with the plasmids containing the CSTs in front of the promoter construct (G-L), then stained with X-Gal. A, G: expression map of the *hsp68-lacZ* promoter. B, H: *CST7*, C, I: *CST10*, D, J: *CST14*, E, K: *CST17*, F, L: *CST18*. The green dots represent *pax2*-specific expression patterns, the red ones represent ectopic activation of the *lacZ* reporter.

4.1.5 Discussion

Conserved non-coding sequences upstream from pax2 genes with enhancer activity

Pax2, together with Pax6, plays an important regulatory role in vertebrate eye development during embryogenesis. *pax2* is expressed in the ventral half of the optic vesicle during early eye morphogenesis (Nornes et al. 1990; Torres et al. 1996), and shortly after the invagination of the optic cup it becomes confined to the optic stalk (Nornes et al. 1990; Torres et al. 1996). The developing optic cup/optic stalk border is marked by overlapping *pax2* and *pax6* expression domains (Nornes et al. 1990; Walther et al. 1991). The *pax6*-expressing pigmented epithelium of the retina has been shown to expand in the *pax2* mutant embryos, invading the optic cup/optic stalk boundary (Torres et al. 1996). Moreover, it was demonstrated that Pax6 was sufficient to repress transcription of a reporter gene driven by *pax2* enhancer sequences and vice versa (Schwarz et al. 2000). Despite the fact that only the ventral retina is *pax2*-specific, I collected the expression information from the whole retina for the cell-counts,due to the potential shift of the expression domains upon injection of promoter and enhancer fragments of the *pax2* gene, potentially disturbing the physiological cis-regulatory element – transcription factor ratio.

Conserved non-coding sequences around the *pax2* loci showed enhancer activity in transient zebrafish tests. Numerous amplified fugu fragments containing the conserved sequences were able to drive expression of reporter genes into *pax2*-specific domains upon interaction with the endogenous zebrafish *pax2a* and the heterologue mouse *hsp68* promoter. I demonstrated that four conserved non-coding sequences (*CST7*, *10*, *14* and *18)* could drive the reporter expression into the developing eye and optic stalk of zebrafish embryos with two different promoter constructs.

Lessons from methodology

I used different techniques to detect the expression of the transient transgene during the enhancer assays. First I detected the fluorescent protein (expressed from the *pax2a-gfp* construct) under epifluorescent microscope, but checking GFP expression at a defined stage in large number of living embryos was not feasible. Thus I fixed the injected embryos at 24hpf stage, and performed in situ hybridisation. This staining is a sensitive method of labelling the transcript of the reporter. This

could be the reason why I could not observe fluorescent signal in the retina or optic stalk of the embryos injected with low plasmid concentration, while the ISH stained embryos showed some background retina expression, even when the same plasmid concentration was used.

As the *hsp68* promoter construct by itself did not show any significant expression in any *pax2*-specific domains, the appearance of signal in the optic stalk or in the retina upon co-injection of fragments clearly indicated eye-specific enhancer activity. The *hsp68* promoter is cloned in front of the *lacZ* cassette, therefore I performed X-Gal staining to detect the enhancer activity of the co-injected fragments. This staining methods is based on enzymatic reaction of the expressed protein, thus gives information about the expression on the protein level. The X-Gal staining is even faster and less laborious then the ISH - the fixation and the following staining is done in one day, moreover it gives the possibility of later ISH over the stained embryos

Verification of the co-injections by covalently cloned fragments

The co-injection experiments with the *hsp68* promoter were verified by constructs where fragments were covalently cloned in front of the promoter. The consequent enhancer activity of *CST7*, *10*, *14* and *18* in the retina and/or optic stalk in these experiments indicate that co-injection of isolated linear DNA sequences is a reliable methods for fast enhancer assay.

Partially different enhancer activity is observed with different promoters

The majority of the CSTs behaved differentially when co-injected with the two promoter constructs. The partially overlapping results (summarized in Table 5.) can arise from different sources. First of all, the 5.3kb zebrafish *pax2a* promoter is not a basal promoter, it contains enhancer elements that drive expression of *pax2* into the MHB, hindbrain, otic placode, spinal cord and pronephros (Picker A et al. 2002), and the conserved sequences may cooperate with these other enhancers. Second, several enhancers were shown to be promoter-specific, including the *pax2* early MHB enhancer (Picker A et al. 2002), so the enhancers may only interact with the endogenous, but not with the *hsp68* promoter. To rule out the first potential cause, extra experiments with the *pax2a* core promoter would be needed to perform.

4.2 Shuffled conserved sequences show enhancer activity, even if not related to transcription factor or developmental regulator genes

Comparative analysis of the mammalian and fish genomes revealed in conserved elements, which shared extremely high degree of homology (Bejerano et al. 2004; Iwama et al. 2004; Plessy et al. 2005; Woolfe et al. 2005). All these studies concluded that conserved non-protein coding sequences working as enhancers were significantly enriched in or around developmental regulators and/or transcription factor genes. But is this observation restricted to developmental genes?

In collaboration with Remo Sanges and Elia Stupka (CBM, AREA Science Park, Basovizza, Trieste, Italy) we were focusing on the extent, mobility and function of conserved non-coding elements across vertebrate orthologous loci. The collaborating partners developed a new tool to identify regionally conserved elements (rCNEs), which were not exclusively associated with genes falling into developmental or transcription factor GO categories. They extensively analyzed these elements, and later I tested a subset of these for enhancer activity in zebrafish embryos.

One of the major drawbacks of current comparative studies is that they rely on methods for local alignment, such as BLAST (Altschul et al. 1990) and FASTA (Pearson et al. 1988), which were developed when the majority of available sequences to be aligned were coding. It has been shown that such algorithms are not as efficient in aligning non-coding sequences (Bergman et al. 2001). Recently two approaches have been published which provide novel alignment strategies: the promoter-wise algorithm coupled with “evolutionary selex” (Ettwiller et al. 2005), and the CHAOS alignment program (Brudno et al. 2003a). Unlike other fast algorithms for genomic alignment, CHAOS does not depend on long exact matches, it does not require extensive ungapped homology, and it does allow for mismatches within alignment seeds, all of which are important when comparing non-coding regions across distantly related organisms. Thus in this project, CHAOS was used for the identification of short conserved regions that have changed their location during vertebrate evolution.

4.2.1 The identification and computational analysis of regionally conserved elements

All genes, for which there were predicted orthologs within the Ensemble (Birney et al. 2006) in the mouse, human and either in the rat or dog genome, were subject to the analysis, 9.749 gene groups in total. For each group of orthologous genes global multiple alignment was performed using MLAGAN (Brudno et al. 2003b). For each locus, the whole repeat-masked sequence was used, containing the transcriptional unit as well as the flanking sequences up to the preceding and following gene. The alignments were parsed using VISTA (Mayor et al. 2000), searching for segments of minimum 100 bp length and 70% identity. From the gained dataset, those regions were chosen, which were found at least in mouse, human, and a third mammalian species, and which overlapped by at least 50bp. 77,3% of the total 364.358 rCNEs were shown to be non-genic. These non-coding conserved elements were further annotated based upon their position in the mouse genome with respect to the gene locus to define categories of pre-gene, intronic or post-gene elements.

Conservation of the rCNEs were found in teleost genomes using CHAOS alignment tool, with the criteria of at least 60% identity over a minimum length of 40bp (the mouse rCNE sequences were used as baselines). Regions of the mouse genome that were conserved at least in the fugu orthologous loci were termed as SCEs (shuffled conserved elements). The analysis identified 21.427 non-redundant non-genic SCEs, which were found in about 30% of the genes analyzed.

The SCEs, which had a median length of 45bp and a median percentage identity of 67%, were investigated if they have shuffled in terms of position and orientation relative to the transcriptional unit. The results of this revealed that only 28% of the elements identified have retained the same orientation and same position (labelled as “collinear”), whereas others have been shifted in terms of orientation (labelled as “reversed”), position (“moved”), or both (“moved-reversed”) (Figure 12.).

The genes associated with the SCEs were analyzed in terms of gene ontology (GO). Although there is a significant over-representation of gene classes of transcription factors and developmental regulators, there are other GO categories that are significantly under-represented in other studies (Woolfe et al. 2005). Most strikingly, there is 54-fold enrichment in genes belonging to the extracellular regions that contain SCEs, and SCEs were identified in 40% of genes assigned to the

behavioural GO class. These results show, that this type of analysis found not only higher number of conserved non-coding elements, but elements, which are assigned to different types of genes as well. Finally the shuffling properties of the SCEs were analyzed in relation to their distance from the transcriptional unit. Collinear elements are distributed significantly closer to the start and the end of the transcriptional unit compared with non-collinear (moved, reversed and moved-reversed) elements. The higher resolution analysis of the regions poor in shuffling revealed that proximal promoter regions (approximately 1000bp upstream of the TSS) contained significantly less shuffled elements.

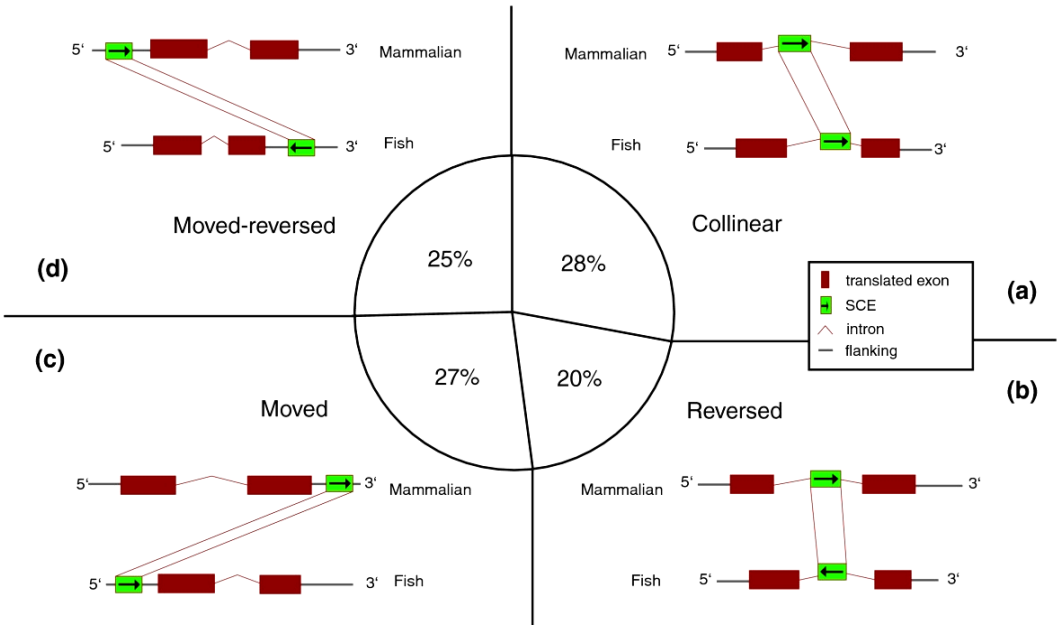


Figure 12: Shuffling categories of SCEs

4.2.2 Analysis of shuffled conserved elements for enhancer activity

To verify the ability of SCEs to predict functional enhancer elements, an overlap analysis was performed with 98 already published mouse enhancers present in Genbank. Compared to the previous datasets of conserved non-coding elements (one) and ultraconserved elements (two), the SCEs contained eighteen of these already known enhancers.

The next step was the in vivo analysis of several shuffled conserved elements. Twenty-eight SCEs were amplified from the fugu genome and I co-injected these fragments into one-cell stage zebrafish embryos together with the isolated fragment of the minimal mouse *hsp68* promoter construct (*hsp68-lacZ*). Four out of these elements overlapped with known mouse enhancers, the activity of these in zebrafish was not previously reported. The remaining 23 elements were assigned to twelve genes; four of them were not belonging to transcription factor or developmental regulator GO categories. As an enhancer control, the already well described *sonic hedgehog activation region C (shh arC)* was used (Muller et al. 1999). As a negative control set twelve non-coding, non-repeated and non-conserved fragments were chosen for co-injections, of which nine were from the same genomic regions as the SCEs, three were from random genes. Table 6. summarises the properties of the SCEs and controls used in the test.

I co-injected all fragments (SCEs and controls as well) at least three times, collected the embryos at 24hpf stage, mildly fixed in BT-Fix, and performed X-Gal staining. I counted the LacZ positive cells per each embryo and determined the tissues where the stained cells appeared. Due to the chosen concentrations of the co-injected DNA molecules, the tested fragments induced the *lacZ* expression only in few cells in the co-injected embryos (Figure 13.), but this setting provided us great specificity. I plotted the X-Gal stained cells into composite expression maps representing approximately 130 embryos per SCEs.

The gained expression patterns were compared with expression data retrieved from the Zebrafish Information Network (http://zfin.org). Based upon the expression domains of the assigned genes, I determined the *lacZ* expression in the following tissues or embryo regions: muscle, notochord, central nervous system (CNS), eye, ear and blood vessels. The YSL was excluded from the cell-counts.

Gene	SCE ID	enhancer effect	Length of element (bp)	Length of PCR (bp)	Mouse region	Fugu region	injected embryos	expressing embryos	% of Embryos Expressing LacZ
no element	-	-	N/A	N/A			161	97	58,2
Shh	**ArC**	+	N/A	500	N/A	N/A	96	87	90,6
	12058	-	45	749	INTRON	INTRON	139	88	62,3
Otx2	**2894**	+	51	527	POST_END	PRE_GENE	111	83	74,7
Gata3	**2755**	-	40	380	POST_END	INTRON1	107	79	73,8
Ets	**1645**	+	40	522	PRE_TSS	POST_GENE	105	66	62,9
	1646	+	46	715	PRE_TSS	INTRON1	133	82	61,7
	1652	+	41	668	INTRON1	POST_GENE	159	93	58,5
	1653	+	48	695	INTRON1	INTRON2	149	80	53,7
Pax2b	**333**	+	39	543	PRE_GENE	PRE_GENE	149	86	57,7
Pax6a	**1194**	+	33	265	INTRON6	PRE_GENE	133	93	69,9
Pax3	**2598**	-	42	670	PRE_TSS	POST_GENE	124	68	54,8
1300007F04Rik*	**44**	-	42	710	POST_END	POST_GENE	107	55	51,4
Zfpm2	**691**	+	48	562	PRE_TSS	POST_GENE	140	94	67,1
	692	+	48	549	PRE_TSS	INTRON2	131	113	86,3
Tmeff2*	**1050**	+	48	771	INTRON4	INTRON4	164	127	77,4
	1051	+	38	648	INTRON4	PRE_GENE	120	108	90
	1052	+	51	570	INTRON4	PRE_GENE	109	83	76,1
Jag1b	**3120**	-	37	453	PRE_TSS	PRE_GENE	136	105	77,2
	3121	+	55	278	PRE_TSS	PRE_GENE	142	91	64,1
	3122	-	44	543	PRE_TSS	PRE_GENE	106	80	75,5
Mapkap1	**1972**	+	37	648	INTRON1	PRE_GENE	143	102	71,3
	1973	+	39	272	PRE_GENE	PRE_GENE	136	93	68,4
Mab21l2	**4939**	+	42	580	PRE_GENE	PRE_GENE	142	123	86,6
	4940	+	37	348	PRE_GENE	PRE_GENE	155	137	88,4
Hmx3	**2032**	+	150	792	PRE_GENE	POST_GENE	165	98	59,4
Lmx1b1	**4049**	+	300	763	INTRON2	PRE_GENE	116	95	81,9
3110004L20Rik*	**VF_5491**	-	45	700	PRE	POST	65	27	41,5
	VF_5492	+	39	863	INTRON	PRE	122	113	92,6
Elmo1	**FF_6026**	-	45	759	INTRON	INTRON	103	75	72,8
Ets ctr	**VC_11216**	-	-	613		INTRON2	104	59	56,7
Gata3 ctr	**VC_3255**	+	-	704		POST_GENE	174	127	72,9
1300007F04Rik ctr	**VC_2797**	-	-	913		POST_GENE	157	120	76,4
Tmeff2 ctr	**VC_198**	-	-	656		INTRON4	145	50	34,5
Mab21l2 ctr	**VC_909**	-	-	576		PRE_GENE	165	108	65,4
3110004L20Rik ctr	**VC_410**	-	-	769		INTRON2	107	34	31,7
Elmo1 ctr	**VC_10157**	-	-	780		PRE_GENE	146	99	67,8
Shh ctr	**VC_11271**	-	-	633		PRE_GENE	165	128	77,5
Impact ctr	**VC_5990**	+	-	596		INTRON4	150	112	74,7
Ubl7 ctr	**VC_268**	+	-	682		POST_GENE	117	93	79,5
Lmx1b1 ctr	**VC_11767**	-	-	536		POST_GENE	116	43	37,1
Irx3 ctr	**VC_5945**	-	-	550		POST_GENE	93	35	37,6

Table 6: The summary of the analysed SCEs and controls

SCEs are grouped under the name of the closest genes. The * mark indicates genes where only a predicted zebrafish homologue found. The length of the element indicates the length of the conserved sequence, while the size of the actual fragment is shown in the length of the PCR product column. The position of the fragments in mouse and in fugu indicates the tendency of the mobility.

In general, the additional DNA molecules (both control or SCE fragments) enhanced the expression of the promoter construct (Figure 13.), thus detailed mapping of the expression patterns and statistical analysis were needed for the determination of significant enhancer activity. Probably due to the properties of the promoter, the reporter expression was upregulated in muscle cells by almost all fragments (Figure 14.A-B), while only some fragments could increase the *lacZ* expression in other tissues like the notochord (Figure 14.C-D), the central nervous system (Figure 14.E-F) or the endothel of the developing vascular system (data not shown).

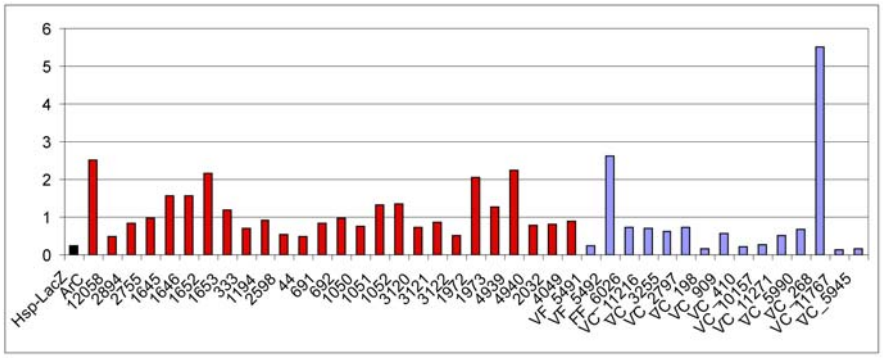


Figure 13: The lacZ-expressing cells per embryo ratios for the analysed SCEs and controls

The cell-counts were used to define statistically which fragments exhibited tissue-restricted or general enhancer activity. The number of expressing cells, for each co-injection and for each tissue, was compared with the number of expressing cells of the negative controls, when the average of cells expressing *lacZ* in the injected embryos were higher than in the control. When *lacZ* expression was increased in particular tissues, Fisher exact tests were used on the dataset, and a P value cut off 0.01 was used to classify a fragment as a tissue-restricted enhancer. The identification of the generic enhancers was performed by establishing the average and standard deviation of the number of expressing cells per embryo in the control group. Those fragments were classified as generic enhancers, in which the number of expressing cells per embryo was higher than the average plus twice the standard deviation of the control fragments. As one control fragment (a control for *ubl7*) had an extremely high activity in the central nervous system and in the notochord, this fragment was excluded from the calculations of the average and standard deviation.

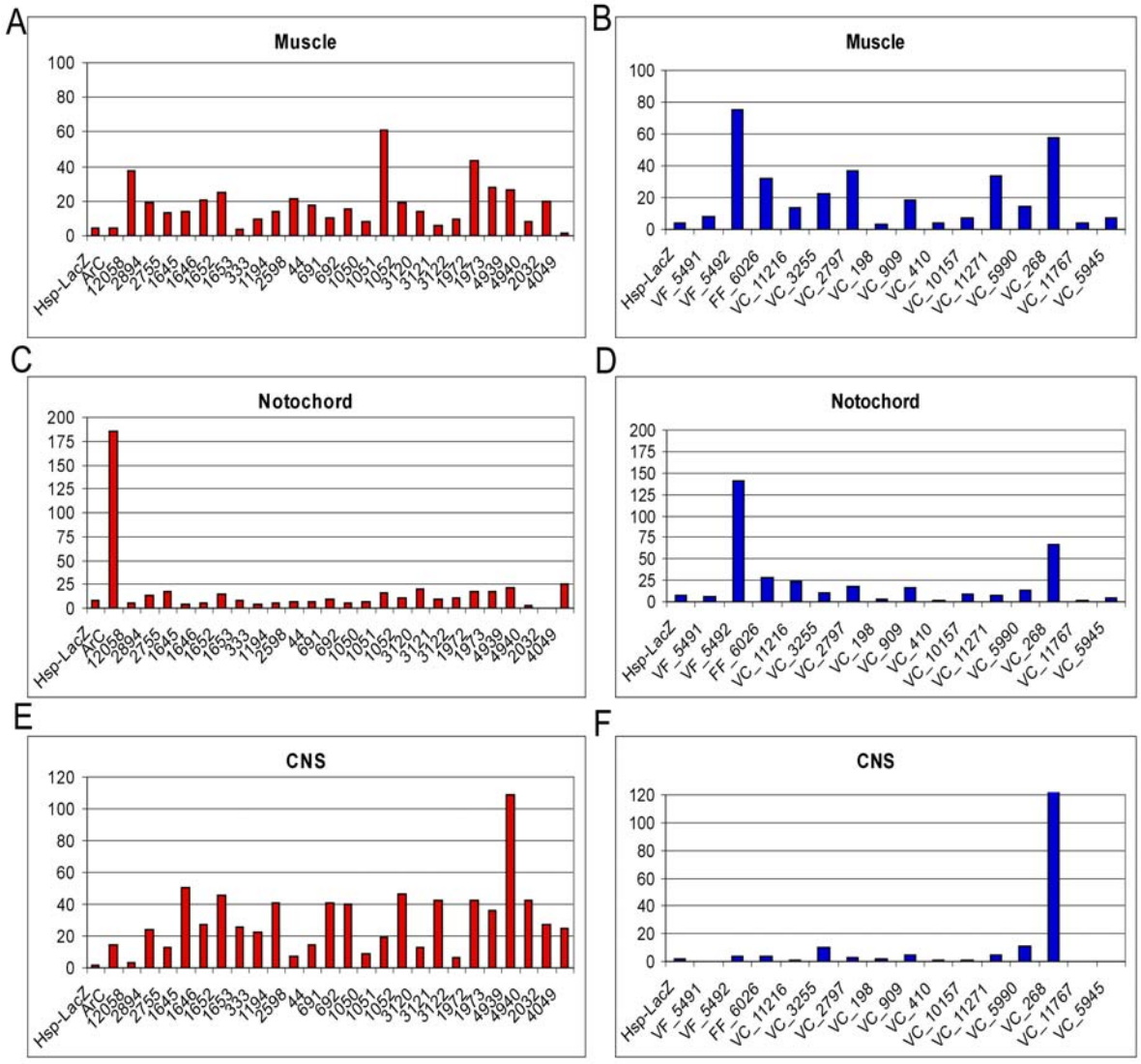


Figure14: The lacZ-positive cells per embryos ratios in distinct tissues

A: The lacZ-positive cells per embryo ratios in the muscle for the analysed SCEs, B: for the controls. C: The expressing cells per embryo in the notochord for the analysed SCEs, D: for the controls. E: The lacZ-positive cells in the central nervous system per embryo ratios for the analysed SCEs, F: for the controls.

Based upon the statistical analysis, 22 out of the 28 fragments (representing 79%) showed enhancer activity; while only 3 out of the 12 investigated controls (25%) were positive in the enhancer assay (Table 7.). Twenty out of the 22 SCEs with enhancer activity turned to be tissue-specific, from which 6 were fragments assigned to non trans-dev genes. From the control group, all three fragments showing enhancer activity were tissue-specific. These results indicate a broader range of conserved cis-regulatory elements than previously described.

Vicinal Gene	trans-dev	SCE ID	injected embryos	muscle	notochord	central nervous system	eye	ear	vessels
no element	N	-	161						
Shh	Y	ArC	96		8,48E-07				
		12058	139	6,86E-09					
Otx2	Y	2894	111	0,644		0,00627	0,5536	0,3155	
Gata3	Y	2755	107			0,398	0,5764	0,1906	
Ets	Y	1645	105			0,00259			4,78E-09
		1646	133			0,1558	0,6015	0,3619	2,15E-06
		1652	159			0,05534	0,6136	0,1485	2,08E-06
		1653	149			0,0444	0,129	0,07924	1,30E-05
Pax2b	Y	333	149			0,00237	0,06327	0,1902	
Pax6a	Y	1194	133			8,2E-06	0,3343	0,01268	
Pax3	Y	2598	124	0,02982		0,5287	1		
1300007F04Rik	N	44	107	0,2929					
Zfpm2	Y	691	140			1,49E-06	0,01296	1	
		692	131			3,58E-04	0,04369	0,1231	
Tmeff2	N	1050	164			0,7654	0,02301	0,3371	
		1051	120	1,04E-03		0,303	0,2088		
		1052	109			6,31E-04	0,0149	0,5862	
Jag1b	Y	3120	136			0,1849	1	1	
		3121	142			5,45E-08	6,52E-03	0,3245	
		3122	106			0,5088	1	0,5058	
Mapkap1	N	1972	143			0,05292	0,3788	0,6065	
		1973	136			4,04E-03	0,5973	0,077	
Mab21l2	Y	4939	142			1,24E-07	4,99E-03	0,2339	
	Y	4940	155			7,85E-08	4,14E-03		
Hmx3	Y	2032	165			0,00103	0,07062	0,01423	
Lmx1b1	Y	4049	116			0,00762	0,1876	1	
3110004L20Rik	N	VF_5491	65	0,2929					
		VF_5492	122	0,1874	0,01209				
Elmo1	N	FF_6026	103	7,13E-03	0,6848				
Ets ctr	Y	VC_11216	104	1					0,6954
Gata3 ctr	Y	VC_3255	174	0,04481		0,281	0,5739	0,0216	
1300007F04Ri ctrk	N	VC_2797	157						
Tmeff2 ctr	N	VC_198	145	0,7448		0,6597		0,3651	
Mab21l2 ctr	Y	VC_909	165	0,06359		1	1	1	
3110004L20Rik ctr	N	VC_410	107						
Elmo1 ctr	N	VC_10157	146	0,287	0,8126				
Shh ctr	Y	VC_11271	165	3,34E-07		1	1	1	
Impact ctr	Y	VC_5990	150	0,6496		0,2754		0,0622	
Ubl7 ctr	N	VC_268	117	3,33E-04		7,15E-11	0,02555	0,6097	
Lmx1b1 ctr	Y	VC_11767	116	0,2743				0,0707	
Irx3 ctr	Y	VC_5945	93	0,03938					

Table 7: The results of the statistical analysis for the SCEs

The red label indicates statistical significant enhancer activity.

In several cases, multiple SCEs found within a single gene locus gave similar tissue-specific enhancer activity. For example, all four SCEs tested from the *ets1* locus gave expression in the endogenous *ets1*-specific expression domains, namely in the developing blood vessels and blood precursors (Thisse 2004) (Figure 15.).

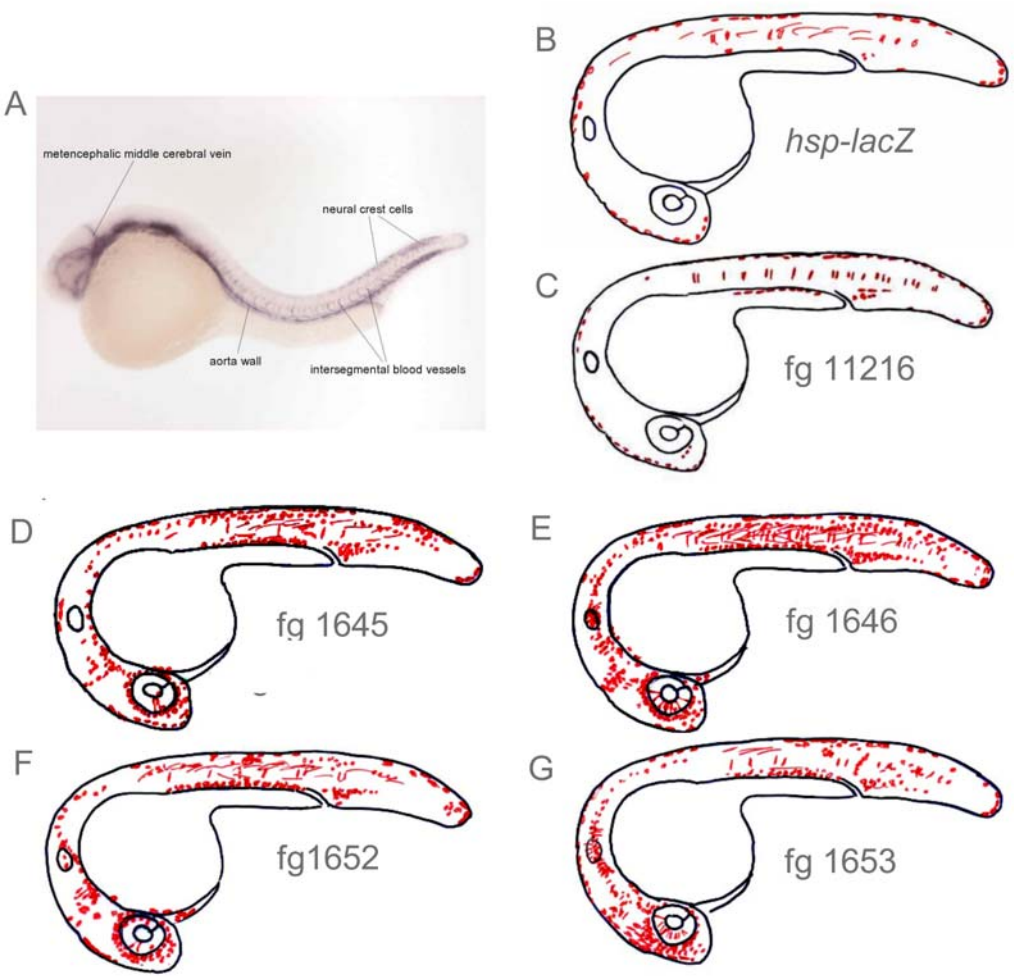


Figure 15: Expression profiles of the embryos co-injected with the *ets1* fragments.
A: ISH performed with *ets1*-specific probe, downloaded from the zfin database. *ets1* is expressed in the neural crest and in the blood vessels around the retina, in the metencephalon, in the ventral truck. The expression maps below represent approximately 120-150 embryos. B: *hsp68-lacZ* minimal promoter-reporter construct, C: fg11216 is a non-conserved control fragment from the *ets1* genomic region, co-injected with *hsp-lacZ*, D-G: SCEs located in the *ets1* genomic region, co-injected with *hsp-lacZ*.

Both elements tested from the *mab21-like2* genomic region gave central nervous system (CNS) and eye specific enhancer activity, but the strength of the two SCEs was significantly different. Both fragments directed the reporter activity into ectopic brain regions as well, as the endogenous *mab21l2* is not expressed in the

telencephalon, while the LacZ staining was quite strong in the forebrain for the embryos co-injected with the *mab21l2* associated fragments. The negative control for this genomic region gave no brain expression at all (Figure 16.).

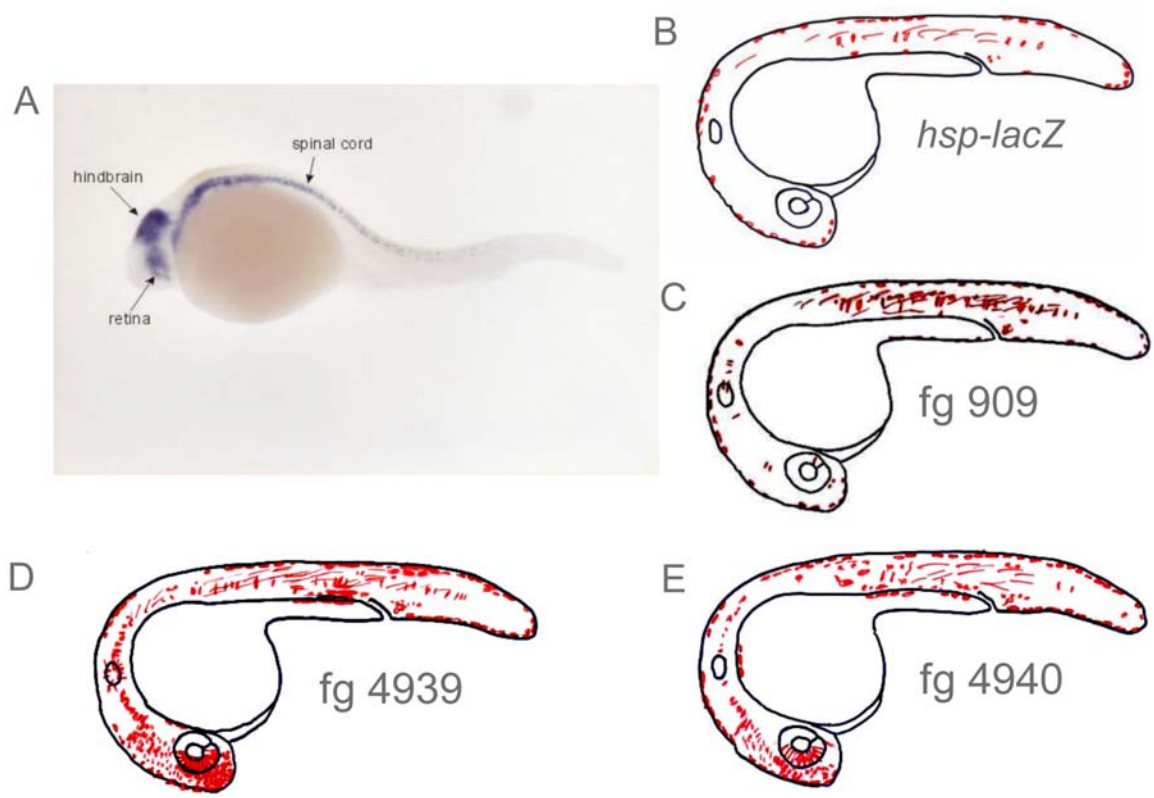


Figure 16: Expression profiles of the embryos co-injected with the *mab21l2* fragments.
A: ISH performed with *mab21l2*-specific probe, downloaded from the zfin database. *mab21l2* is expressed in restricted areas of the CNS, like the eyes, midbrain and some neurons in the spinal cord. The expression maps below represent approximately 120-150 embryos. B: *hsp68-lacZ* minimal promoter-reporter construct, C: fg909 is a non-conserved control fragment from the *mab21l2* genomic region, co-injected with *hsp-lacZ*, D-E: SCEs located in the *mab21l2* genomic region, co-injected with *hsp-lacZ*.

Both elements for the *zfpm2* (the homologue of *fog2*) gave CNS-specific enhancer activity, which is in accordance with the reported expression pattern of the endogenous gene in zebrafish (Figure 17.).

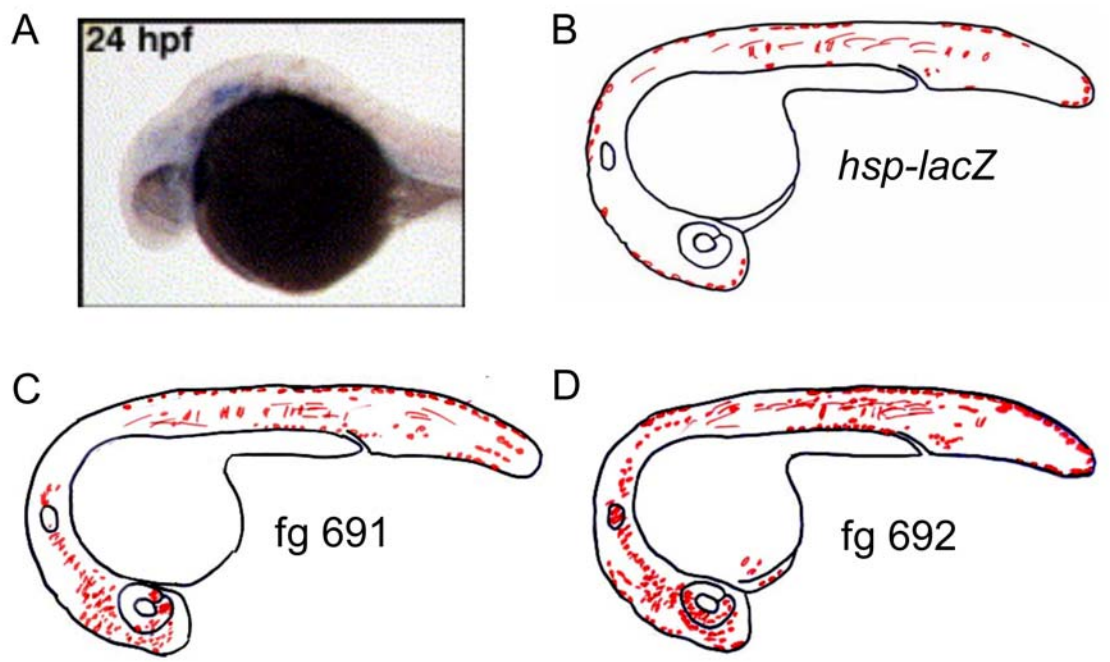


Figure 17: Expression profiles of the embryos co-injected with the *zfpm2* fragments.
A: ISH performed with *zfpm2*-specific probe, from (Walton et al. 2006). *zfpm2* is expressed in some brain regions, in the heart and in the intermediate cell mesoderm. B: *hsp68-lacZ* minimal promoter-reporter construct, C-D: SCEs located in the *zfpm2* genomic region, co-injected with hsp-lacZ.

In contrast, there were genomic regions from which only one out of several SCEs showed tissue-specific enhancer activity, the effect of the other fragments were comparable with the controls, based on the statistical calculations (Table 10.), like in the case of *jag1b*. *Fragment 3121* gave specific expression in the CNS and in the eye, which is partially overlapping with the endogenous *jag1b* expression; it is expressed in the rostral end of the pronephric duct, in nephron primordia and in brain regions extending from the otic vesicle to the eye (Thisse 2004) (Figure 18.).

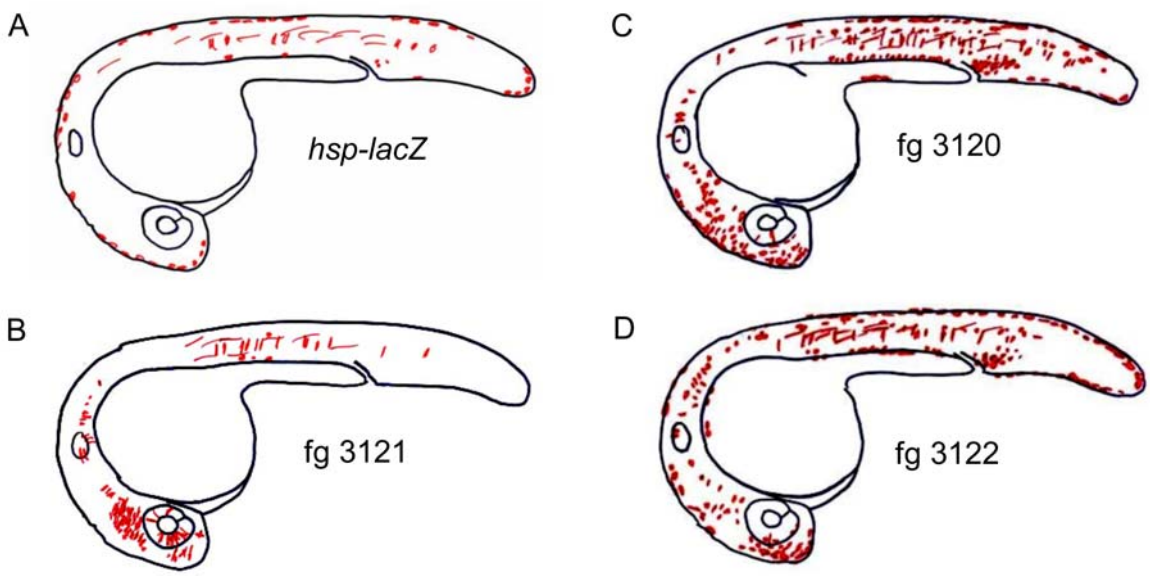


Figure 18: Expression profiles of the embryos co-injected with the *jag1b* fragments.
At 24-hpf-stage *jag1b* is expressed in the pronephros and in the region extending from the otic vesicle to the eye. The expression maps represent approximately 120-150 embryos. A: *hsp68-lacZ* minimal promoter-reporter construct, B-D: SCEs located in the *jag1b* genomic region, co-injected with *hsp-lacZ.*

SCEs assigned to genes, which do not belong to transcription factor and/or developmental regulator ("trans-dev") GO categories (*mapkap1*: mitogen-activated protein kinase associated protein 1, an integral membrane protein; *tmeff2*: a putative transmembrane protein, predominantly expressed in the mouse brain; *elmo1*: engulfment and cell motility 1 gene, homologue of C. elegans *ced-12,* involved in actin cytoskeleton organisation; and *3110004L20Rik:* a transmembrane transporter protein), were tested for enhancer activity as well. *mapkap1* is ubiquitously expressed (Thisse 2004), while *elmo1'*s expression is restricted to the central nervous system, lateral line primordial, lens, olfactory placode and blood vessels (Thisse 2004). For *tmeff2* and *3110004L20Rik* no expression data is available in zebrafish. Two SCEs assigned to *mapkap1*, two out of three SCEs assigned to *tmeff2*, two fragments from the *3110004L20Rik* genomic region and one from the *elmo1* locus activated the *lacZ* expression in distinct domains, showing significant enhancer activity (Figure 19.).

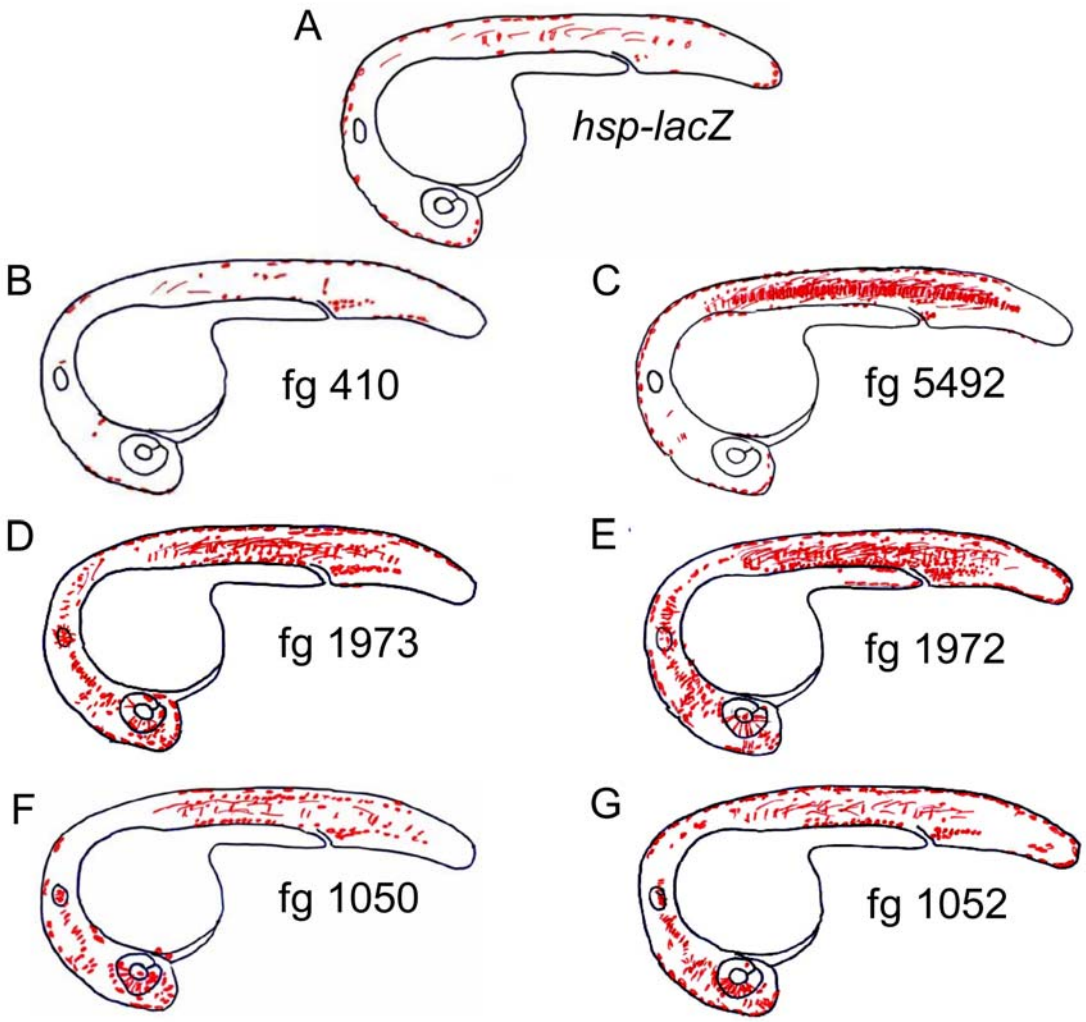


Figure 19: Expression profiles of the embryos co-injected with SCEs of non- "trans-dev" genes. A: *hsp68-lacZ* minimal promoter-reporter construct, B: a control fragment from the *3110004L20Rik* genomic region, co-injected with *hsp-lacZ*. C: an SCE located in the *3110004L20Rik* genomic region, co-injected with *hsp-lacZ*. No expression information is available for this gene. D-E: SCEs located in the *mapkap1* genomic region, co-injected with *hsp-lacZ*. *mapkap1* expression is not spatially restricted, based on (Thisse 2004). F-G: SCEs located in the *tmeff2* genomic region, co-injected with *hsp-lacZ*. No expression information is available for this gene.

4.2.3 Discussion

A new alginment method found more and diversed types of conserved noncoding sequences

With the combination of a global and a local alignment method at the genome level, in this screen 21.427 non-genic shuffled conserved elements (SCEs) were identified; approximately 30% of the analysed genes presented at least one SCE. This number is roughly a magnitude higher than the number of highly conserved non-coding sequences found by a previous similar approach (Woolfe et al. 2005). 72% of the elements identified were shifted in terms of orientation and/or position when compared in different species. Approximately 50% of SCEs do not overlap with previously reported datasets of conserved sequences, suggesting that the use of nonexact seeds for the initial local alignments has a significant impact on the analysis of noncoding DNA harbouring short, well conserved elements. This dataset overlaps only 45% of the UCE elements (Bejerano et al. 2004), and 51% of the CNEs (Woolfe et al. 2005) within the loci analyzed, probably because of the regional approach taken, which disregards elements conserved across nonorthologous loci. Detailed analysis of the shuffled elements showed that 1kb regions upstream of the TSSs showed less mobility, suggesting that the majority of non-coding conserved elements are located outside of these regions.

Although there is a significant over-representation of gene classes of transcription factors and developmental regulators among the genes assigned to the SCEs, using this type of analysis we have found not only higher number of conserved non-coding elements, but elements assigned to different types of genes as well, like extracellular or behavioural genes.

The majority of the tested shuffled consreved elements show enhancer activity in zebrafish

The set of SCEs contained 18 from the 98 already published mouse enhancers present in Genbank, ten times more as the conserved non-coding (Woolfe et al. 2005) or the ultraconserved elements (Bejerano et al. 2004). Four out of these 18 mouse enhancers overlapped with the SCE-series I tested, and I could show the enhancer activity of their fugu counterparts in zebrafish. The relative evolutionary closeness of

fugu and zebrafish implies that expression and regulation of expression of developmentally regulated genes is probably well conserved (Miles et al. 2003; Woolfe et al. 2005). Only one of the fragments that we tested (SCE 1973 from the *mapkap1* gene) overlaps with a UCE element. The overlap is only 33 bp, but the element nonetheless acted as a tissue-restricted enhancer *in vivo*. A region adjacent to the UCE in mouse (SCE 1973), although not ultraconserved, is also conserved in fish and acted as a generic enhancer in our assays, highlighting the complexity of these regions.

Based on my previous results with the functional analysis of the *pax2* CSTs, I performed the enhancer test as a co-injection assay, using the basal mouse *hsp68* promoter linked to a *lacZ* gene. In general, the additional DNA molecules (both control or SCE fragments) enhanced the expression of the promoter construct. Probably due to the properties of the promoter, the reporter expression was upregulated in muscle cells by almost all fragments, while other tissues were activated by only a subset of the tested fragments. Similar phenomena were experienced in transgenic mice when the *hsp68* promoter was used to generate the transgenes. Ectopic expression patterns were observed in a number of founder embryos in the spinal cord, which was independent of the enhancer elements used. When the *hsp68* promoter in construct was replaced with the *En-2* promoter, the spinal cord expression was lost and the *En-2*-specific expression pattern was retained. So therefore it was suggested that the *hsp68* promoter fragment used in these studies contained an element that is capable of directing expression to the spinal cord and that such expression is only detectable when the promoter is flanked by a strong enhancer element (Logan et al. 1993).

Altogether, 79% of the amplified and tested fugu SCEs (22 out of 28) showed significantly enhanced reporter expression, from which 20 showed tissue-specific enhancer activity. Multiple SCEs assigned to *ets1*, *mab21l2* and *zfpm2* genes gave similar expression patterns, indicating that a single gene can have several enhancers with similar activities. This functional redundancy is a well-described phenomenon (Jongens et al. 1988; Tebb et al. 1989; Buttgereit 1993). For example the two enhancer elements of *math1* (a transcription factor of the bHLH class, which is expressed during development in multiple neuronal domains), while dissimilar in sequence, appear to have redundant activities in the different *math1*-specific expression domains except the spinal neural tube (Helms et al. 2000). Redundant

enhancers were found to control the *shh* expression in the ventral spinal cord, hindbrain and regions of the telencephalon (Jeong et al. 2006).

In other genomic regions just the subset of the fragments showed significant tissue-specific enhancer effect (e.g. *jag1b, tmeff2*). From the fragments assigned to non- "trans-dev" genes (*mapkap1, tmeff2, elmo1 and 3110004L20Rik)* at least one fragment per gene showed significant enhancer activity.

Three nongenic nonconserved control fragments have been shown to act as tissue-specific enhancers, one showed even higher activity then all the tested conserved sequences (the control fragment for *ubl7*). These are probably nonconserved enhancers hit by the random sequence choice.

In this screen, SCEs were assigned to the closest genes, although it has been shown that enhancers may act across intervening genes (Spitz et al. 2003), or can also be located within the introns of neighbouring genes (Lettice et al. 2003), located at distances of several hundred kilobases to over a megabase (Bishop et al. 2000; Jamieson et al. 2002; Lettice et al. 2003). Moreover, a recent report suggested, that only half of the cis-regulatory elements is located in a 250kb radius from the target promoter (Vavouri et al. 2006). For those genes, where expression information exist in zebrafish, we could confirm the target gene choice, in all cases the expression patterns gained with the enhancers were partially recapitulating the expression patterns of the closest genes. For those genes, of which we lack the expression information (*elmo1, 0300007F04Rik* and *3110004L20Rik*), I could not perform this comparison, thus I cannot be confident, that these are regulated by the assigned sequences.

Although conserved non-coding sequences were reported to harbour other functions (Bernstein et al. 2006; Feng et al. 2006; Lee et al. 2006; Calin et al. 2007; Lareau et al. 2007; Ni et al. 2007), I only tested these elements for enhancer activity, and I only used one developmental stage (24 hpf), thus the elements turned to be silent in this screen, still can harbour function.

4.3 A high throughput screen to investigate promoter-enhancer specificity

In the original genomic context, one enhancer generally has only one target gene, although a cis-regulatory element can be located even in a megabase range around the promoter, surrounded with other potential targets. There are mechanisms that can restrict the promiscuity of the cis-regulatory elements, described in details under point 1.8. Insulators or boundary elements can block the interaction between promoters and enhancers (Levine et al. 2003), specific regulatory elements can "guide" the enhancers to specific promoters (Zhou et al. 1999), or cis-regulatory elements can compete with each other for the interaction with an enhancer or a promoter (Kmita et al. 2002; Lin et al. 2007). Several studies have shown that the core promoter sequence context can significantly influence the responsiveness of a given gene to gene-specific DNA-binding activators and repressors (Simon et al. 1988; Metz et al. 1994; Ernst et al. 1996). Studies in Drosophila have provided evidence that core promoter structure plays an important role in selectivity of enhancers for their target genes (Li et al. 1994; Ohtsuki et al. 1998; Butler et al. 2001).

To elucidate if the sequence of the cis-regulatory elements already determines the interaction specificity in transcription regulation during vertebrate development, we addressed the following questions. Do promoters isolated from their original genomic context show specificity towards interacting with different enhancers? Do isolated enhancers "select" from a set of promoters? Based on the textbook knowledge, the answer to these questions would be no, as enhancers by definition should activate any promoter (Banerji et al. 1981; Atchison 1988), but experimental results do not uniformly confirm this (Wefald et al. 1990; Li et al. 1994; Keplinger et al. 2001). Furthermore we wanted to know, whether an isolated interdigitate enhancer is able to interact with both promoters of its target and bystander genes. To answer these questions 13 enhancers and 20 zebrafish promoters (including controls) have been selected to generate the 260 possible combinations in Multisite Gateway expression vectors. We have tested the transcriptional activity and strength of these constructs by generating transient transgenic zebrafish, and the analysis of this high throughput screen revealed that the sequence of the regulatory elements is an important determinant of the interaction specificity.

4.3.1 Identification of enhancers

I have chosen enhancers with published expression patterns from the literature (Table 8. summarizes the enhancers selected). As the scientific interest is much greater in relation to genes expressed in the nervous system, our collection is also overrepresented with enhancers of these genes (like the *arC*, driving the expression of the *shha* gene into the floorplate, notochord and hypothalamus; the eye enhancer of *pax6b;* the midbrain-hindbrain boundary specific *CXE* enhancer of *eng2;*, the brain enhancer of the *dre-mir9-1* microRNA gene; the *isl1 zCREST2*, a sensory and motor neuron enhancer; the forebrain-specific *ei* enhancer of the *dlx2b/dlx6a* gene cluster and the *regB* enhancer, driving the expression of *mnx1* into the spinal motor neurons). The only enhancer element that is not from a publication is the *eng2b reg5*, which was identified and tested in our laboratory.

	Name	Size (bp)	Origin	Reference	Nature of functional verification
00	ctr (VC_909)	576	Takifugu rubripes	(Sanges et al. 2006)	co-injection
01	*shha arC*	462	Danio rerio	(Muller et al. 1999)	co-injection, cloned fragment injection
04	*bactin2 intron1*	508	Cyprinus carpio	(Muller et al. 1997)	co-injections
06	*pax6b eye enh.*	343	T. rubripes	(Woolfe et al. 2005)	co-injections
07	*eng2b CXE*	968	D. rerio	(Song et al. 1996)	
08	*eng2b reg5*	416	D. rerio	New enhancer	co-injection cloned fragment injection
09	*dre-mir9 brain enh*.	365	D. rerio	(Kikuta et al. 2007)	enhancer trap
10	*myl7 heart enh.*	285	D. rerio	(Huang et al. 2003)	deletion, cloned fragment injection
11	*myf5 somite enh.*	308	D. rerio	(Chen et al. 2007)	deletion
15	*isl1 enh.*	724	D. rerio	(Uemura et al. 2005)	cloned fragment injection
16	*dlx2b/dlx6a ei*	479	D. rerio	(Zerucha et al. 2000)	cloned fragment injection
17	*mnx1 regB*	215	D. rerio	(Nakano et al. 2005)	tg mice with cloned fragment
18	*kdrl enh.*	812	D. rerio	(Choi et al. 2007)	deletion, cloned fragment injection

Table 8: Summary of the enhancers used in the project

The list of enhancers also contains elements with expression domains outside of the nervous system, like the intronic enhancer of the *ß-actin* gene, with general expression pattern, the heart enhancer of the *myl7* gene, the somite enhancer of *myf5* and the enhancer driving expression of the *kdrl* gene to the developing vascular endothel. For having more diverse types of cis-regulatory elements, I cloned responsive elements of inducible genes as well, like the estrogen responsive element

of the *cytochrome P450 oxidase* gene or the xenobiotic responsive element (*XRE)* element of the *metallothionenin2* gene, but these elements did not show upregulation of the reporter upon published induction circumstances (estradiol or zink treatment) in the pre-screen tests, so these were omitted from the later assay. As a negative control, I used a nonconserved noncoding Fugu sequence (VC_909) showing no significant enhancer activity in my previous assay (Sanges et al. 2006). Figure 20. illustrates the expected expression domains of the enhancers.

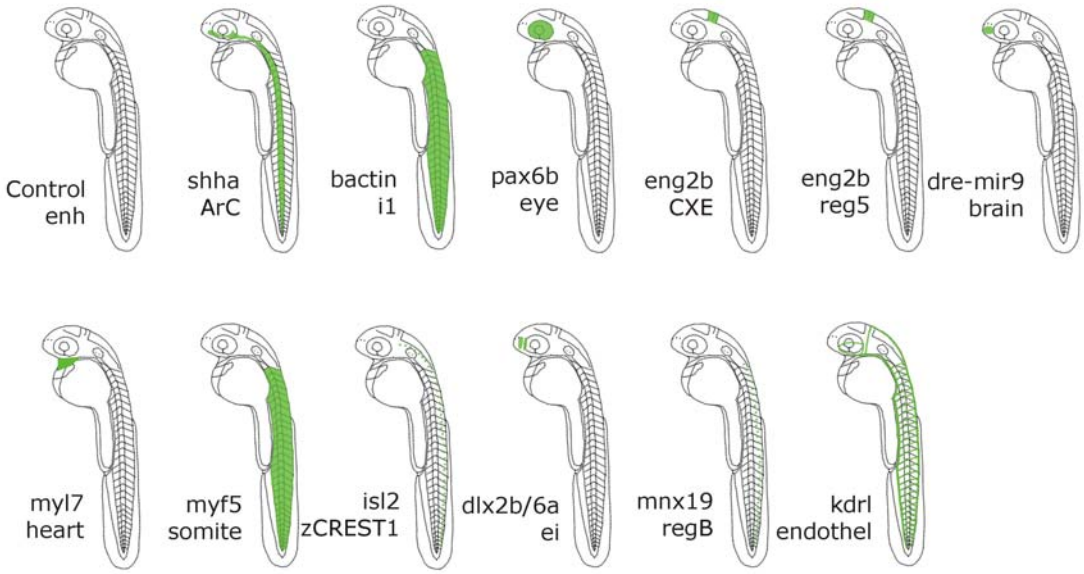
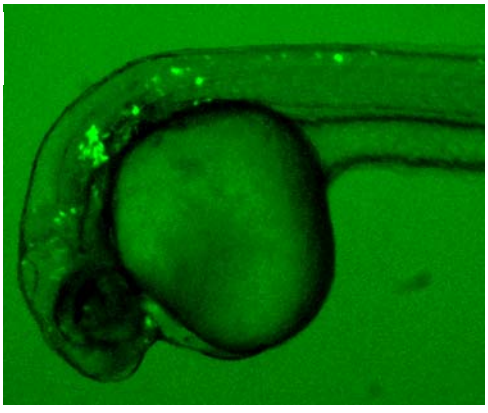


Figure 20: The expected expression domains of the enhancers at prim 16 stage

Sonic hedgehog (**Shh**) is a signalling molecule expressed in the midline mesoderm of vertebrates (Strahle et al. 1996), playing a crucial role in the induction of floor plate and motor neurons (Echelard et al. 1993; Chiang et al. 1996), and in the proper development of the limbs (Riddle et al. 1993). The analysis of the introns of the zebrafish *shha* gene revealed in identification of multiple enhancers. *Activator region C* (*arC*) was shown to direct expression into the notochord (Muller et al. 1999). **β-actin** is expressed in nonmuscle cells (Clarke et al. 1977). The regulatory elements (a proximal promoter, an upstream negative regulatory element, and an orientation- and position-dependent enhancer-like element) responsible for the expression of the common carp (*Cyprinus carpio*) *β-actin* were identified in the first intron (Liu et al. 1990b). The intronic enhancer element was showing non-restricted expression pattern in zebrafish in a co-injection experiment (Muller et al. 1997). **Pax6** is a highly conserved protein, with paired- and homeodomain DNA-binding regions;

it functions as a transcription factor with a major role in eye and brain development from Drosophila to humans (Callaerts et al. 1999; van Heyningen et al. 2002). Sequence comparison of fugu and human DNA using MegaBlast resulted in seven conserved sequences around the fugu *pax6* genes. The element called *pax6_19* (EMBL accession number: CR847483) is located in an intron of the neighbouring housekeeping gene, *elp4*. This element drove the expression of the *β-actin* promoter-containing reporter construct into the eye, forebrain, hindbrain, spinal cord and skin in a co-injection assay (Woolfe et al. 2005). **Engrailed** is a homeoprotein with multiple roles in directing anterior-posterior patterning during vertebrate (Joyner et al. 1985) and invertebrate (Nusslein-Volhard et al. 1980) development. The zebrafish *engrailed2* genes are expressed across the presumptive midbrain–hindbrain boundary (MHB) with distinct temporal and spatial profiles (Ekker et al. 1992). A 1-kb enhancer element (called *CX*) directing the mouse *engrailed2* expression into the MHB was found to be conserved in human, and the human enhancer showed MHB-specific enhancer activity in transgenic mice (Song et al. 1996). This cis-regulatory element shares high similarity with a sequence upstream of the zebrafish *eng2b* locus. This zebrafish homologue (named as *CXE*) was used in our screen, as in co-injection tests it was able to drive reporter expression int. o the MHB (Figure 21.A). During the comparisons of the non-coding sequences in and around the zebrafish, human and mouse *engailed2* locus we have found an intronic conserved element. This sequence, called *reg5,* showed enhancer activity in co-injection experiments (Figure 21.B).

A



B

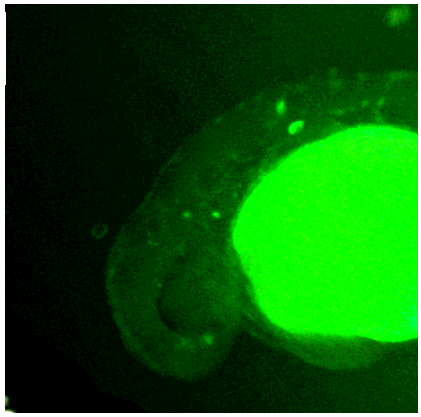


Figure 21: Fluorescent pictures of embryos co-injected with *eng2b*-promoter-*yfp* and *eng2b CXE* (A) and *reg5* (B) enhancers. The fluorescent signal was detected in the MHB and in other neural tissues such as hindbrain and spinal cord, and in other cell types (like the YSL) as well.

Mir9 is a microRNA playing role in brain development (Krichevsky et al. 2003). The enhancer, which is located between the *mef2d* and *rhgb* genes, drives the expression of the *dre-mir9-1* into the developing brain of zebrafish embryos was discovered in an enhancer trap experiment (Kikuta et al. 2007). Regulatory myosin light polypeptide 7, (**Myl7**), formerly called as Cardiac myosin light chain (Cmlc2) is the major contractile component of cardiac striated muscle in zebrafish (Yelon et al. 1999), a homologue of human and murine Mlc-2. It is expressed in the zebrafish cardiac cells fused into a single heart tube at 24hpf stage (Huang et al. 2003). In deletion series, a 244-bp sequence, located upstream of the core promoter, was identified as an enhancer driving *myl7* in the heart. This fragment was tested in combination with the endogenous promoter and with the CMV promoter as well (Huang et al. 2003). **Myf5** is a basic helix-loop-helix transcription factor that functions as a myogenic factor in the specification and differentiation of muscle cells (Pownall et al. 2002). Its expression is finely tuned by cis-regulatory control mechanisms. Deletion analysis of the 80-kb upstream sequence of the zebrafish *myf5* gene resulted in the characterization of distinct enhancers. The segment responsible for the basal transcription in somites and the presomitic mesoderm was positioned into the –290/-1 fragment (Chen et al. 2007). Islet-1 (**Isl1**) is a LIM-homeodomain protein, expressed at the earliest stage of neural differentiation, and is highly conserved during evolution (Ericson et al. 1992; Inoue et al. 1994; Thor et al. 1997; Jackman et al. 2000). The *isl1 zCREST2* enhancer, which is conserved in human, mouse and chick, located at 53kb downstream of the zebrafish *isl1* locus, was described as an enhancer element activating gene expression in primary sensory neurons and in spinal motor neurons innervating the abductor muscle of the pectoral fin bud and the ventral trunk muscles (Uemura et al. 2005). In stable transgenic fish the reporter expression was observed also in other tissues, such as notochord and commissural interneurons in the spinal cord (Uemura et al. 2005). Four out of the six mammalian *distall-less-related homeobox* (***dlx***) genes are arranged in a head-to-tail manner, and the gene pairs show overlapping expression domains in the ventral telencephalon and diencephalons (Liu et al. 1997). This genomic arrangement is conserved in distantly related vertebrates as zebrafish (Ellies et al. 1997). A regulatory element between the zebrafish *dlx2a/dlx6a* gene pair was identified as an enhancer

element responsible for the ventral forebrain activity in developing transgene zebrafish embryos (Zerucha et al. 2000). Motor neuron and pancreas homeobox 1 (**Mnx1**), previously called as Hb9, is a transcription factor serving as a marker for motor neurons in developing vertebrate embryos (Tanabe et al. 1998), probably playing role in the consolidation and maintenance of motor neuron identity (Arber et al. 1999). Using a cross-species homology analysis, a highly conserved 125-bp sequence was identified, and this sequence was able to drive expression into the motor neurons of the transgene zebrafish (Nakano et al. 2005). Kinase insert domain receptor (**Kdr**) is the major receptor for Vascular endothelial growth factor (VEGF) on endothelial cells in vertebrates (Marcelle et al. 1992). During embryogenesis it is required for both vasculogenesis and angiogenesis (Shalaby et al. 1995). In zebrafish, two genes have been identified with similar sequence and function (Habeck et al. 2002; Covassin et al. 2006; Bussmann et al. 2008). Deletion analysis of the 6.4-kb genomic sequence upstream of the TSS revealed that an approximately 800-bp DNA fragment (at -4.3kb position) is sufficient to drive expression of *kdrl* in endothelial cells. GFP expression was detected in transgenic fish in the dorsal aorta, posterior cardinal vein, intersomitic vessels, endothelial cells in the brain and in neural tissues such as brain, eyes and neural tube in addition to endothelial cells by 24hpf (Choi et al. 2007).

4.3.2 Identification of promoters

Basal promoters with different TSS distribution, core promoter composition and strength have been selected from the already cloned and tested set of promoters present in the Müller lab. The TSSs have been identified by using the ESTs present in the dbTSS or the ENSEMBL databases (Figure 22.). Promoters were classified into the following four distinct categories based on Kawaji et al 2006.: core promoters showing the TSS distribution of a.) a single dominant peak, b.) a general broad distribution, c.) a broad distribution with a dominant peak, and d.) a bi-or multimodal distribution (Figure 2). Depending on the TSS distribution of the promoters, I amplified roughly 120-200-bp piece of the core promoter regions. In case of promoters with a dominant TSS peak (when the majority of the ESTs were clustering around one position), I took ~75 base pairs upstream and ~50 downstream from the major TSS. For the bimodal promoters (in which two major TSS peaks could be observed, Figure 22.), the amplified fragment contained ~75 base pairs upstream from

the 5` TSS and ~50 downstream from the 3` TSS. In case of promoters showing broad TSS distribution (when the EST cluster did not show sharp peak), I have defined arbitrary the 5` and 3` ends of the core promoter regions: ~50 base pairs upstream from the first EST-alignment, and ~50 base pairs downstream from the last one. In all TSS-distribution categories, if the first ATG was within the downstream 50 base pairs, the last bases upstream to the ATG were used as 3' end of promoter region.

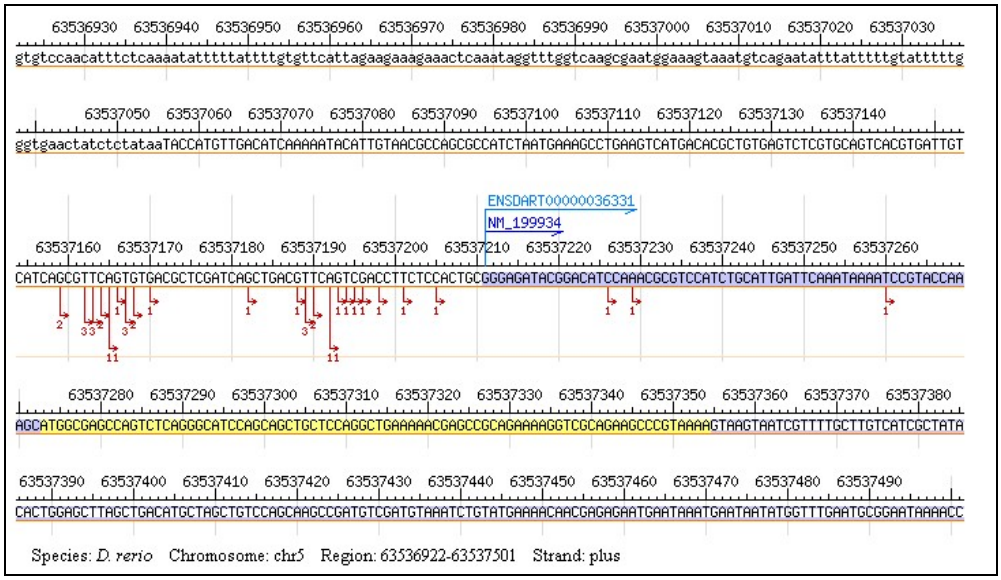


Figure 22: The bimodal promoter region of the *at6v1g1* gene

The light and the dark blue arrows represent the start points of the transcripts present in the Ensemble and NCBI databases. The red arrows represent the starting positions of all the known *atp6v1g1* ESTs aligned to the genomic DNA, showing a dominant bimodal TSS distribution. Downloaded from the dbTSS database

These promoters belong to genes from diverse gene ontology classes such as tissue-specific (*apoeb*), developmentally restricted (*ndr1*) or ubiquitous (*tbp*) genes. For some enhancers the endogenous promoters (*eng2b* for *eng2b CXE* and *reg5* enhancers, *shha* for the *shha arC* enhancer) and for others the target and the bystander promoter pairs were chosen (*dre-mir9-1* and *mef2d* promoters for the *dre-mir9-1* enhancer). As a negative control, I used the basal promoter of the *Ciona intestinalis fog (friend of gata)* gene (Roure et al. 2007), as my previous experiments showed that Ciona promoters were inactive in zebrafish embryos (data not shown). Table 9. summarizes the properties of the promoters used in the project.

	Gene symbol	Chromosomal position	Promoter size (bp)	Total number of ESTs	TSS distribution
00	ctr				-
01	*apoeb*	chr16:24388715-24388903	181	624	dominant
02	*atp6v1g1*	chr5:50839078-50839225	148	195	dominant, bimodal
03	*gtf2a1*	chr20:12836731-12836899	169	73	broad
04	*klf4*	chr2:27760887-27761017	131	130	broad
05	*krt4*	chr6:29046051-29046216	166	601	dominant
06	*ndr1*	chr21:11382510-11382652	143	3	not conclusive
07	*pcbp2*	chr9:614661-614812	152	107	broad
08	*rdh10*	chr2:24616214-24616320	107	210	dominant
09	*tbp*	chr13:24702224-24702404	181	106	dominant
10	*tram1*	chr24:17550732-17550953	222	102	broad
11	*c20orf45*	chr6:58935813-58935969	157	88	broad
12	*ccne*	chr7:43205849-43206021	173	20	broad
13	*shha*	chr7:36764532-36764712	181	3	not conclusive
15	*mef2d*	chr16:20571360-20571520	161	20	broad
16	*dre-mir9-1*	chr16:20549606-20549799	194	2	not conclusive
17	*elp4*	chr7:8747729-8747891	163	24	broad
19	*hsp70*	chr3:23548288-23548436	149	4	not conclusive
20	*lmbr1l*	chr7:38089239-38089403	165	0	not conclusive
21	*eng2b*	chr2:24409903-24410129	227	8	broad

Table 9: The promoters used in the project

Apoeb is an extracellular protein responsible for lipid transport. In zebrafish the expression of *apoeb* gene is very strong in the yolk syntitial layer (YSL) from blastula stage until larval development. Between the first and the third days of development, a new domain of *apoeb* gene expression appears in the head region, in the facial ectoderm, and in some cells of the retina and brain (Babin et al. 1997). ***atp6v1g1*** encodes a component of vacuolar ATPase (V-ATPase), a multisubunit enzyme that mediates acidification of eukaryotic intracellular organelles (Finbow et al. 1997). V-ATPase dependent organelle acidification is necessary for such intracellular processes as protein sorting, zymogen activation, receptor-mediated endocytosis, and synaptic vesicle proton gradient generation (Nishi et al. 2002). During zebrafish embryogenesis, it is expressed by 30hpf in the central nervous system (Rauch et al. 2003). The ***gtf2a1*** gene codes the general transcription factor IIA 1, a factor important in the transcription initiation from RNA Pol II promoters and shows non-restricted expression pattern throughout the embryonic development of the zebrafish (Thisse 2004). Krüppel-like factor 4 (**Klf4**) is related to the erythroid cell specific Krüppel-like factor EKLF in mammals (Kawahara et al. 2000). It is expressed in the hatching gland, blood, lateral line primordium and neuromasts of the Prim15-stage zebrafish

embryos (Thisse 2004). Keratin 4 (**Krt4**) belongs to the protein family of intermediate filaments and it is a component of the cytoskeleton of epithelial cells. Gene expression analysis during embryonic development revealed that this gene is expressed in all surface cells, notably in those of the enveloping layer (EVL) and of the periderm (Imboden et al. 1997). Nodal-related 1 (**Ndr1**), previously called as Squint is a member of the nodal-related subclass of the TGF-β family, and is essential for early steps in dorsal mesoderm development (Feldman et al. 1998). *ndr1* is expressed throughout the whole embryos at Prim5 stage (Hagos et al. 2007), no information is available for the later stages. Poly(rC) binding protein 2 (**Pcbp2**) is an RNA-interacting protein with a specificity for poly(C) homopolymer (Swanson et al. 1988). Its function is linked to mRNA stabilization, translational silencing and translational enhancement (Makeyev et al. 2002). In zebrafish it has a general expression pattern (Thisse 2004). Retinol dehydrogenase 10 (**Rdh10**) is a crucial protein in embryonic organ development of placental vertebrates by being involved in retinoic acid synthesis from maternal retinol. It is expressed in the developing brain and sensory organs in mouse embryos (Cammas et al. 2007). In zebrafish, *rdh10* was shown to be expressed in the notochord, tail bud and YSL in the 14-19-somite stage (Thisse 2001), no information is available about its expression at 30hpf stage. TATA-binding protein (**TBP**) mediates transcription initiation from TATA-box containing promoters (Buratowski et al. 1988). In zebrafish the isolated *tbp* promoter was shown to direct reporter expression throughout the developing embryo (Burket et al. 2008). Cotranslational translocation of most, but not all secretory proteins across the mammalian endoplasmic reticulum membrane requires the Translocating chain-associating membrane protein 1 (**Tram1**) (Voigt et al. 1996). *tram1* is expressed in the central nervous system, otic vesicle, pectoral fin musculature and pharyngeal arch3-7 skeleton by 30hpf in zebrafish embryos (Thisse 2001). **C20orf45** is a predicted orthologue of the vertebrate Slowmo homologue 2 protein. *slowmo* encodes a mitochondrial protein of unknown function in *Drosophila melanogaster*, which is essential for the development of the central nervous system (Dee et al. 2005). In vertebrates the in vivo function and the expression pattern of this gene is yet unknown. Cyclin E (**Ccne**), a protein essential for the control of the cell cycle at the G1/S transition (Sherr 1993), shows a restricted expression in the central nervous system in the 30-hpf-stage zebrafish embryo (Thisse 2004). The Myocyte enhancer

factor 2d (**Mef2d**) transcription factor plays role in muscle thick filament assembly. In zebrafish it is expressed in the heart and in slow and fast muscles (Hinits et al. 2007). Elongation protein 4 homolog (**Elp4**), previously called as Paxneb (Pax6 neighbour) is a part if the Elongator holoenzyme complex, regulating the transcription elongation from RNA Pol II promoters (Winkler et al. 2001). It is ubiquitously expressed in the developing zebrafish embryo (Thisse 2004; Kikuta et al. 2007). The 70-kDa Heat shock cognate protein (**Hsp70** or more precisely Hsc70) is a non-heat-inducible chaperone from the Hsp70 gene family (Ingolia et al. 1982), playing role in protein folding. The expression of zebrafish *hsp70* starts at 72 hpf during the embryonic development (Yeh et al. 2002). Limb region 1 (**Lmbr1**) is a transmembrane protein, playing role in limb development (Clark et al. 2000). The zebrafish homologue, *lmbr1l* is expressed in the whole embryo at Prim 15 stage (Thisse 2001).

The transcriptional activity of the *apoeb, atp6v1g1, gtf2a1, klf4, krt4, ndr1, pcbp2, rdh10, tbp, tram1, c20orf45, ccne* and *shha* promoters (as 1-kb-fragments) was previously verified by assaying reporter expression (Gehrig in preparation).

I amplified the *mef2d, dre-mir9-1* and *elp4* promoters as 500-bp fragments, and cloned into the CLGY vector (Ellingsen et al. 2005). After linearization by restriction digest, I injected these constructs to one-cell zebrafish embryos either alone or in combination with the isolated *shha arC* enhancer. The embryos were fixed at 24hpf stage, and were subject to antibody staining with a first antibody specific to GFP, but also recognizing YFP. The basal activity of these promoters were low and rather unspecific, but in all three cases the YFP expression was directed into the notochord upon the enhancer co-injection, suggesting that these promoters are able to interact with the *shh arC* enhancer. The *dre-mir9-1* promoter in combination with the *shha arC* not only gave notochord and hypothalamus expression, but the YFP was also present in spinal cord motor neurons (Figure 23.).

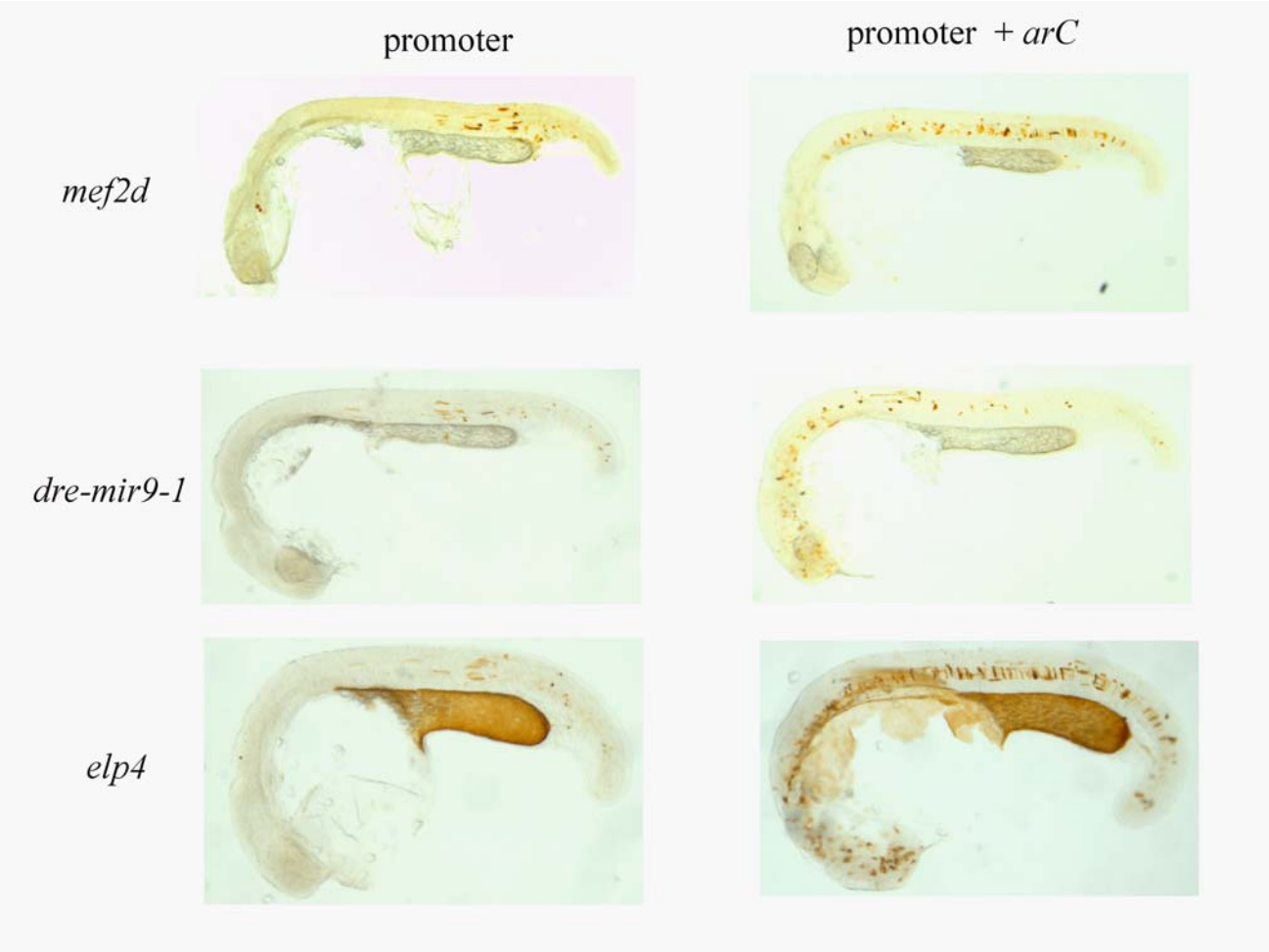


Figure 23: The expression pattern of the *mef2d*-, *dre-mir9-1*- and *elp4*-CLGY constructs
The isolated plasmids were injected either alone or with the isolated *shha arC* fragment. The embryos were subject to antibody staining.

I cloned the *hsp70, lmbr1l* and *eng2b* promoters into the Gateway vectors, and tested in combination with the control enhancer for basal activity. The *eng2b* core promoter gave some skin expression by itself, while no signal was detected in the embryos injected with the *hsp70* and *lmbr1l* promoters in combination with the control enhancer. The *hsp70* and the *eng2b* promoters were injected in combination with the *shha arC* enhancer, and both of them were activated in the notochord, floorplate and hypothalamus (Figure 24.). The *lmbr1l* promoter was tested with the *mnx1 regB* enhancer, and this construct showed YFP expression in the spinal cord, skin and muscle (Figure 24.).

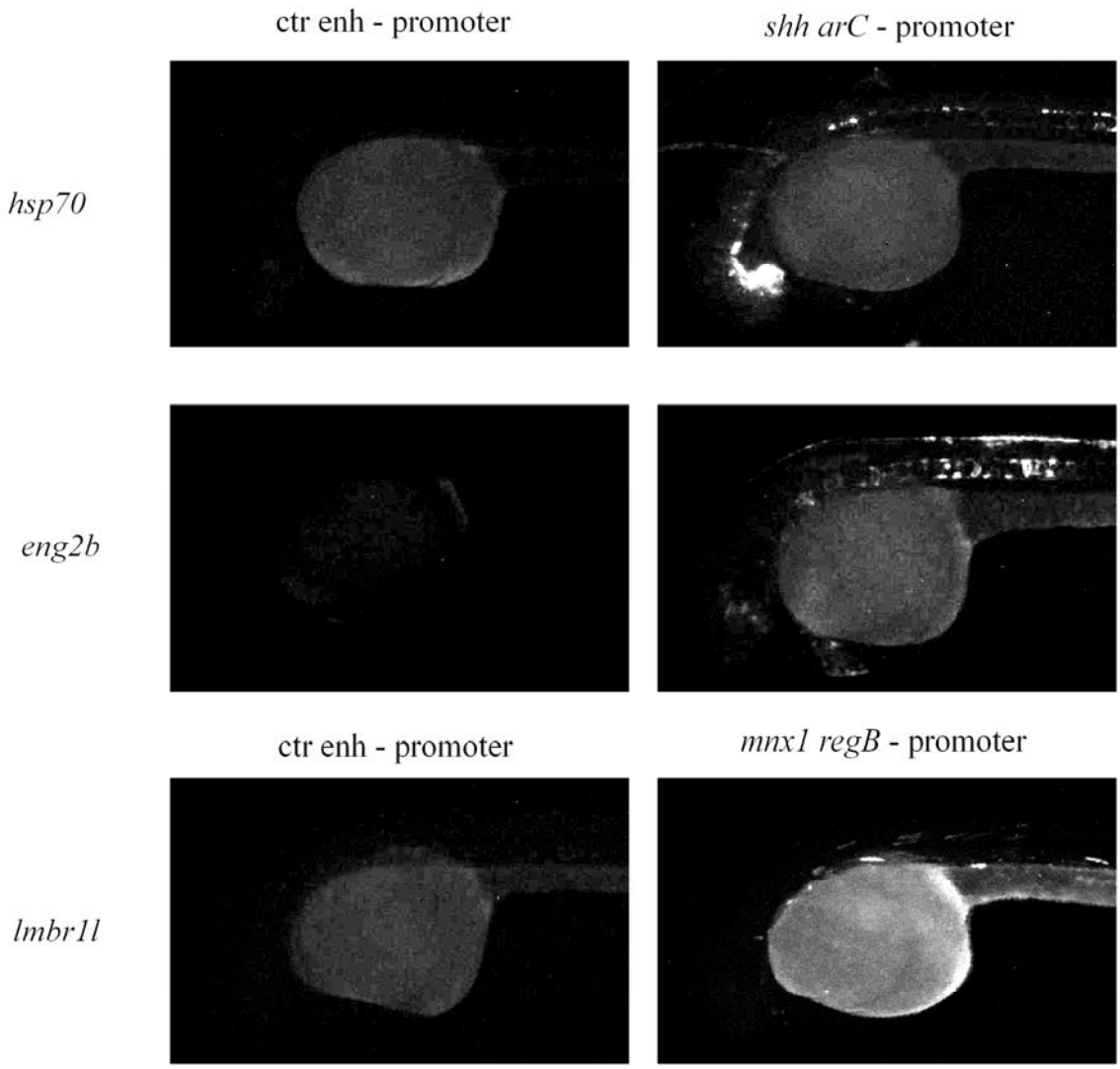


Figure 24: The activity of the *hsp70*, *eng2b* and *lmbr1l* promoters

The promoters were combined either with the control enhancer, or with the *shha arC* (*hsp70* and *eng2b*) or with the *mnx1 regB* enhancer (*lmbr1l*).

4.3.3 Multisite Gateway expression vectors do not affect the expression pattern of the transgene

We decided to use the Multisite Gateway vectors for the generation of the different enhancer-promoter combinations, as this system provides the possibility of cloning several fragments of interest into the same plasmids without using restriction enzymes (Hartley et al. 2000; Walhout et al. 2000). Once distinct fragments are cloned into the intermediate vectors called entry clones, they can be used for the generation of any possible combination in the second recombination reaction without any further modifications. (For the overview of the Multisite Gateway cloning, see the Materials and Methods chapter.)

I used a version of the Multisite Gateway system modified by the group of Patrick Lemaire (IBDM, Marseille, France), which was designed for *Ciona intestinalis* (Roure et al. 2007). Before starting the cloning I checked whether the expression pattern of a given promoter is altered when cloned into a Multisite Gateway destination vector. I cloned the 2.4kb fragment of the *shha* promoter into distinct sites of the Multisite Gateway destination vectors. I microinjected the generated expression vectors harbouring *shha* 2.4 promoter followed by the *Venus yfp* as a reporter gene to zebrafish embryos, and detected the expression patterns under epifluorescence microscope at 24hpf stage. Compared to a pCS2 vector containing the same *shha* promoter in front of a *gfp* reporter, both Multisite Gateway constructs gave the same results in respect of the ratio of the expressing per total embryo number (Table 10.), and there were no difference in the gained expression patterns (Figure 25.). The generated Multisite Gateway expression vectors did not disturb the transcription regulation compared to previously used vector.

construct	**injected embryos**	**expressing embryos**	**%**
shh2.4-gfp-pCS2	270	153	67,7
B3-*shh2.4*-B5::B1-*venus*-B2	235	123	52,3
R3-R5::B1-*shh2.4*-B2-*venus*	205	117	57,1

Table 10: Comparison of the expressing per injected embryo ratio for the different *shha2.4* constructs

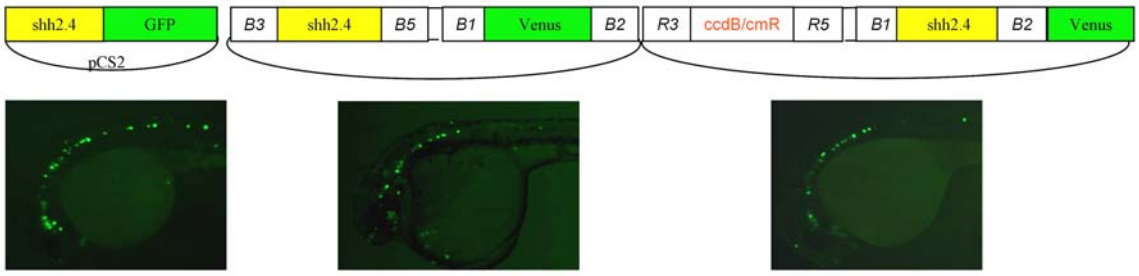


Figure 25: Comparison of the expression patterns of zebrafish embryos injected with different vectors containing the *shha* promoter.

Based upon these observations I have started to clone the isolated promoters and enhancers into the Multisite Gateway vectors system. I amplified the promoters and enhancers from zebrafish or fugu genomic DNA or from plasmids, and I used the PCR products containing the correct attachment sites and the p221-P1-P2 or p221-P3-P5 donor vectors for the first recombination reaction. I checked the generated entry clones by sequencing, and then used these to generate the 260 expression vectors (pSP72-B3-enhancer-B5::B1-promoter-B2-*venus*) in the second recombination reaction.

4.3.4 Pre-screen test with the *shha arc*-series

Before starting the high throughput screen, I performed a test-injection series with the *shh arC*-promoter constructs. I microinjected the circular expression vectors into zebrafish eggs, and checked the reporter expression at 30hpf embryonic stage. I counted the number of the expressing embryos, and checked the different domains specific to the *arC* enhancer: the notochord, floorplate and the hypothalamus. The results of the pre-screen test were promising: the *shha arC* enhancer was able to drive the expression of the *venus* into *shha*-specific domains in combination with ten promoters, while only three promoters (*tbp*, *tram1* and *c20orf45*) showed expression comparable with the control (Figure 26.A and B).

As Figure 26.B illustrates, the distribution of the reporter expression varied between the constructs containing different promoters. There were promoters that have been activated by the enhancer preferably in the notochord (*apoeb*), others in the floorplate and/or hypothalamus (*hsp70*, *klf4*), or in the case of the *krt4*, *ccne*, *shha* and *eng2b* promoters, the reporter expression was detected more or less equally in all *shha*-specific domains.

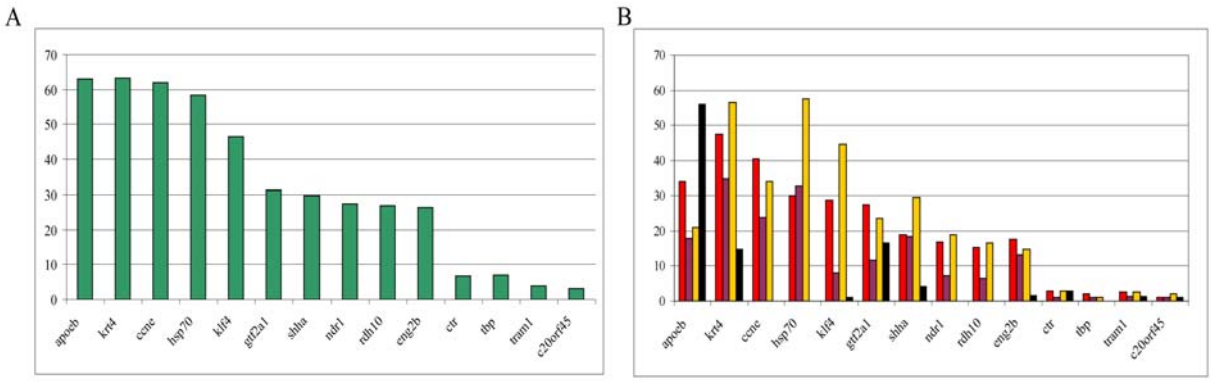


Figure 26: The YFP expression gained with the *shha arC* enhancer in combination with distinct promoters. A: Expressing per total embryo number ratio. B: Expressing per total embryo number ratios for distinct tissues. Red bars represent the notochord, pink bars the floorplate, yellow bars the hypothalamus and black ones the ectopic domains such as skin and muscle.

Regarding the ectopic expression in the skin and muscle, there were major differences between the promoters: while the majority of the constructs showed no or low level of ectopic activity, the *apoeb, krt4* and *gtf2a1* promoters were highly active in these tissues.

These results suggested us that there would be promoter-specific differences in the enhancer activities in the forthcoming high throughput screen.

4.3.5 The high throughput screening and the data analysis

Our laboratory team: Ferenc Müller, Jochen Gehrig, Marco Ferg, Yavor Hadzhiev, Andreas Zaucker, Simone Schindler, a guest researcher Chengyi Song and Nadine Gröbner participated in the high throughput screen. We microinjected all the 250 generated expression vectors into one-cell stage zebrafish embryos (200 eggs per construct), five to fifteen minutes after the eggs have been laid. We found that this time-window is crucial to gain high level of reporter expression. The injection solution also contained *cfp* mRNA to trace those embryos that were correctly injected; the CFP-negative embryos were manually selected out at 24hpf stage. Then we dechorionated the embryos and pipetted them into plates containing delves with 500μm in diameter in 92 wells (Figure 27.). At 30hpf stage, when all the enhancers should be active, we manually oriented the anesthetised embryos into the delves, and the embryos were subject to automated image acquisition with the kind help of Urban Liebel (ITG, FZK, Karlsruhe).

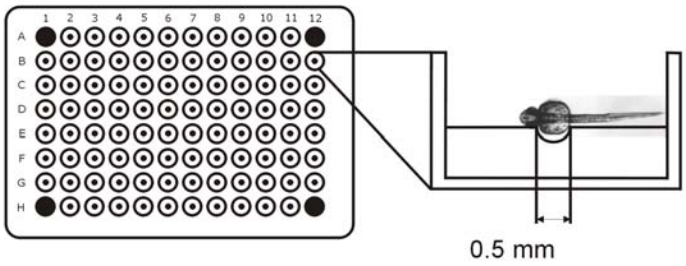


Figure 27: The 92-well agar-filled plate containing 500-nm diameter delves in the middle of each well.

Plates were handled by a Hamilton robotic arm, and central focal plane of the embryos was detected by an object detection auto-focus algorithm. Pictures were acquired from each embryo in 4 z-slices (55µm) in 3 channels: CFP, YFP and bright field. Markus Reischl (IAI, FZK, Karlsruhe) processed the gained pictures. Because the embryos were randomly oriented on the microscopic pictures, first they were to be registered and oriented. Then the 4 z-slices of a single embryo were projected into one picture with an extended focus algorithm (Figure 28.). To get the expression pattern of multiple embryos injected with the same construct, the extended focus pictures from embryos injected with one expression vector were projected into one composite expression picture (Figure 29.)

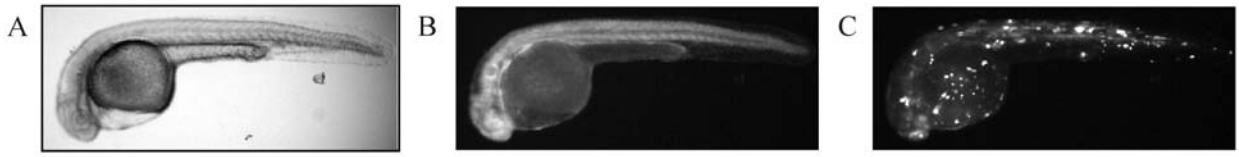


Figure 28: Extended focus pictures of an embryo injected with *shha arc-tram1* construct
A: bright field B: CFP C: YFP

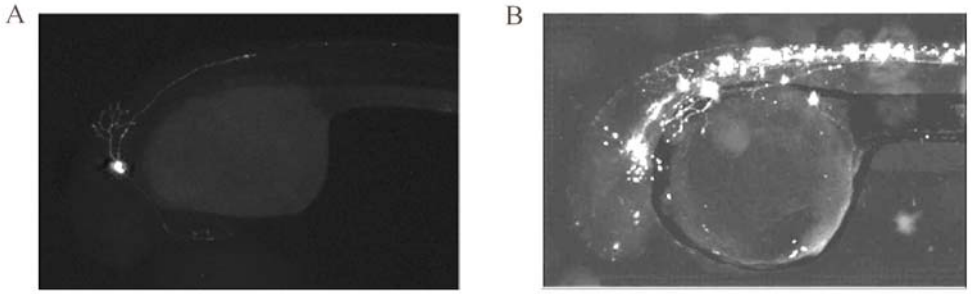


Figure 29: Generation of composite expression pictures in the example of the *isl1 zCREST2-eng2b* construct A: Venus signal of a single embryo. B: projection of approximately 60 embryos.

4.3.6 The activity and interactivity of promoters

The analysis of the extended focus pictures of several ten thousand embryos revealed that the ability of promoters to drive expression of the reporter and their responsiveness to enhancers varied in a wide range (Figure 30. and Table 11.). From the nineteen promoters seven (the *apoeb, krt4, atp6v1g1, klf4, gtf2a1, c20orf45* and *tbp*) showed background activity when tested with the control enhancer, and showed high expression in combination with several enhancers. On the other end of the scale stand the *pcbp2* (interacting only with *shha arC), dre-mir9-1 (eng2b CXE)* and *lmbr1l* (*dlx2b/dlx6a ei*) promoters with a single interacting enhancer. The control promoter did not show reporter expression in combination of any of the enhancers.

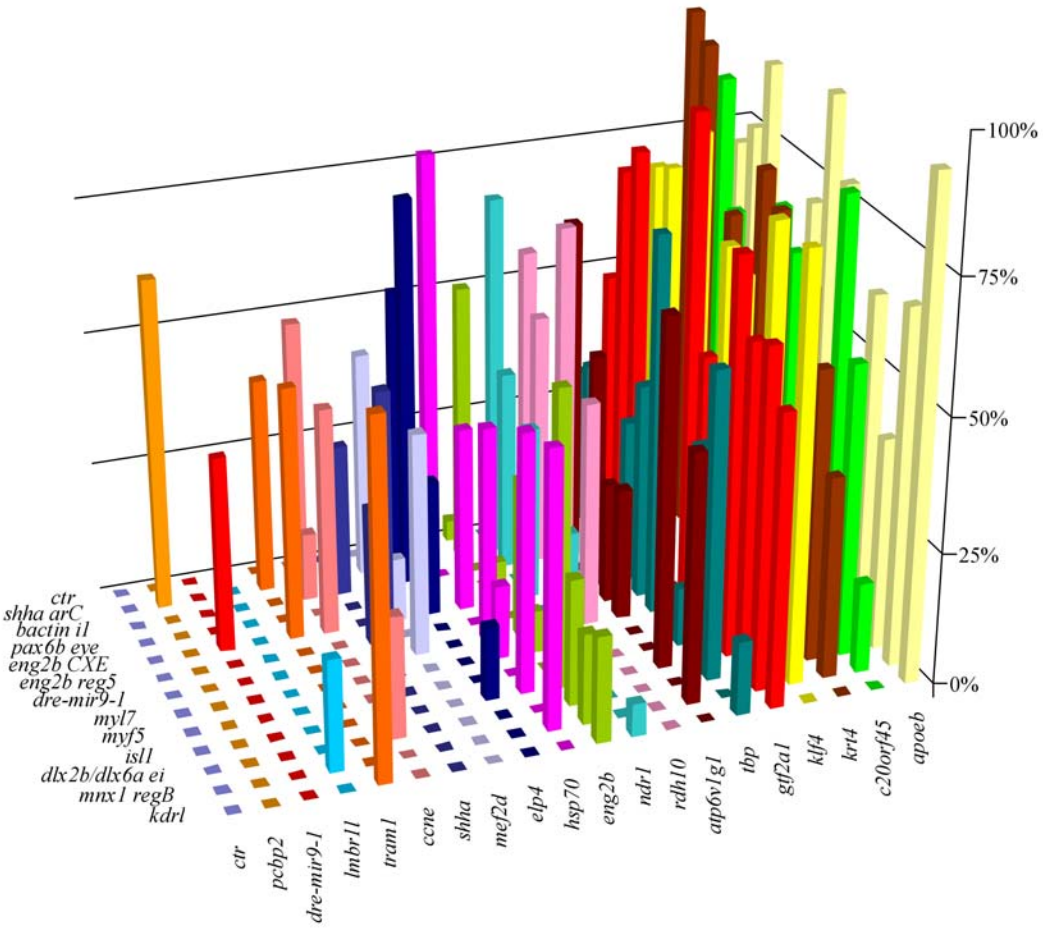


Figure 30: Expressing per total embryo number ratios for all the enhancer-promoter combinations

	ctr	*shha arC*	*eng2b CXE*	*eng2b reg5*	*isl1 zCREST2*	*dlx2b/ dlx6a ei*	*mnx1 regB*	*dre-mir9-1*	*ß-actin intron1*	*myl7*	*myf5*	*kdrl*
apoeb	+	+	+	+	+	-	+	+	+	+	-	+
atp6v1g1	+	+	+	+	+	-	+	+	+	-	-	-
gtf2a1	+	+	+	-	+	-	+	+	+	+	-	-
klf4	+	+	+	-	+	-	+	+	+	+	-	-
krt4	+	+	+	+	+	+	+	+	+	+	-	-
ndr1	-	+	+	-	-	-	-	+	+	+	-	+
pcbp2	-	+	-	-	-	-	-	-	-	-	-	
rdh10	-	+	+	-	-	-	-	-	+	+	-	-
tbp	+	+	-	+	+	+	NA	+	-	-	-	-
tram1	-	+	+	-	-	NA	NA	+	-	NA	-	-
c20orf45	+	+	+	+	+	-	-	+	-	+	+	-
ccne	-	+	+	-	-	+	-	-	+	+	-	-
shha	-	+	-	+	+	-	-	+	+	-	-	-
mef2d	-	+	+	-	-	-	-	-	+	-	-	-
dre-mir9-1	-	-	+	-	-	-	-	-	-	-	NA	-
elp4	-	+	+	NA	+	-	-	-	+	-	-	-
hsp70	-	+	+	-	+	-	+	+	-	+	-	-
lmbr1	-	NA	NA	NA	-	+	-	-	NA	-	-	-
eng2b	+	+	+	-	+	+	+	+	-	+	NA	-

Table 11: Summary of the promoter-enhancer interactions

The *apoeb* promoter showed a weak background activity in the central nervous system, including the brain, eye and spinal cord. The high level of expression driven by the enhancers came upon this background. This promoter was able to interact with the *shh arC, ß-actin intron1, kdrl, eng2b reg5, dre-mir9-1, myl7, isl1 zCREST2* and *mnx1 regB* enhancers (Figure 31.).

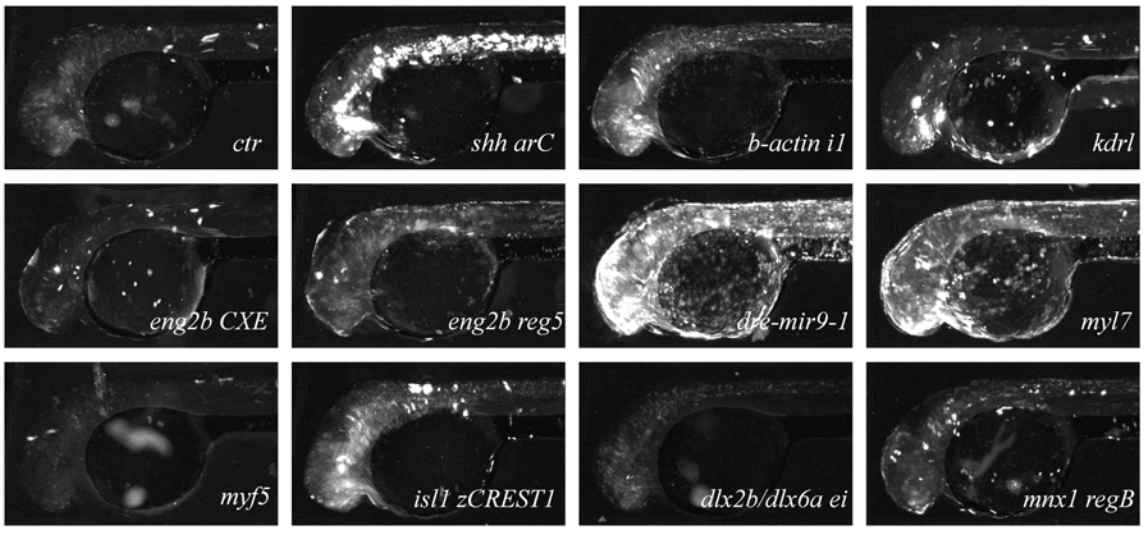


Figure 31: The expression patterns of the constructs containing the *apoeb* promoter.

The *atp6v1g1* promoter gave yolk expression when tested with the control enhancer. Combination of this element with the *ß-actin intron1*, *eng2 CXE*, *eng2b reg5*, *dre-mir9-1*, *isl1 zCREST2* and *mnx1 regB* enhancers resulted in enhancer-specific expression of the reporter. The *shh arC* enhancer activated the expression only in the muscle (Figure 32.).

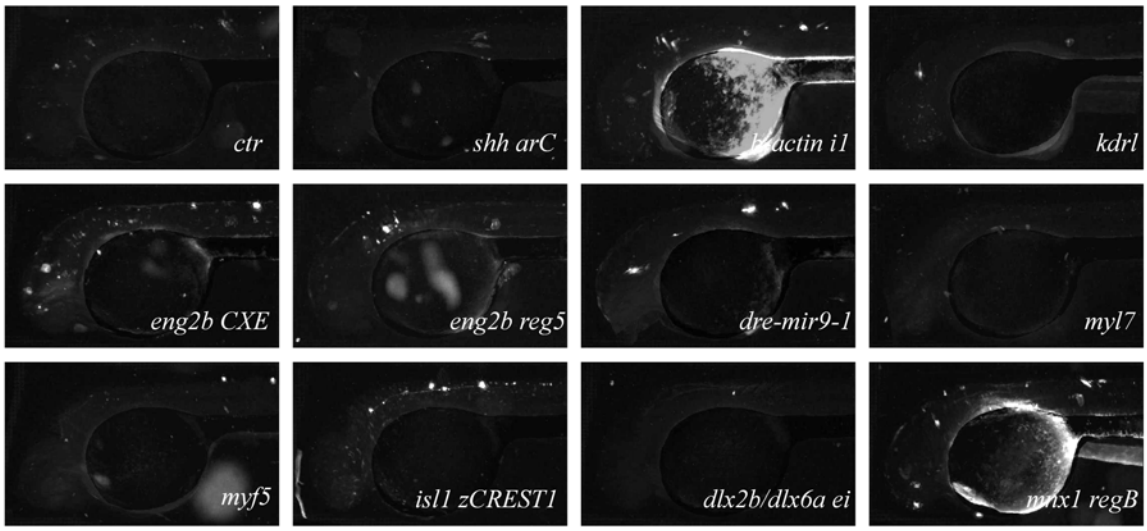


Figure 32: The expression patterns of the constructs containing the *atp6v1g1* promoter

The *gtf2a1* promoter showed background expression in the brain, eye and spinal cord, while the gene itself is expressed throughout the embryo. This promoter showed expression in the enhancer-specific domains with the following enhancers: *shh arC*, *ß-actin intron1*, *dre-mir9-1*, *myl7*, *isl1 zCREST2* and *mnx1 regB*. The *eng2b CXE* activated the promoter to drive expression in ectopic brain regions and in the notochord (Figure 33.).

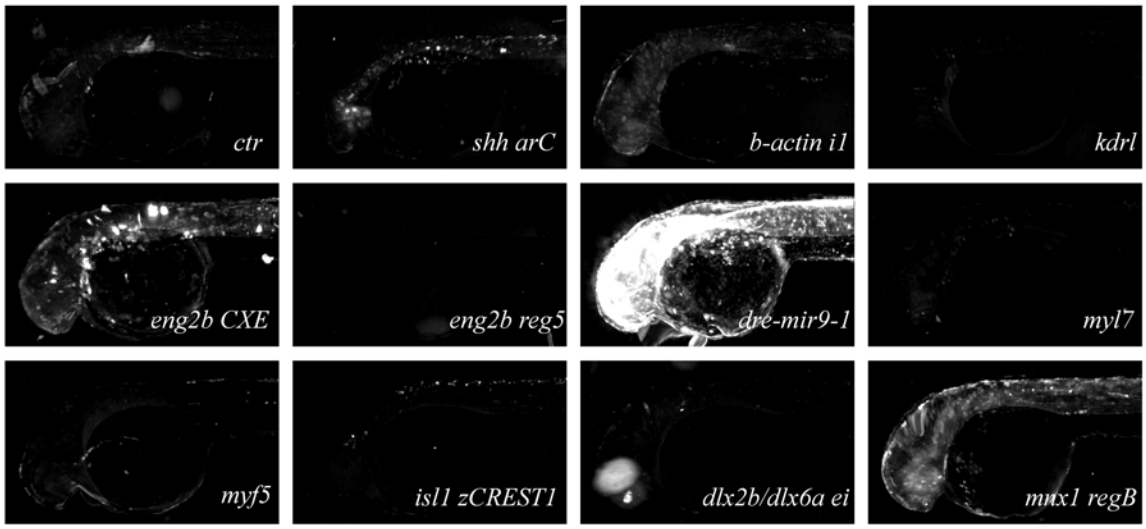


Figure 33: The expression patterns of the constructs containing the *gtf2a1* promoter

The *klf4* promoter showed a strong background activity in the skin, eye and brain. The skin expression was highly upregulated when it was combined with the *shh arC, ß-actin intron1, dre-mir9-1, myl7* and *mnx1 regB* enhancers. The *shh arC, eng2b CXE, myl7* and *mnx1 regB* enhancers drove the reporter expression to their specific domains as well (Figure 34.).

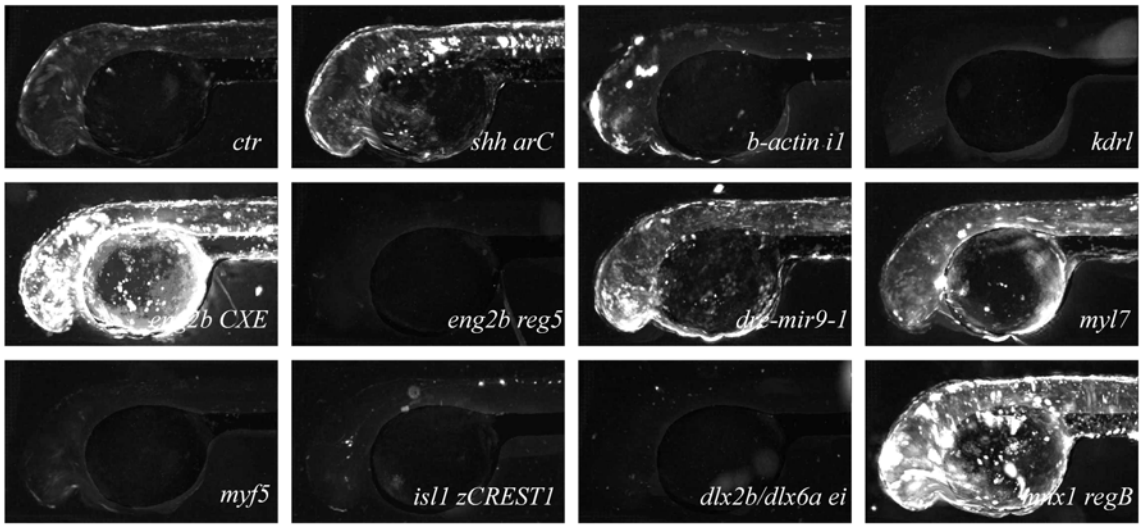


Figure 34: The expression patterns of the constructs containing the *klf4* promoter

The epidermis-specific *krt4* promoter showed skin, brain and spinal cord background expression. The combination of this promoter with the *shh arC* and *isl1 zCREST2* resulted in highly specific and strong *venus* expression, while the *ß-actin intron1, eng2b CXE, dre-mir9-1, dlx2a/dlx6b ei* and *mnx1 regB* enhancers boosted up the background activity as well (Figure35.).

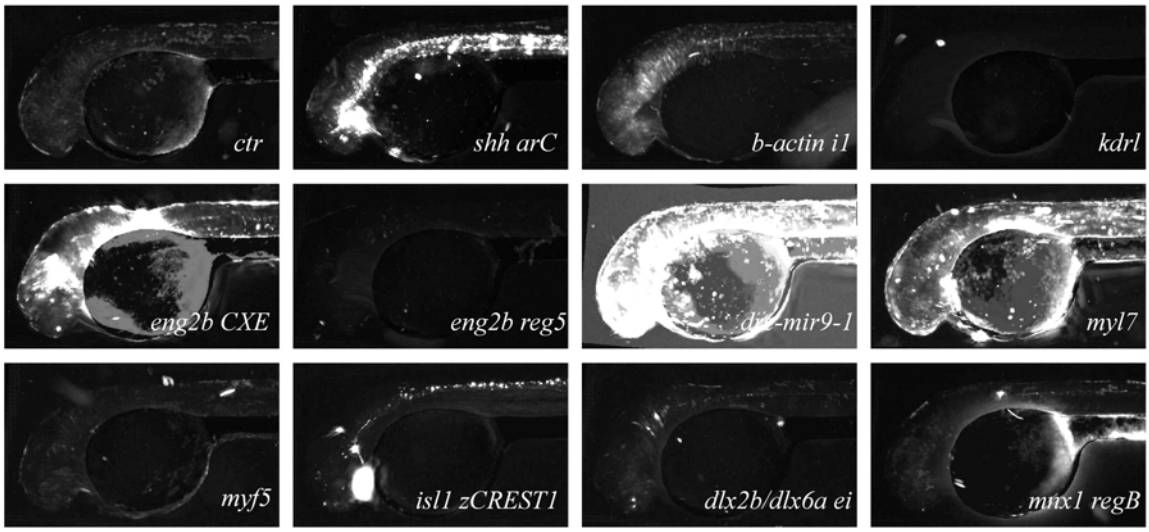


Figure 35: The expression patterns of the constructs containing the *krt4* promoter

ndr1 worked as a weak promoter, the expression driven to the specific domains by the *shh arC* and *myl7* enhancers was really faint. The *ß-actin intron1*, *dre-mir9-1* and *kdrl* enhancers slightly enhanced the Venus expression throughout the whole embryo, while the *eng2b CXE* enhancer activated strong expression in the brain, eye and skin domains (Figure 36.).

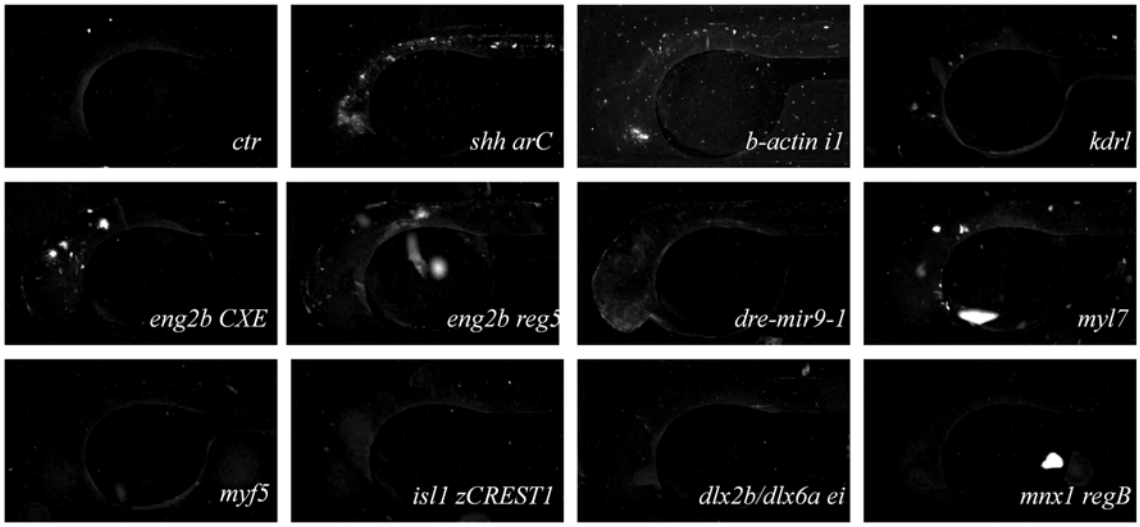


Figure 36: The expression patterns of the constructs containing the *ndr1* promoter

The *rdh10* promoter was able to interact with the *shh arC*, *ß-actin intron1*, *eng2b CXE* and *dre-mir9-1* enhancers. The *CXE* enhancer activated the reporter expression into the MHB and ectopic brain domains, skin and muscle, while the activation by the *shh arC* enhancer was highly specific (Figure 37.).

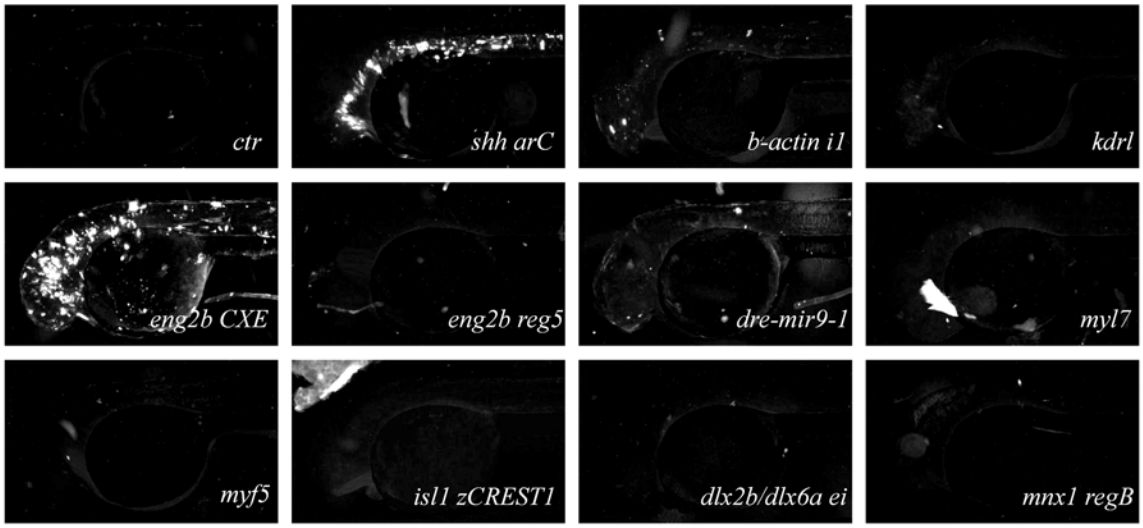


Figure 37: The expression patterns of the constructs containing the *rdh10* promoter

The *tbp* promoter in combination with the control enhancer gave some weak background activity in the CNS, skin, muscle and notochord. It worked as a weak promoter in combination with the *shh arC, dre-mir9-1* and *dlx2b/dlx6a ei* enhancers. The *isl1 zCREST2* and *eng2b CXE* enhancers activated ectopic expression when attached to this promoter (Figure 38.).

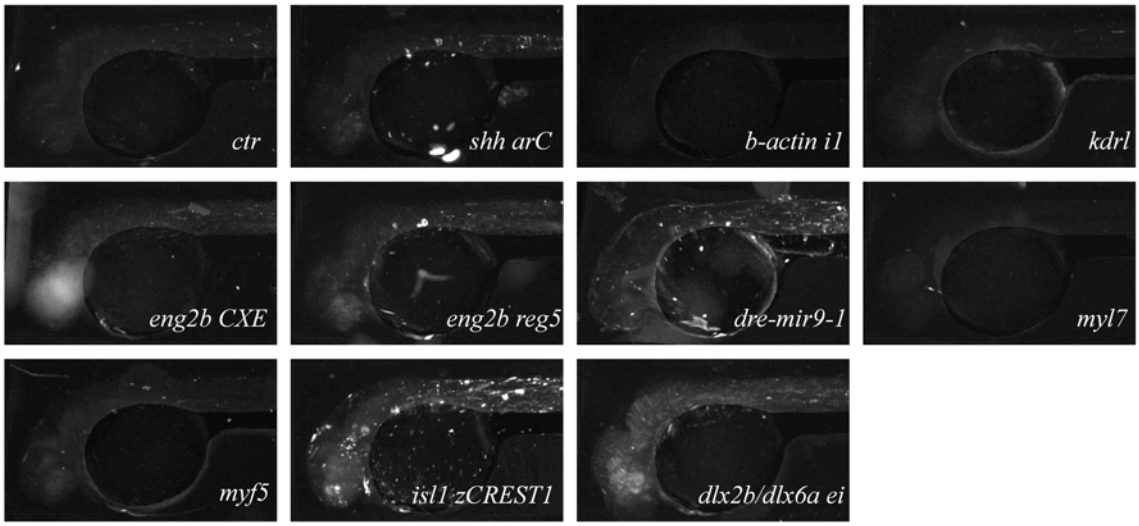


Figure 38: The expression patterns of the constructs containing the *tbp* promoter

c20orf45 was one of the strongest promoters, giving high level of background expression in the CNS, skin, muscle, notochord and hatching gland. This background activity was highly enhanced by the *shh arC, eng2b CXE and reg5, dre-mir9-1, myl7, myf5* and *isl1 zCREST2* enhancers (Figure 39.).

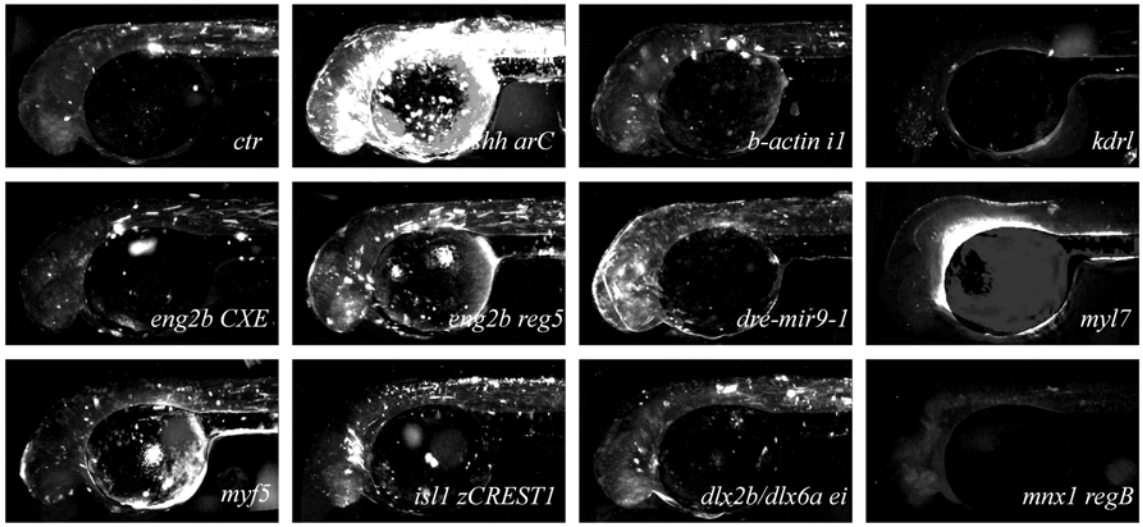


Figure 39: The expression patterns of the constructs containing the *c20orf45* promoter

The *shha* promoter was able to interact with the *shh arC, ß-actin intron1, eng2b reg5, dre-mir9-1* and *isl1 zCREST2* enhancers. The expression of these constructs was quite weak, but in the case of the *arC* and *zCREST2* enhancers the expression was highly specific (Figure 40.).

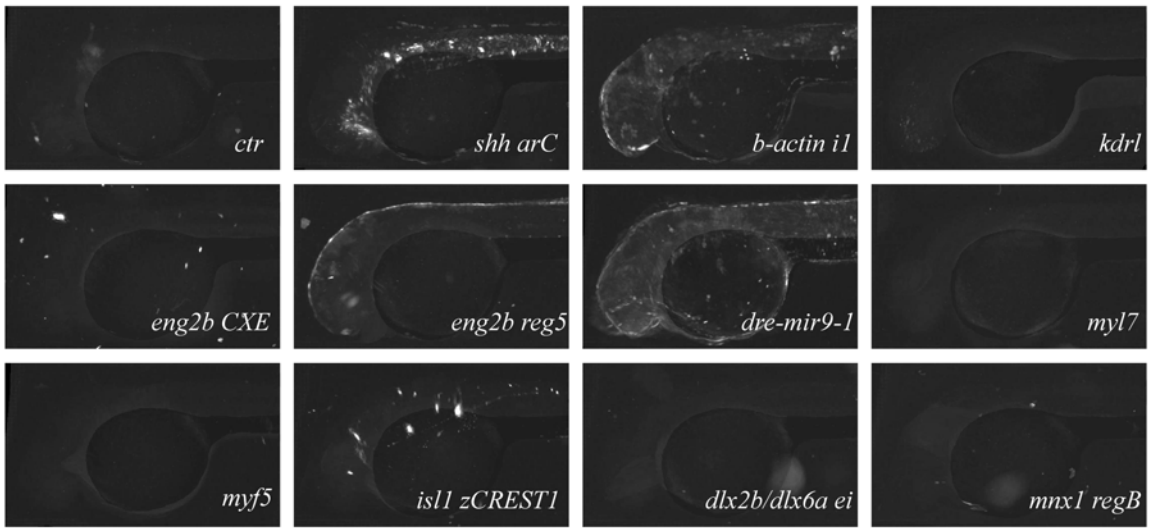


Figure 40: The expression patterns of the constructs containing the *shha* promoter

The *hsp70* promoter was activated by the *shh arC, eng2b CXE, dre-mir9-1, myl7, isl1 zCREST2, dlx2b/dlx6a ei* and *mnx1 regB* enhancers. In the cases of the *shh arC, eng2b CXE, myl7, isl1 zCREST2* and *mnx1 regB* enhancers the expression was highly specific (Figure 41.).

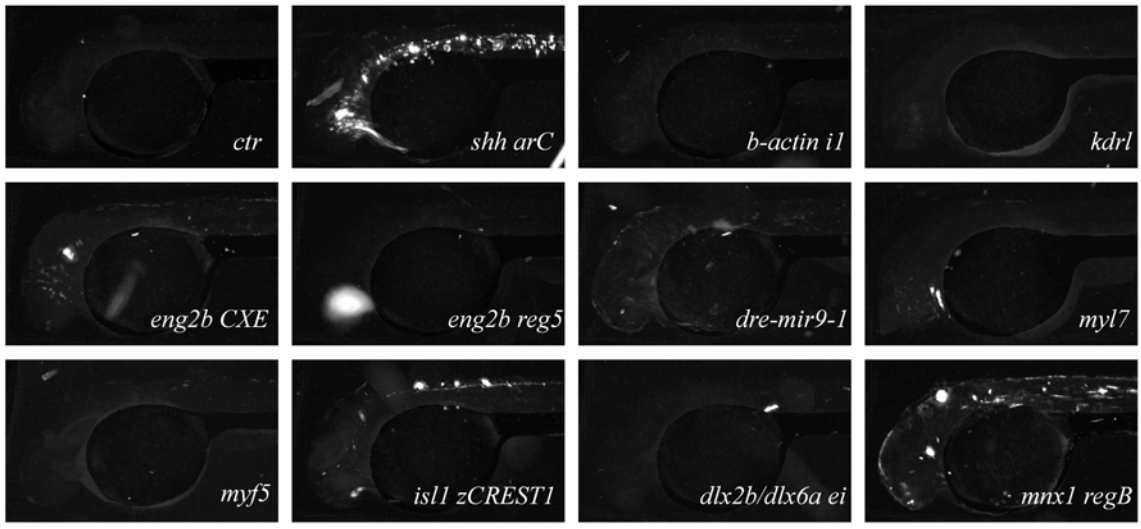


Figure 41: The expression patterns of the constructs containing the *gtf2a1* promoter

A weak skin background expression was detected in embryos injected with the *eng2b-ctr enhancer* construct. The reporter expression was enhanced by the *shh arC,*

eng2b CXE, *dre-mir9-1*, *myl7*, *zCREST2*, *dlx2b/dlx6a ei* and *mnx1 regB* enhancers, but the *CXE* enhancer was not able to direct the expression into the MHB (Figure 42.)

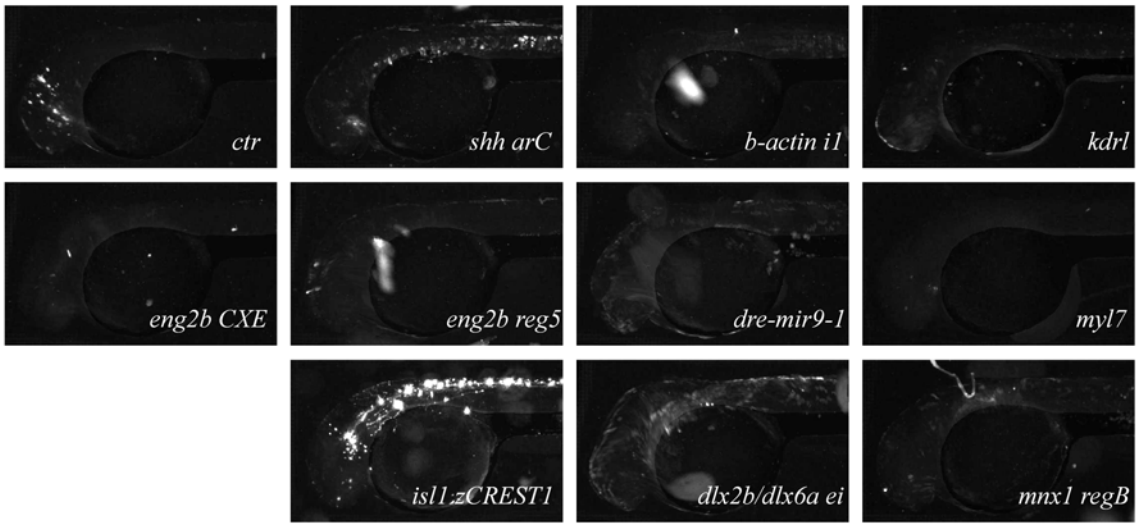


Figure 42: The expression patterns of the constructs containing the *eng2b* promoter

The *tram1*, *ccne*, *mef2d* and *elp4* promoters showed weak, generally ectopic activity in combination with several enhancers (Table 14.).

4.3.7 Promoter-specific differences in enhancer activity and strength

Based on my analysis of the extended focus pictures for each construct, from the twelve enhancers seven (the CNS-specific *shh arC, eng2b CXE and reg5, isl1 zCREST2* and *mnx1 regB*, the general β-*actin intron1* element and the heart-specific *myl7* enhancer) directed the reporter expression into the previously described domains; two (*dre-mir9-1* and *dlx2b/dlx6a ei*) worked as general enhancers rather than CNS-specific (Figure 30. and Table 14.). The eye enhancer of the *pax6b* gene did not show enhancer activity at all, while the somite enhancer of the *myf5* and the enhancer element driving the *kdrl1* gene to blood vessel endothel directed the reporter expression into ectopic domains, and in combination with only two promoters.

The *shh arC* enhancer activated the reporter expression in the notochord, floorplate and in the hypothalamus in most of the tested promoter combinations, while *shh arC-atp6v1g1* showed some muscle expression in few embryos, while injection of the *shh arC-dre-mir9-1* construct resulted in no YFP signal (Figure 43.).

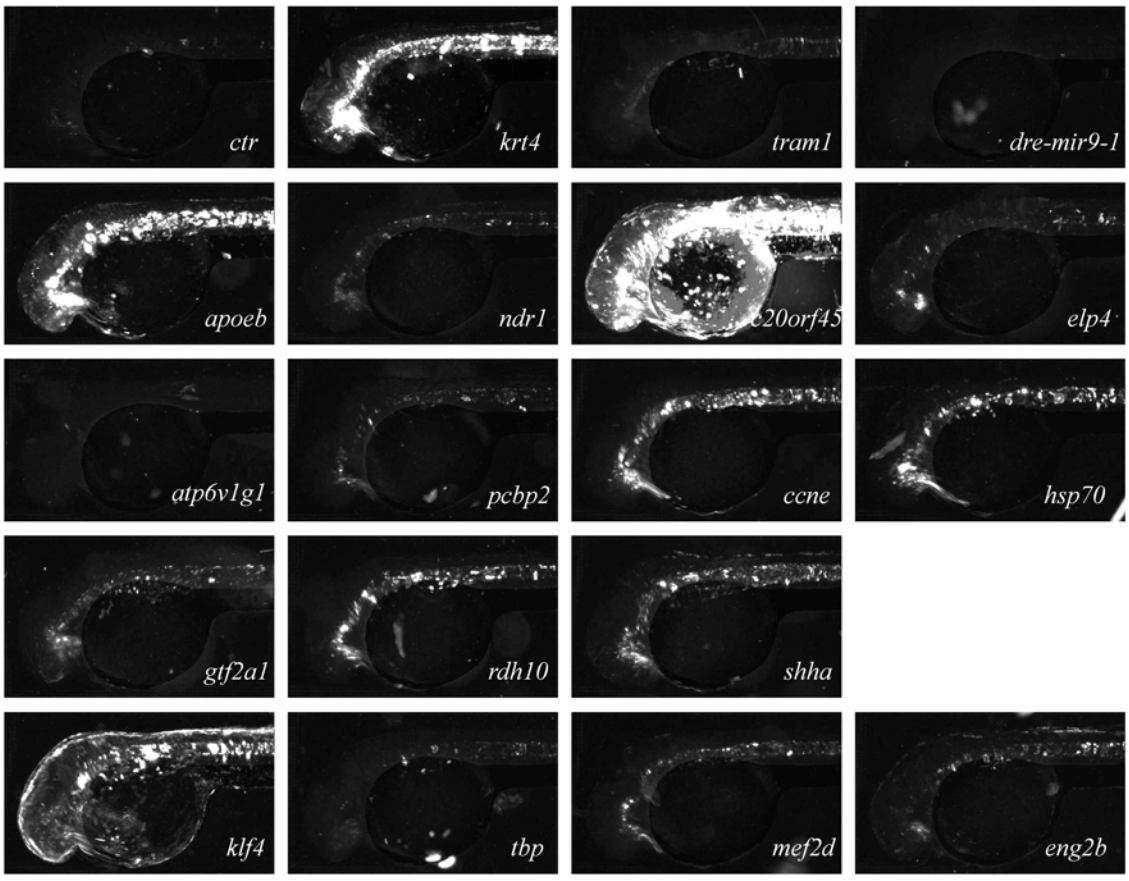


Figure 43: The projected expression maps of the embryos injected with the *shha arC* constructs

The general β-*actin* enhancer enhanced the background activity of the *apoeb, gtf2a, klf4, krt4, tbp* and *c20orf45* promoters and in general, increased expression could be observed in the skin and muscle. While the *atp6v1g1* promoter only had some yolk expression by itself, its combination with the β-*actin* intronic enhancer resulted in expression in novel domains: in the skin, muscle, notochord and in the CNS. In case of the *ndr1, rdh10, ccne, shha* and *elp4* promoters, where there was no detectable background, the enhancer activated the reporter expression in the brain, retina, skin and in some cases in the muscle. No expression was detected in the combination with the *pcbp2, tbp, tram1, mef2d, dre-mir9-1, hsp70* and *eng2b* promoters (Figure 44.).

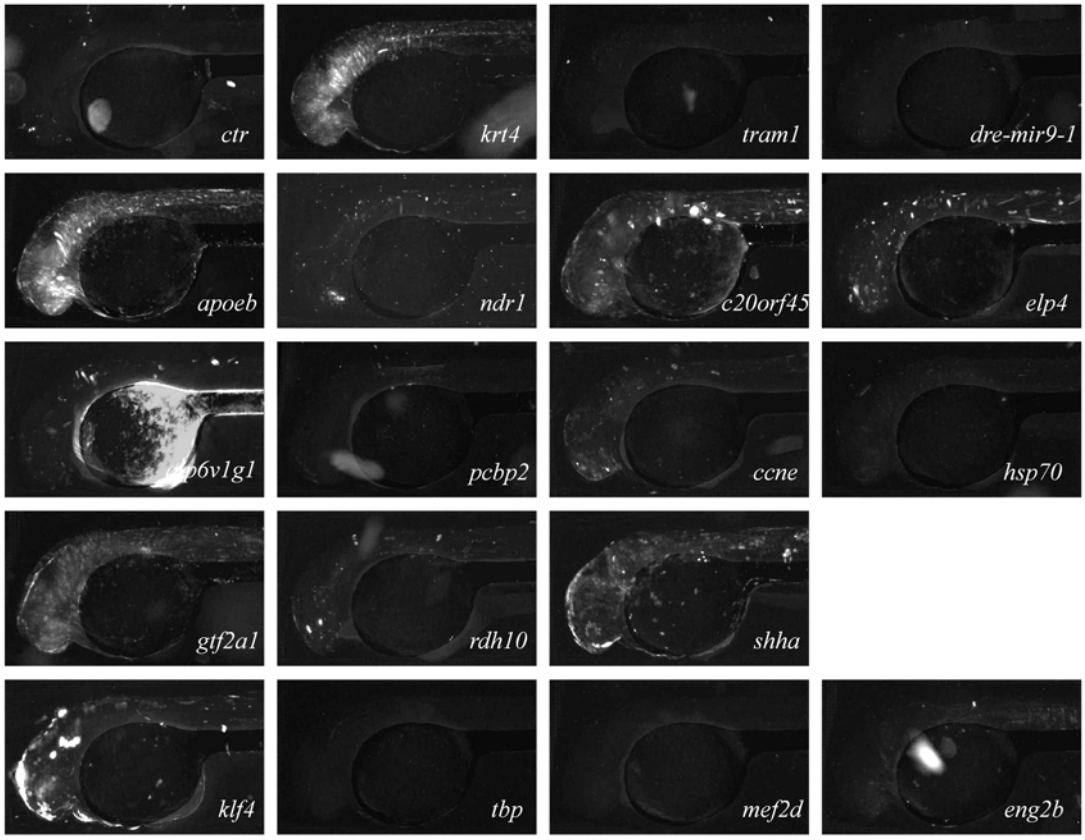


Figure 44: The projected expression maps of the embryos injected with the β-*actin intron1* constructs

The two *eng2b* enhancers directed the *venus* expression into the midbrain hindbrain boundary (MHB) in combination with distinct promoters (Figure 45. and 46.).

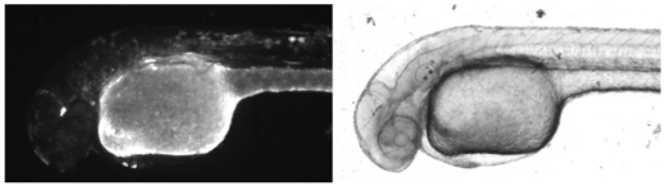


Figure 45: Pictures of an embryo injected with the *eng2b CXE-klf4* construct, showing *venus* expression in the MHB. Left side: extended focus picture from the YFP channel, right side: bright field picture of the same embryo.

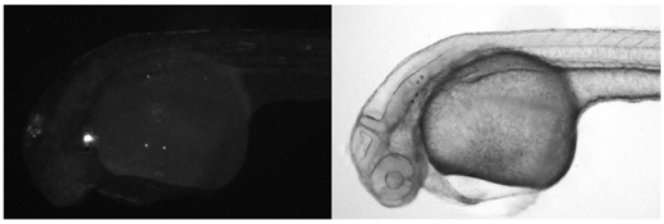


Figure 46: Pictures of an embryo injected with the *eng2b reg5-c20orf45* construct, showing *venus* expression in the MHB. Left side: extended focus picture from the YFP channel, right side: bright field picture of the same embryo.

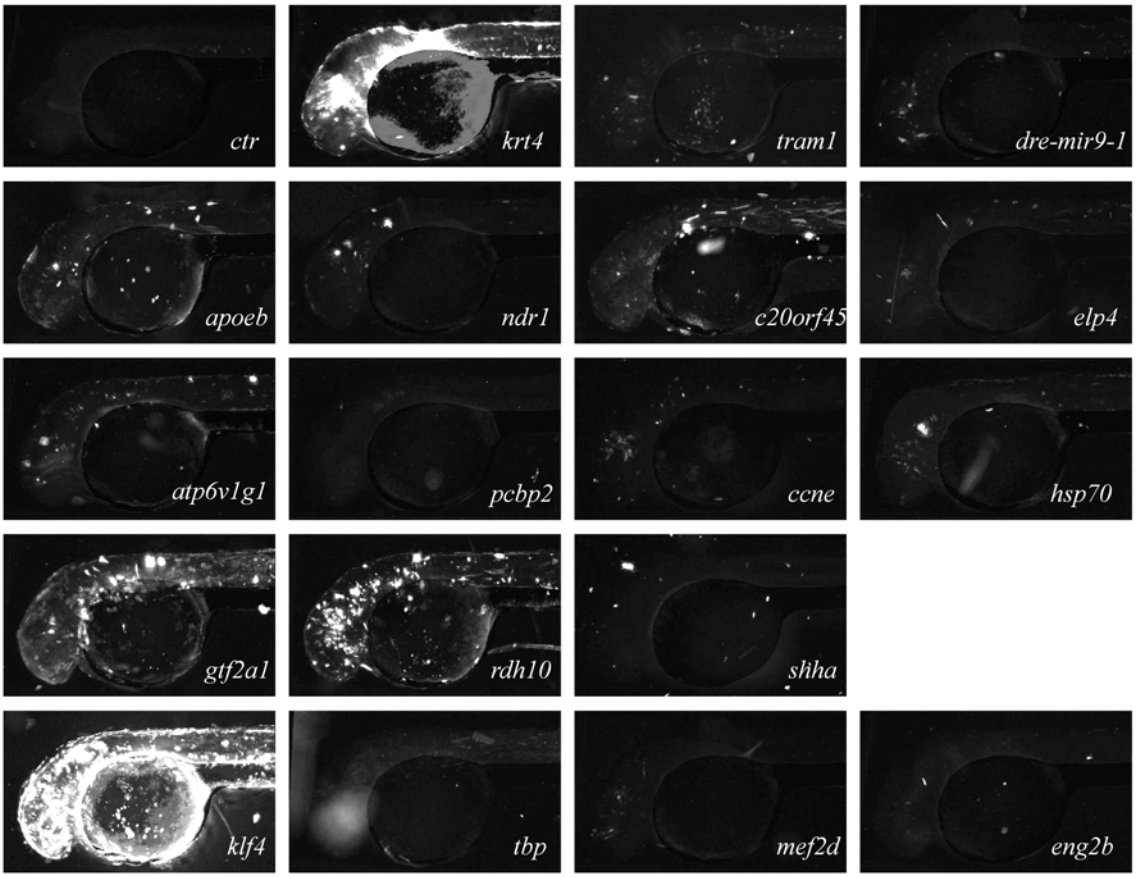


Figure 47: The projected expression maps of the embryos injected with the *eng2b CXE* constructs

The *CXE* enhancer, which was described in mouse, was able to interact with the *atp6v1g1, klf4, krt4, rdh10, c20orf40, ccne, mef2d, dre-mir9-1, elp4* and *hsp70* promoters to drive expression into the MHB, while the interaction with the *tram1* and the endogenous *eng2b* core promoters resulted in general brain and ectopic skin expression of the reporter. The *ccne* promoter, which did not show any background expression, was activated in brain regions other then the MHB and in other tissues as well, such as retina, spinal cord, epithel and muscle. These ectopic expression domains were observed with the MHB-specific constructs as well (Figure 47.).

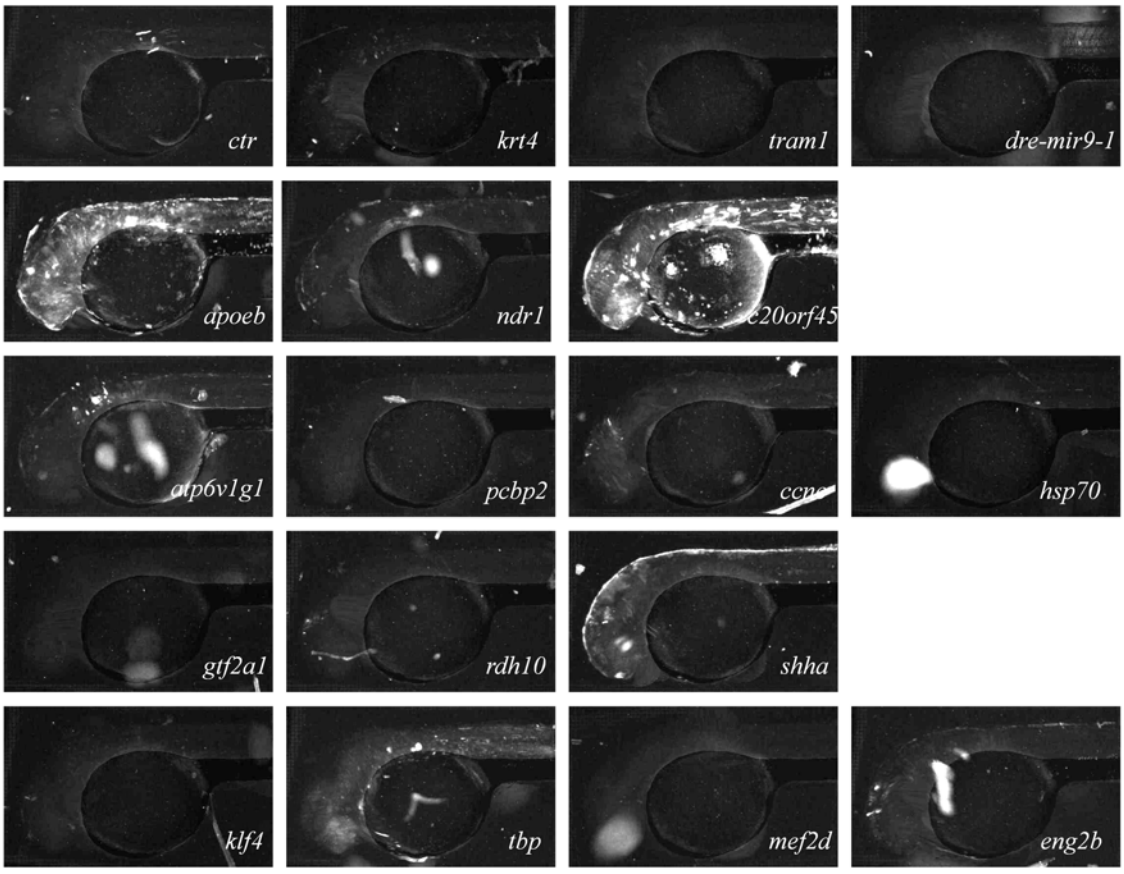


Figure 48: The projected expression maps of the embryos injected with the *eng2b reg5* constructs

The *eng2b CXE-pcbp2* and *–shha* combinations showed no expression. The *reg5* enhancer, which was identified as a conserved intronic element, activated expression in combination with fewer promoters, compared to the *CXE*. No expression was detected with the *gtf2a1, klf4, ndr1, pcbp2, rdh10, tram1, ccne, mef2d, dre-mir9-1, hsp70* and *eng2*b core promoters, although the *gtf2a, klf4* and *eng2b* promoters have background activity. The *reg5* enhancer directed the *venus* expression

into the MHB in combination with the *apoeb*, *atp6v1g1*, *krt4*, *c20orf45* and *shha* promoters (Figure 48.), from which the *atp6v1g1*, *krt4* and *c20orf45* promoters gave MHB-specific expression with the *CXE* enhancer, while none of them gave expression with *pcbp2* promoter.

The *dre-mir9-1* enhancer element, identified as a brain enhancer in an enhancer trap experiment did not show expression in distinct brain domains, but rather worked as a general enhancer in our system. In combination with the a*poeb*, *atp6v1g1*, *klf4*, *krt4*, *tbp* and *eng2b* promoters, it gave ectopic skin, yolk, muscle and neural expression upon the enhanced background activity, while with the *ndr1*, *rdh10*, *shha*, *mef2d* and *hsp70* promoters, where no background was detected, the *dre-mir9-1* enhancer activated the expression in the brain, retina, spinal cord, skin and muscle (Figure 49.).

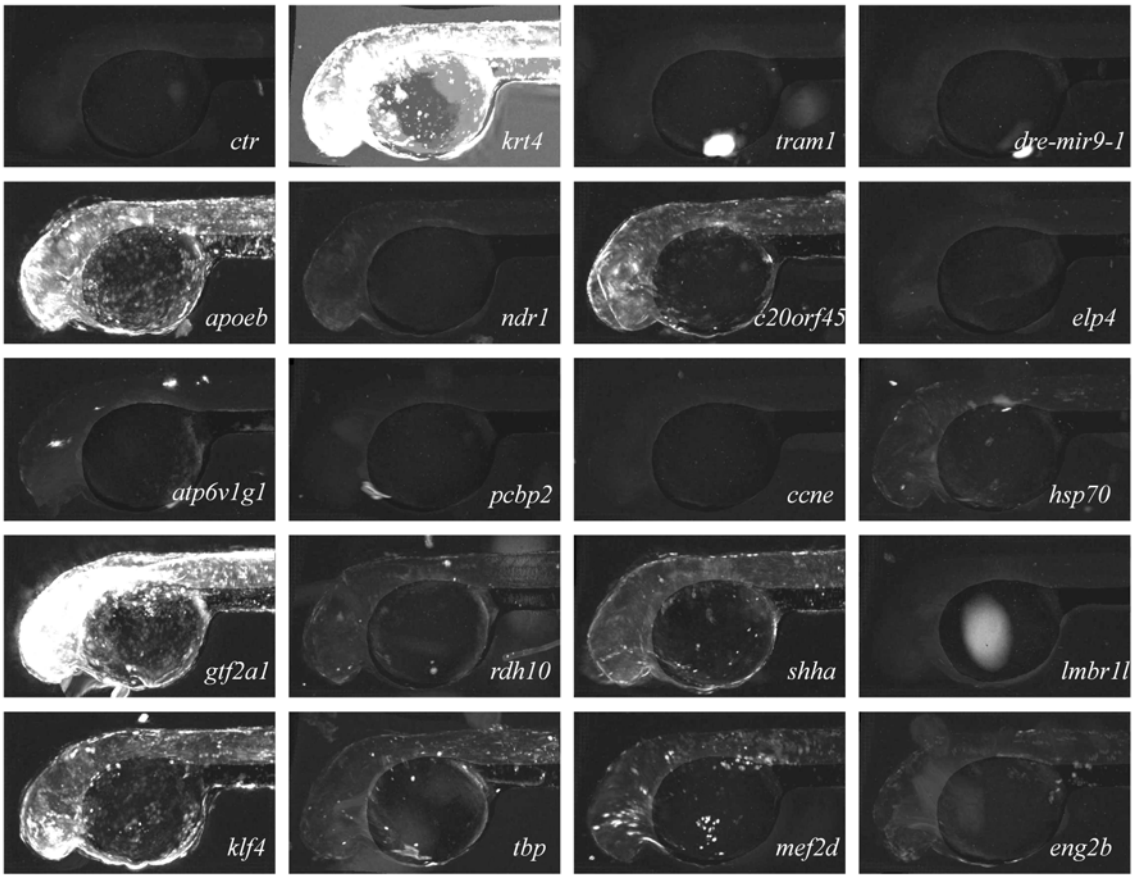


Figure 49: The projected expression maps of the embryos injected with the *dre-mir9-1* constructs

Venus signal in the developing heart tube (Figure 50.) was observed in the following promoters in combination with the *myl7* enhancer: *apoeb*, *gtf2a*, *klf4*, *krt4*, *ndr1*, *hsp70* and *eng2b*, but only in few embryos in the case of the last three promoters. Interestingly no expression at all was observed in the embryos injected with al the other constructs (Figure 51.).

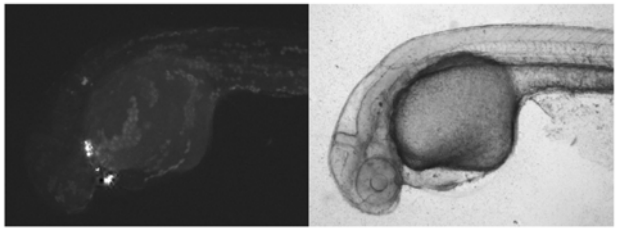


Figure 50: Pictures of an embryo injected with the *myl7-krt4* construct, showing *venus* expression in the MHB. Left side: extended focus picture from the YFP channel, right side: bright field picture

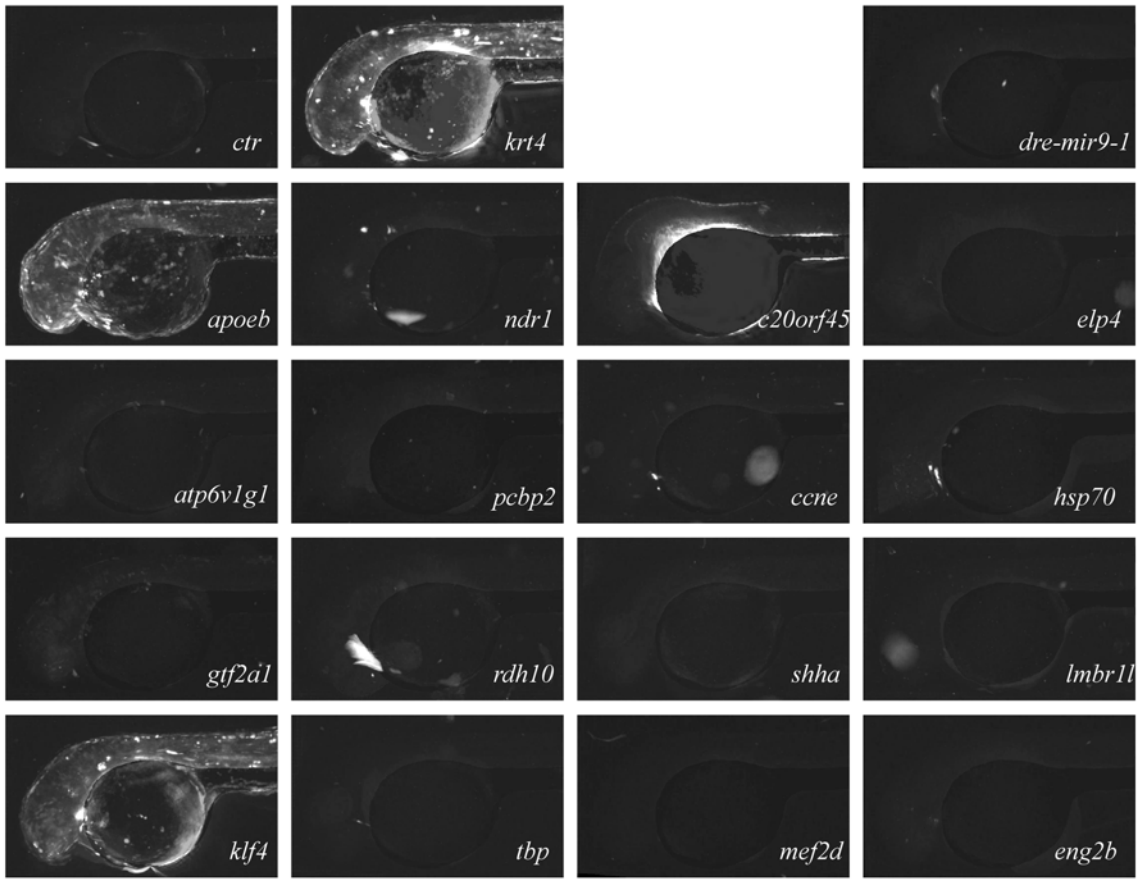


Figure 51: The projected expression maps of the embryos injected with the *myl7* constructs

The somite-specific *myf5* enhancer was not able to direct the expression specifically into the developing muscle, nor enhance the background activity of the promoters, and no vascular epithel-specific Venus expression was observed with the *kdrl* enhancer.

The *isl1 zCREST2* enhancer directed the reporter expression into motor and sensory neurons of the brain and spinal cord in combination with the *atp6v1g1*, *gtf2a*, *klf4*, *krt4*, *c20orf45* and *eng2b* promoters, while it activated the *venus* expression in ectopic domains when combined to *elp4* and *hsp70* promoters. No Venus signal was detected in embryos injected with all the other constructs (Figure 52.).

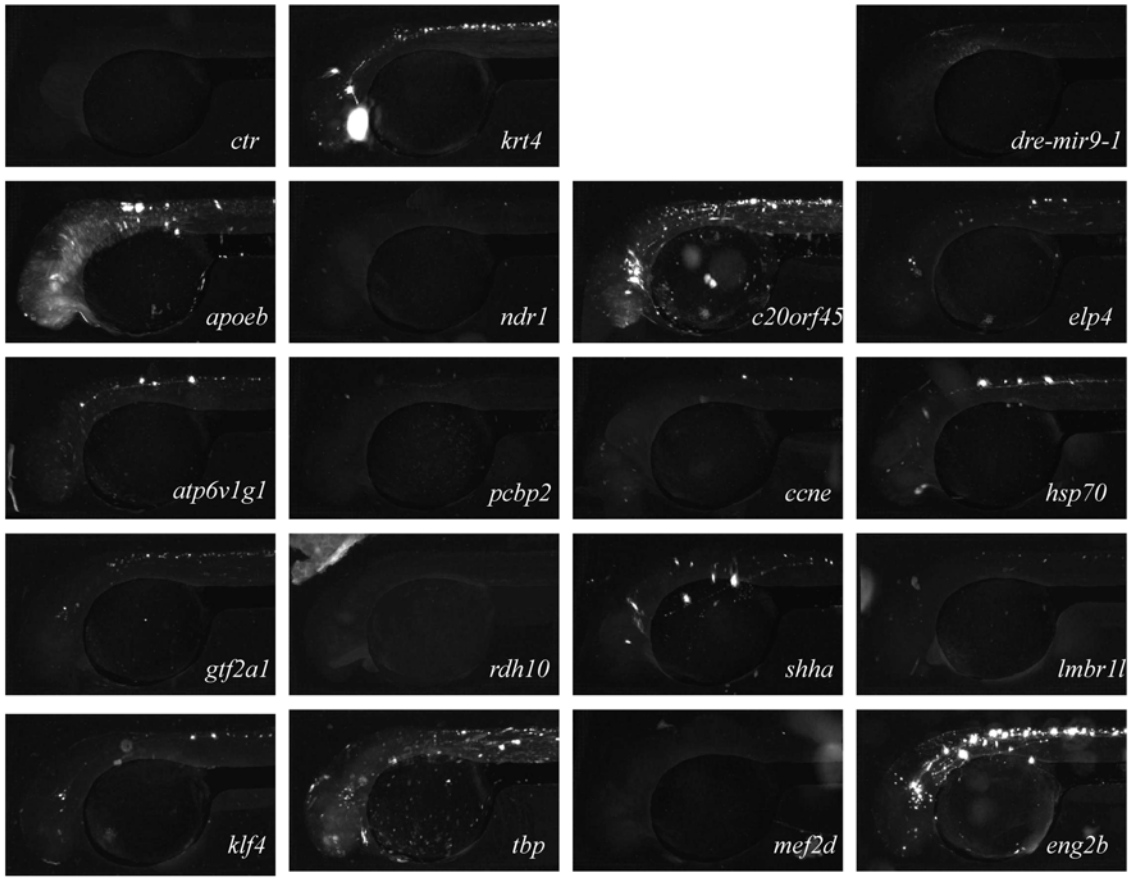


Figure 52: The projected expression maps of the embryos injected with the *isl1 zCREST2* constructs

No forebrain-specific reporter expression was observed in embryos injected with expression vectors containing the *dlx2b/dlx6 ei* enhancer. It could only activate the *ccne* promoter in the brain, retina, muscle and epithel, the strength and the specificity of the expression gained in combination with the *apoeb*, *gtf2a*, *krt4* and *tbp* promoters were comparable with their background activity.

The motor neuron specific *mnx1 regB* enhancer was able to direct the *venus* expression into the spinal cord (Figure 53.) in combination with the *apoeb*, *atp6v1g1*, *gtf2a* and *klf4* promoters, while it showed a non-specific activator effect when combined to the *hsp70*, *lmbr1l* and *eng2b* promoters (Figure 54.).

Figure 53: Pictures of an embryo injected with the *myl7-krt4* construct, showing *venus* expression in the MHB. Left side: extended focus picture from the YFP channel, right side: bright field picture of the same embryo.

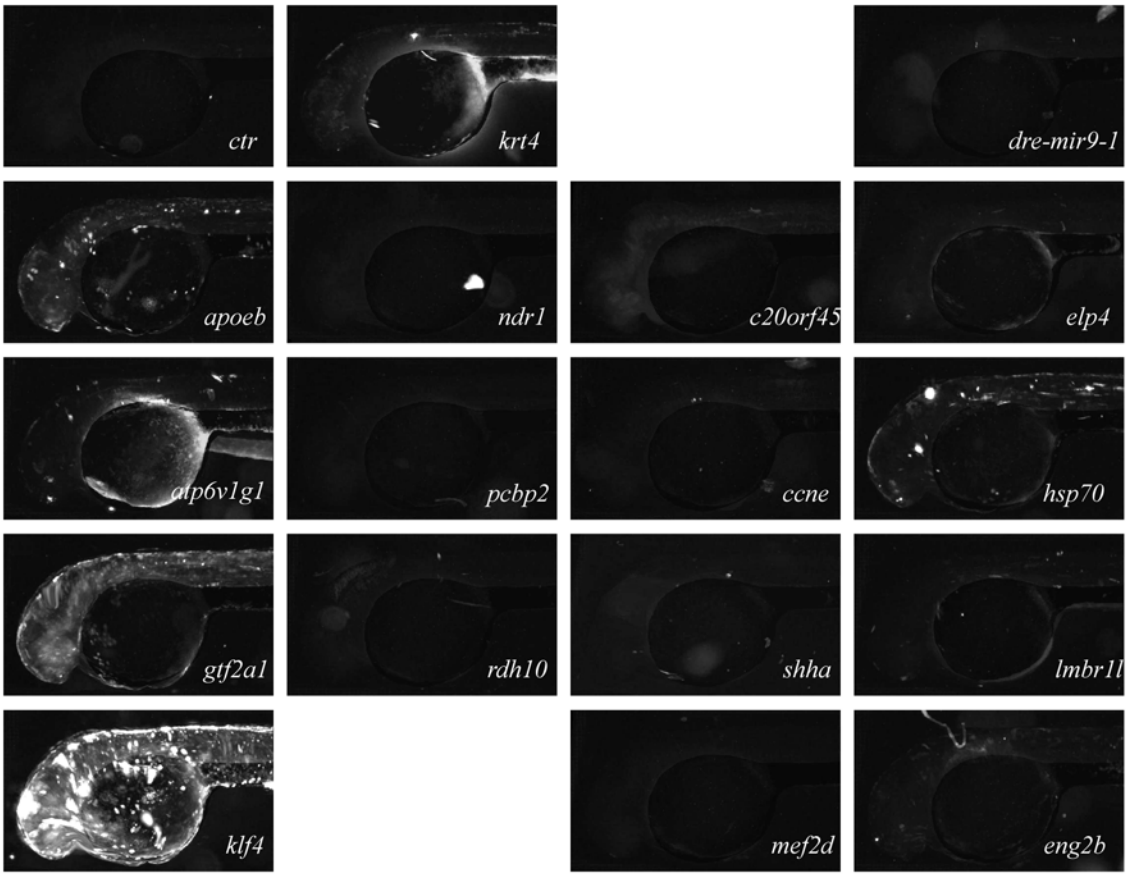


Figure 54: The projected expression maps of the embryos injected with the *mnx1 regB* constructs

4.3.8 Quantification of the expression

To generate numeric data from the expression patterns, embryo structures were detected on the extended focus pictures and the fluorescence was measured in the following domains (for domain definition see Materials and Methods point 4.6.3): eye, midbrain hindbrain boundary, brain, spinal cord, heart, notochord, skin, yolk and yolk plug. The Venus signal originating from the yolk structures were removed from the overlay analysis. To obtain a quantitative readout of reporter activity, the mean of Venus expression in the total number of embryos was calculated for each tissue domains as well as for the whole embryo. The mean of the fluorescent pixels detected within a tissue domain were normalised to the size proportion of the domain as compared to the whole of the embryo and the normalised values were expressed in a chart and expressed in colour intensity codes. The brightness of a square in the colour code system was used to indicate the extent of expression as measured by normalised pixel intensity counting (Figure 55.). The complete results of the quantification of the entire experiment are shown in Figure 56.

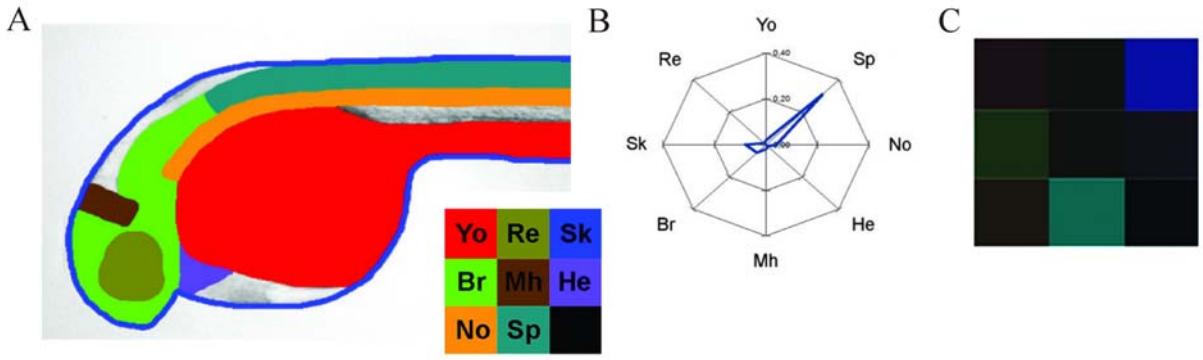


Figure 55: Quantification of the expression was performed in distinct domains of the zebrafish embryo

A: The colour code of the defined tissue domains: red – yolk and yolk plug (Yo), greyish green – retina (Re), dark blue – skin (Sk), light green – brain (Br), brown – MHB (Mh), blue – heart (He), orange – notochord (No), bluish green – spinal cord (Sp). B: The quantification of the projection picture shows that the majority of the signal is located in the spinal cord, and there were some Venus-positive cells in the brain and skin domains. C: The colour code representation of the signal strength in the different domains.

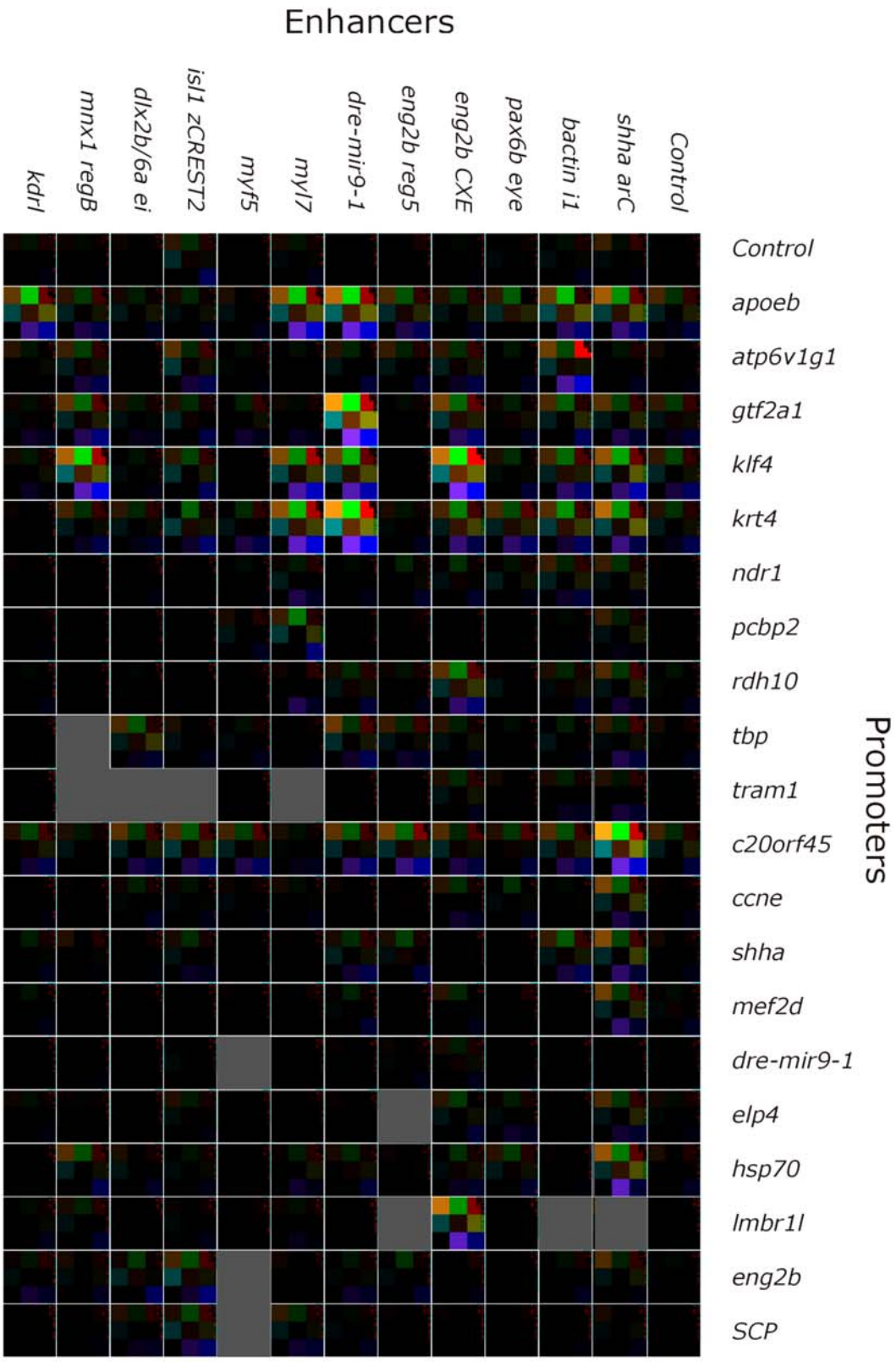


Figure 56: Overview of the analysis of 250 reporter constructs. Colour intensities represent pixel counts per embryo for each domain as described in Figure 29**.** Grey boxes indicate constructs not assayed.

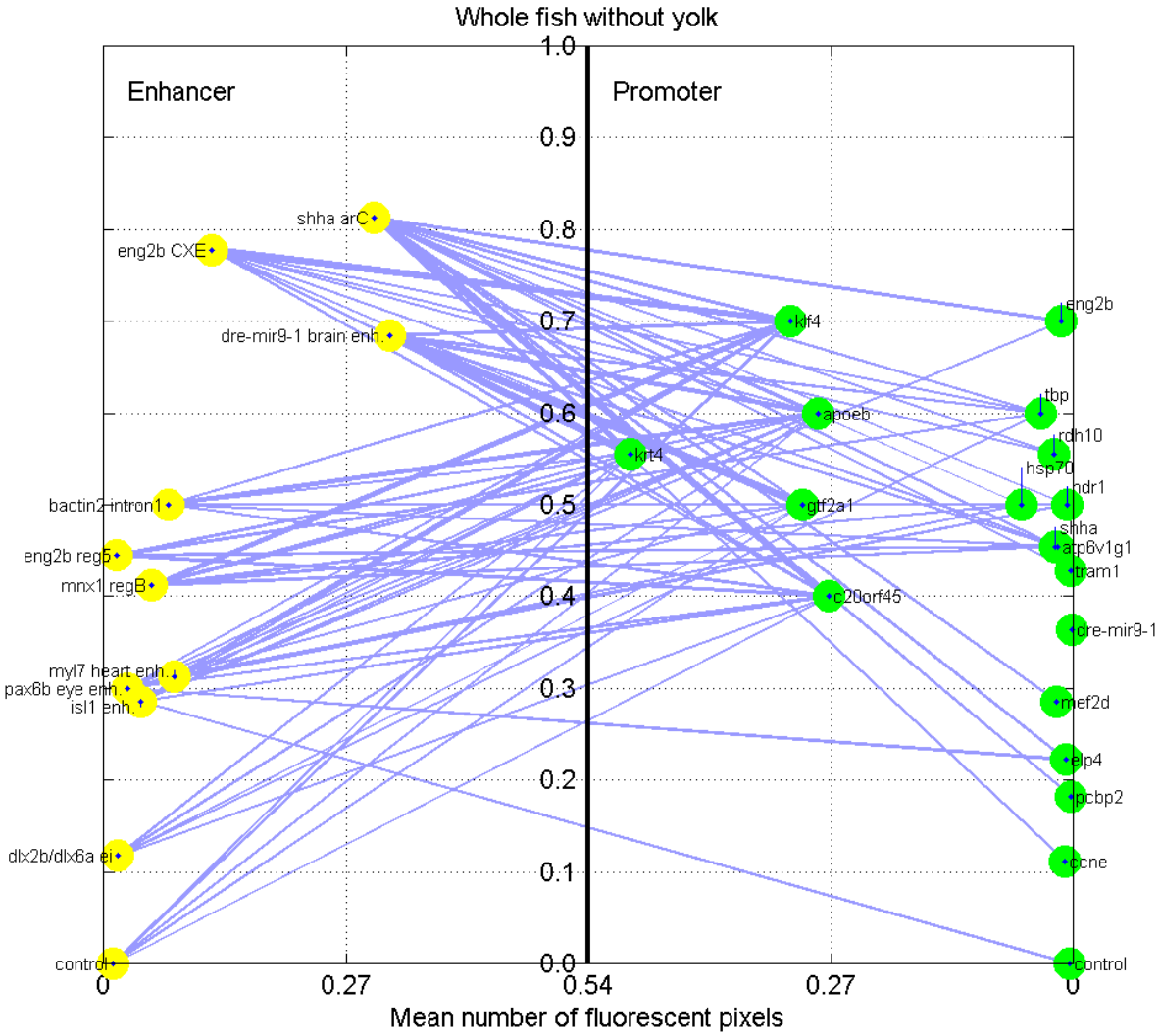


Figure 57: Interaction map

Reporter gene expression activities (total number of YFP-positive pixels per embryo) were calculated for each cis regulatory combination. Blue lines indicate reporter activity of a specific enhancer-promoter combination, where the thickness of the line is proportionate to the strength of activity as measured by pixels counts of YFP signal. Position of promoters (right half, yellow discs) and enhancers (left half, green discs) on the y-axis is proportionate to their ability to generate interaction with each other (interactivity) and is measured as percentage of positive reporter activity obtained of total interactions tested.

The interactions between enhancers and promoters observed by analyzing the projection map pictures were visualized on an interaction map. The interaction map was created by plotting enhancers and promoters on the x-axis representing strength of reporter gene activity (expressed as a mean of all interactions studied with the given enhancer or promoter). All the promoters series were compared to the ctr

enhancer-promoter constructs. Position on the x-axis demonstrates the strength of enhancer-promoter interaction as measured by the average number of YFP-positive pixels in all interaction experiments for the given enhancer or promoter. On the y-axis, the position of an enhancer/promoter was determined by the percentage of combinations showing more expressivity than the controls (Figure 57.).

Based on the results of the quantification and the interaction map, strong core promoters were identified, that showed comparably high interaction capability with several enhancers (e.g. *apoeb*, *klf4* and *krt4*) while other promoters were weakly active and only in conjunction with a small number of enhancers (*pcbp2, tram1, elp4* and *dre-mir9-1).* Furthermore, differential interaction-specificity could be observed with several enhancer-promoter combinations. For example, the *ndr1* and *engrailed2b* promoters had differential ability to interact with the *isl1 zCREST2* and *shha arC* enhancers. While the *eng2b* promoter was efficiently activated by the *zCREST2* in the motor neurons, this enhancer was almost inactive in combination with the *ndr1* promoter. In contrast, the *shha arC* enhancer was better activating the Venus expression in the notochord and ventral brain in conjunction with the *ndr1* promoter than with the *eng2b*. Thus, both the overall strength of activity, as well as the tissue specificity of enhancer-promoter interactions was dependent on the identity of the core promoters applied.

4.3.9 The ability of promoters to respond to enhancers depends on the promoter strength and usage

I have asked whether any characteristics of core promoters used in this study show correlation with the responsiveness of promoters to enhancers. The following properties were taken into account: tissue-specificity of the gene the promoters belong to, TSS distribution and core promoter composition of the promoters and the total number of ESTs mapping to the core promoter region (Table 15.).

	Gene symbol	Number of interactions	Tissue-specificity	Core promoter composition	TSS distribution	Number of ESTs
01	*apoeb*	9	CNS	TATA-box, Inr	dominant	624
05	*krt4*	9	tissue-spec	-	dominant	601
02	*atp6v1g1*	7	CNS	-	dominant, bimodal	195
03	*gtf2a1*	7	general	TATA-box	broad	73
04	*klf4*	7	tissue-spec	-	broad	130
11	*c20orf45*	7	NA	-	broad	88
12	*ccne*	7	CNS	-	broad	20
21	*eng2b*	7	CNS	Inr	broad	8
06	*ndr1*	6	general	Inr	not conclusive	3
19	*hsp70*	6	CNS	TATA-box	not conclusive	4
09	*tbp*	5	general	BRE[d]	dominant	106
13	*shha*	5	CNS	-	not conclusive	3
08	*rdh10*	4	tissue-spec	BRE[d]	dominant	210
17	*elp4*	4	general	BRE[d]	broad	24
10	*tram1*	3	CNS	-	broad	102
15	*mef2d*	3	tissue-spec	-	broad	20
07	*pcbp2*	1	general	BRE[d]	broad	107
16	*dre-mir9-1*	1	CNS	-	not conclusive	2
20	*lmbr1l*	1	general	BRE[d]	not conclusive	0
00	ctr	0	-	-	-	

Table 15: The properties of the promoters used in the screen. The promoters are ranked by the number of interacting enhancers.

To investigate whether the tissue-specificity of the regulatory elements influenced the formation of interactions, based on the embryonic expression patterns of the zebrafish genes, the promoters were grouped into three categories: 1) expression domain in the central nervous system (CNS), 2) expression domain in tissues other than the CNS (called tissue-specific) and 3) general expression pattern. The promoters showed no clustering in terms of tissue-specific expression patterns when the expression profiles of the interacting enhancers were checked (Figure 58.).

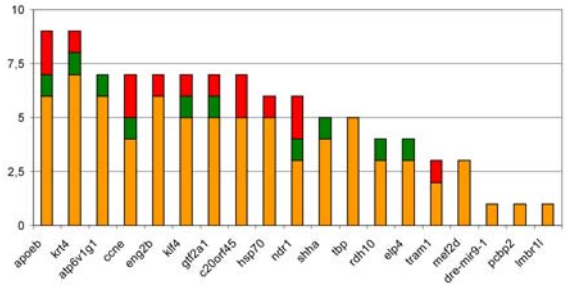


Figure 58: The number of interacting enhancers for each promoter. Orange bars represent the CNS-specific enhancers, green ones the general enhancer and the red bar are for the non-CNS tissue-specific enhancers.

Core promoters of genes from all categories were able to interact with tissue-specific or general enhancers, although general promoters showed a tendency of interacting with less CNS-specific enhancers, compared to CNS-specific or general promoters (Figure 59).

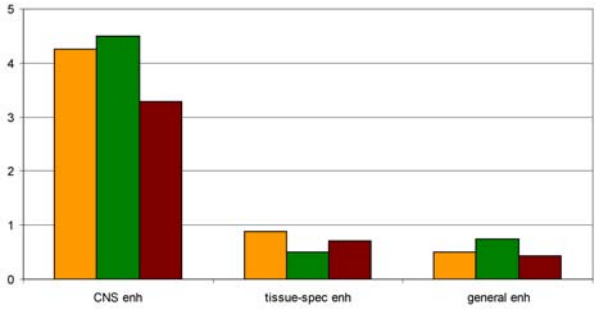


Figure 59: The average of interacting enhancers for the three promoter categories. The orange bars represent the core promoters of genes expressed in the CNS, the green bars are for the promoters expressed in tissues other than the CNS and the purple ones are for the promoters of generally expressed genes.

Next the core promoter composition of the promoters was analyzed. None of the promoters contained DPE, DCE, BRE^{u} or MTE elements. The sequence analysis did not detect any known core promoter elements in the *atp6v1g1*, *tram1*, *c20orf45*, *mef2d* and *dre-mir9-1* promoters. The *apoeb* (at –30 position), *gtf2a1* (-29), and *hsp70* (-42) harbour TATA-box (Bucher 1990), the *pcbp2* (at –28 position), *rdh10* (-20, 19), *tbp* (-15, 1), *elp4* (-7) and *lmbr1l* (-21) promoters contain BRE^{d}, while the *apoeb* (at 13), *ndr1* (-84, 8, 50) and *eng2b* (10) promoters have initiator sequence.

No correlation could be observed between the core promoter composition and the strength of promoters, although BRE^{d} shows a tendency to be present in promoters that are rather weak or having few interacting partners.

To get the information about the TSS distribution of the amplified core promoters ESTs mapping to the promoter regions were retrieved from the dbTSS database or were manually checked in the ENSEMBL database. The TSS distribution did not correlate with the strength, nor with the interactivity of the promoters. Although the strength (expressivity) of the promoters was found to be dependent on the number of EST evidences available for a given promoter. A significant correlation was observed between the ability of promoters to interact with enhancers and the number of ESTs. These results suggest that more active (strong) promoters are more likely to be able to interact with enhancers (Figure 60.).

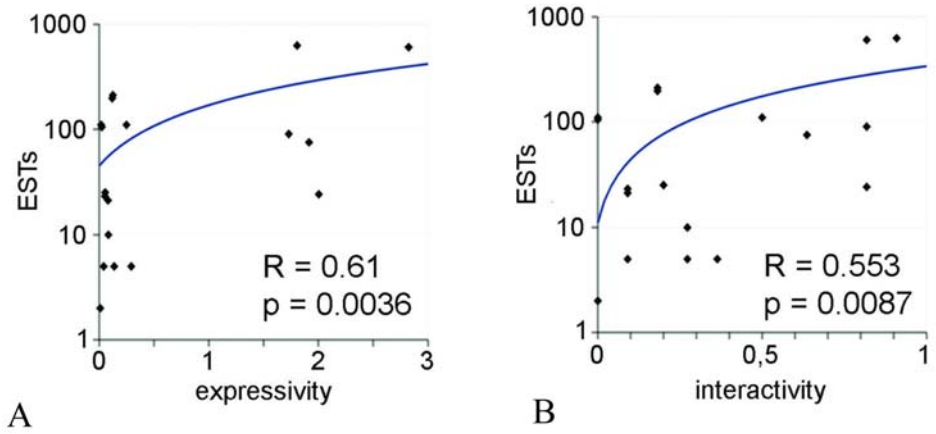


Figure 60: Correlation between the promoter activity (A) or responsiveness to enhancers (B) and the number of ESTs mapped to the promoter region.

4.3.10 Discussion

Large scale cloning of 250 enhancer-promoter combinations

To investigate what determines the specificity of the interaction between cis-regulatory elements I have generated 250 constructs containing different core promoter-enhancer combinations. We decided on using covalently joint fragments for this screen to better control the ratio of the cis-regulatory elements injected. I used the Multisite Gateway vectors system, which allows cloning of fragments in large scale, as it relies on site-specific recombination instead of restriction digest and ligation. I cloned the promoters to the site where usually the gene of interest is cloned, in front of the C-terminal *venus* tag, and the enhancers were recombined into the regulatory element site.

High throughput screening of the generated expression vectors

We have injected these constructs into zebrafish eggs, and detected the expression of the *venus* reporter and the co-injected *cfp* mRNA in pim15 stage embryos. For the automated microscopy embryos were plated into 92-well plates with agarose filling, each containing a pit in the middle for the yolk. The embryos were anesthetised and then the yolks were manually oriented into the wholes. By using a robotic arm to handle the plates in the cooled microscope room, we could even screen 30 plates per night. Pictures were taken with a "Scan^R" high content screening microscope with a 2x objective in bright field and with CFP, YFP filter cubes. After the central focal plane of the embryos was detected by an object detection auto-focus algorithm. Each embryo was acquired with four z-slices (55µm).

The generated pictures were processed by using several algorithms: the embryos were automatically detected and oriented, the bad quality pictures were removed and the pictures taken at 4 z-slices from each and every embryo were projected into one extended focus picture. I used these files to evaluate the interaction of the different cis-regulatory element combination, approximately 15000 extended focus pictures were analysed. The extended focus pictures were then merged into one overlay picture for demonstration purposes, as the overlay of single embryos with mosaic expression could highlight the whole expression patterns.

Automated data analysis

The mean fluorescent pixels were counted in distinct expression domains and for the whole embryo from extended focus pictures for every construct, then were normalised by the area of the domain. These numbers were used to generate a colour coded plot and the interaction map.

The automated data quantification resulted in slightly different results as the manual analysis of the extended focus pictures. For example, the *atp6v1g1*, *ccne* and *ndr1* promoters are ranked as weak promoters with few interacting partners by the automated quantification, while these promoters were activated by seven (*atp6v1g1* and *ccne*) or six (*ndr1*) enhancers. The differential output could arise from the major drawback of the automated analysis. The algorithm cannot distinguish between signals coming from the surface or from the middle of the embryo, the total number of fluorescent pixels were counted per expression domains. This means that the potential background activity in the epithel strongly modified the quantification results. The strongest and mostly interactive promoters, the *apoeb, klf4, krt4, tbp* and *c20orf45* promoters all showed epithel expression with the control and with several other enhancers as well, so the activity of the enhancers in combination with these promoters is probably overestimated. The future challenge is to approve the quantification to overcome these imperfections.

Seven enhancers directed the expression into specific domains

From the twelve enhancers eleven showed enhancer activity when combined to different promoters, while the *pax6b* eye enhancer did not work as an enhancer at all. The *pax6b* eye enhancer was identified as a non-coding sequence highly conserved between fugu and human, located at the *elp4* locus (Woolfe et al. 2005), where a *pax6* eye-enhancer element was previously predicted (Kleinjan et al. 2001). This element was tested in a co-injection experiment with the following parameters: conserved element DNA at 150–300 ng/µl, reporter construct (*β-actin* promoter) DNA at 25 ng/µl concentration (Woolfe et al. 2005). In this screen we covalently joined the enhancer fragment to the promoters, so the molar ratio was obviously much less, compared to the original publication. The differential molar ratio or the use of another promoter could be the reason why this element did not show enhancer activity in combination of any of the tested promoters.

Regarding the specificity, seven enhancers out of the working eleven directed the expression into the previously described tissue/domains, showing tissue-specific enhancer activity with at least one promoter. The *dre-mir9-1* and *dlx2b/dlx6a ei* enhancers worked as general enhancers rather than CNS-specific, and the *myf5* and the *kdrl1* enhancers showed weak and ectopic enhancer activity in combination with few promoters. The enhancer element driving the *myf5* expression into the somites was identified in deletion series (Chen et al. 2007). As it is located close to the core promoter element, it is possible that this element is not an enhancer, but a proximal promoter, working only in the context of the endogenous *myf5* promoter. No vascular epithel-specific Venus expression was observed with the *kdrl* enhancer. This could be due to the fact that this enhancer element was investigated in its original context, in combination with the 1.5-kb *kdrl* promoter (Choi et al. 2007) that could harbour other regulatory elements assisting in the interaction of the enhancer with the promoter.

There were three enhancers for which I have cloned their endogenous promoters as well. Interestingly, the *shh arC-shha* combination did not reveal the strongest signal from the *arC*-series, but the reporter activation was strongly enhancer-specific, no significant background expression was detected (Figure 30.). For the other two enhancers, neither the *eng2b CXE-eng2b,* nor the *eng2b reg5-eng2b* combinations resulted in activation of the Venus expression in the midbrain hindbrain boundary (Figure 34. and 36.). The activity of the two *eng2b* enhancers was tested in co-injection experiments with their endogenous promoter prior to the screen (Figure 21.), and in this experiment they directed the reporter expression into MHB. The promoter fragment was then 1kb long, while in this screen a 227-bp-long core promoter was used. The longer fragment could harbour sequences the enhancers were able to interact with, possibly missing from the basal promoter.

The strenght and the interactivity of promoters varied in a wide range

Three promoters (*pcbp2, dre-mir9-1* and *lmbr1l*) from the nineteen showed expression only in combination with one enhancer. The *pcbp2* promoter could interact only with the *shha arC* enhancer, driving the *venus* expression into *shha*-specific tissues such as hypothalamus, floorplate and notochord. The brain-specific *dre-mir9-1* promoter was only active with the *eng2b CXE* enhancer, and this interaction was also enhancer-specific, while the *lmbr1l* promoter was activated by the *dlx2b/dlx6a ei* in the muscle, yolk, skin and spinal cord, where the *dlx2b* and *dlx6a* genes are not

expressed. On the other hand, several promoters showed high background activity, and these were able to interact with numerous enhancers (*apoeb, krt4, atp6v1g1, ccne, eng2b, klf4, gtf2a1* and *c20orf45*).

Interestingly, the expression patterns gained with the core promoters were not in all case overlapping with the endogenous expression of their genes. For example the *atp6v1g1* gene is expressed at 30hpf in the central nervous system, while its core promoter showed activity in the yolk when tested with the control enhancer. The *gtf2a1* and the *tbp* promoter showed reverse effect, they were expressed in the brain, eye and spinal cord when combined with the control enhancer, while the genes themselves show general expression.

The *dre-mir9-1* and *mef2d* promoters and the *dre-mir9-1* enhancer were chosen to represent gene interdigitation in the dataset. None of the two promoters were activated when combined with the *dre-mir9-1* enhancer, thus no further conclusions can be drawn in terms of interaction specificity of enhancers with their target and bystander promoters.

Enhancer trap experiments performed with different basal promoters showed preference for enhancers driving expression into different tissues. The screen using the basal promoter of the *gata2* gene could identify enhancers assigned to regulators of early development (Ellingsen et al. 2005). The promoter of the *krt4* (previously called as *krt8*) gene was previously used in transposon-mediated enhancer trap (Parinov et al. 2004). The 460-bp promoter construct effectively detected expression patterns in tissues derived from all three germ layers: ectoderm, endoderm, and mesoderm, but the CNS was the main target for the reporter expression (Parinov et al. 2004). The *krt4* core promoter was one of the strongest and most active promoters in our screen: it could interact with nine enhancers, including general, tissue-specific and CNS-specific enhancers. Only the *myf5* and *kdrl* enhancer did not activate expression in combination with the *krt4* promoter, but these enhancers rather worked as weak and unspecific regulatory elements in the screen.

The *hsp70* promoter is widely used as a basal promoter in reporter constructs in enhancer tests and in generation of transgenes (Miyashita et al. 2004; Aizawa et al. 2005; Thummel et al. 2005; Sanges et al. 2006). In this screen, this promoter was able to interact with seven (*shh arC, eng2b CXE, dre-mir9-1, myl7, isl1 zCREST2, dlx2b/dlx6a ei* and *mnx1 regB)* from the eleven working enhancers. In the case of *shh arC, eng2b CXE, myl7, isl1 zCREST2* and *mnx1 regB* enhancers the expression was

highly specific (Figure 41). Most of these enhancers are CNS-specific, but the *myl7* element, driving the expression to the developing heart tube is also in this list. Interestingly, the *ß-actin intron1* enhancer, which worked as a general enhancer in combination with several promoters, did not activate the *hsp70* promoter.

Enhancer action is promoter-specific

I could show that there are differences between different promoters in terms of ability to interact with a given enhancer. These results are contradictory to a previous report, in which a total number of 27 combinations of different enhancers and promoters were tested. This experiment was performed in cell culture, and they have chosen promoters (β-globin and Ig kappa) and enhancers of the immunoglobulin heavy and light chain, SV40 and Moloney sarcoma virus. The combinations of these cis-regulatory elements could evenly activate the reporter expression in the transfected cells (Kermekchiev et al. 1991). In contrary, our screen was performed in a developing vertebrate embryo, which system is much more complex, then tissue-culture. Second, the majority of promoters and enhancers are cis-regulatory elements of developmentally regulated genes, completely differing of the elements of genes expressed in terminally differentiated cells from an adult organism. Thus the two experimental systems are not directly comparable.

Butler et al performed an elegant experiment in Drosophilal by using FLP/Cre excision and enhancer-trapping techniques. They could demonstrate the existence of DPE-, and TATA-box-specific enhancers (Butler et al. 2001). Our results do not show this kind of straightforward correlation between the core-promoter composition and the ability of promoters to interact a specific enhancer. The different classes of promoters do not tend to cluster in terms of interacting with different enhancers. This could be due to the fact, that from the 20 promoters 10 do not contain any of the know core promoter motives (for TATA.box, only the TATAWADR sequence was accepted, based on Bucher's definition) at the right position, so the sample number is too low to detect such a correlation. Despite of few tendencies observed (BRE^d is present is promoters that are rather weak or having few interacting partners, and general promoters showed a tendency of interacting with less CNS-specific enhancers, compared to CNS-specific or general promoters), the only significant correlation found was that the more EST evidences a promoter region has, the stronger and more interactive the promoter is.

5) Conclusions

5.1 Four conserved non-coding elements form the pax2 locus show eye enhancer activity

Sequence comparison of genomic sequences upstream of the orthologuous *pax2* genes from fugu, mouse and human revealed in non-coding sequences (CSTs) with conservation. Previously several enhancers have been identified, which direct the transcription of *pax2* into distinct domains. The element regulating the optic stalk expression was located into a 9kb DNA sequence (Schwarz et al. 2000), but this element was not further characterized. I tested the conserved noncoding sequences in a co-injection assay, with a 5.3kb *pax2a* promoter. To test whether the same enhancer activity is gained with a different promoter, I performed experiments with the mouse minimal *hsp68* promoter as well. The CSTs showed overlapping but not identical results with the two different constructs. I could demonstrate that four of these elements were able to direct *pax2a* expression into the optic stalk and retina in the developing zebrafish embryos at 24hpf stage with both of the tested promoters. These four elements were cloned in front of *hsp68* promoter, and the expression of these constructs was compared to the results gained with the co-injections. The consequent enhancer activity of these conserved non-coding sequences in the retina and/or optic stalk confirm that co-injection of isolated linear DNA sequences can be used for enhancer-assays.

5.2 Combined alignment approach reveals in increased number and variety of conserved non-coding sequences with enhancer function

Genomic sequence aligmnent tools eighter generate false hits (local tools), or miss consreved sequences due to their insensitivity to elements that are shifted in position (global tools). The development of a new sequence comparison method, by combining a global and a local alignment, 21.427 non-genic conserved elements were identified, ten times more, than found by similar approaches. Two thirds of the elements were shuffled during evolution, suggesting that enhancer shuffling is widespread in vertebrates. This type of analysis not only revealed in higher number of conserved non-coding elements, but elements assigned to different types of genes as

well, emphasising the importance of the alignment tool choice in sequence comparison.

22 out of 28 shuffled conserved elements showed significant enhancer activity, from which 20 were tissue-specific. I presented here some examples of functional redundancy of genes, where multiple elements assigned to a single gene showed similar enhancer activity. Gene assignment of the SCEs was confirmed by comparing the expression domains of the enhancers with the endogenous expression patterns of the genes.

5.3 Promoter-specific differences in enhancer action

To investigate whether core promoters and enhancers isolated from their original genomic context show interaction specificty, we have performed a high throughput screen with approximately 23.000 zebrafish embryos. I have generated 250 expression vectors with the Multisite Gateway system, and we have injected these to zebrafish embyos. Pictures were taken with an automated microscope from the 30hpf stage embryos in a special 92-well format, and the generated more than 275.000 pictures were processed by computer algorithms. The fluorescent pixels were counted in distinct domains and in the whole embryos. The data generated by the quantification was used to draw an interaction map.

From the 13 enhancers nine showed enhancer activity consistent with the published activities, while two acted as general enhancers instead of being tissue-specific. Each working enhancer could interact with only a subset of promoters, and the expression was not in all case drirected into the enhancer-specific domains. The enhancer action thus was shown to be promoter-specific, but the properties of core promoters to determine the interaction specificity could not be identified. The only significant correlation observed was that the promoters having more EST evidences in their promoter regions, meaning transcribed more often, were more active in our screen and had inetraction with more enhancer elements.

6) Publications related to the thesis

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

I performed the in vivo enhancer-tests, the microinjections and the following analysis of the transient transgenic fishes.

b) A. Roure, U. Rothbächer, F. Robin, C. Lamy, E. Kalmar, G. Ferone, C. Missero, F. Mueller, and P. Lemaire
"A multicassette Gateway vector set for high throughput and comparative analyses in ciona and vertebrate embryos."
2007. PLoS ONE **2**(9): e916.

I generated the pSP1.72-B3-zf-shh2.5-B5::B1-kozak-Venus-Stop-B2 and the pSP1.72-B3-zf-shh2.5-B5::B1-Kozak-NLSLacZ-Stop-B2 expression vectors containing the zebrafish 2,5 kb shh promoter, performed the micro-injections and analyzed the transient transgenic fishes.

c) E. Kalmar, J. Gehrig, M. Reischl, M. Ferg, Y. Hadzhiev, A. Zaucker, C. Song, S. Schindler, U. Liebel and F. Müller
"High throughput mapping of promoter enhancer interaction specificity by automated spatial registration of reporter gene activity in virtual zebrafish embryos"
under preparation

I generated the expression vectors, organized the screen, took part in the microinjections, the fish-care, the embryo sorting, the plating, the microscopic analysis and took part in the final data-analysis.

7) Acknowledgements

First of all, I would like to express my sincere thanks to Ferenc Müller for his supervision, support, ideas, discussions and for everything I learned for him.

I am thankful to Prof. Uwe Strähle for providing me opportunity to work in the ITG, FZK, for his supervision, and useful suggestions.

I would like to thank Prof. Jochen Wittbrodt for being my second supervisor. I thank all the members of the jury for accepting to preside over my thesis defence.

I would like to thank the members of the Müller lab: Jochen Gehrig, Simone Schindler, Marco Ferg, Andreas Zaucker, Yavor Hadzhiev and Chengyi Song their contribution to the high throughput screening, for the cafe breaks and for the nice atmosphere in the lab. I would like to thank Péter Kóbor his help in the identification and isolation of the *eng2b CXE* and *reg5* enhancers. From the ITG stuff, I am grateful to Nadine Gröbner for her assistance in the high throughput screen and to Nadine Borel for providing help with the fish.

My deepest regards go to Urban Liebel (ITG, FZK, Karlsruhe, Germany) for his enthusiasm and help in setting up and troubleshooting the automated microscopic analysis of several thousand zebrafish embryos, and to Markus Reischl (IAI, FZK, Karlsruhe, Germany) for his knowledge and helpful assistance in the data processing and analysis in the high throughput screen. Without them, I would still sit in front of the microscope.

I would like to say thank you to Remo Sanges, Elia Stupka (CBM, AREA Science Park, Basovizza, Trieste, Italy) and Sandro Banfi (TIGEM, Naples, Italy) for the possibility of working in their project, and to Agnes Roure and Patrick Lemaire (IBDM, Marseille, France) for providing me the modified Multisite Gateway system and for teaching me in the always sunny Marseille how to get over the first difficulties of the cloning.

I thank to Agnes, Anne, Sylwia, Sarah and Vivienne for sharing their time and tricks “how to survive Germany as a foreigner”, and to Wilko and Natascha for their friendship.

Finally, but not final I thank my husband, Peter and my daughter, Hanna their patience, support and understanding, and my parents their belief in me.

8) References

Ahituv, N., et al. (2004). "Exploiting human--fish genome comparisons for deciphering gene regulation." Hum Mol Genet **13 Spec No 2**: R261-6.

Ahituv, N., et al. (2007). "Deletion of ultraconserved elements yields viable mice." PLoS Biol **5**(9): e234.

Aizawa, H., et al. (2005). "Laterotopic representation of left-right information onto the dorso-ventral axis of a zebrafish midbrain target nucleus." Curr Biol **15**(3): 238-43.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P. (2002). "The molecular biology of the cell." Garland Publishing, New York.

Alexander, J., et al. (1998). "Screening mosaic F1 females for mutations affecting zebrafish heart induction and patterning." Dev Genet **22**(3): 288-99.

Altschul, S. F., et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Amsterdam, A. (2006). "Insertional mutagenesis in zebrafish: genes for development, genes for disease." Brief Funct Genomic Proteomic **5**(1): 19-23.

Amsterdam, A., et al. (2005). "Transgenes as screening tools to probe and manipulate the zebrafish genome." Dev Dyn **234**(2): 255-68.

Amsterdam, A., et al. (1999). "Retrovirus-mediated insertional mutagenesis in zebrafish." Methods Cell Biol **60**: 87-98.

Amsterdam, A., et al. (1995). "The Aequorea victoria green fluorescent protein can be used as a reporter in live zebrafish embryos." Dev Biol **171**(1): 123-9.

Anney, R. J., et al. (2002). "Characterisation, mutation detection, and association analysis of alternative promoters and 5' UTRs of the human dopamine D3 receptor gene in schizophrenia." Mol Psychiatry **7**(5): 493-502.

Antonarakis, S. E., et al. (1984). "beta-Thalassemia in American Blacks: novel mutations in the "TATA" box and an acceptor splice site." Proc Natl Acad Sci U S A **81**(4): 1154-8.

Aparicio, S., et al. (1995). "Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes." Proc Natl Acad Sci U S A **92**(5): 1684-8.

Aranda, A., et al. (2001). "Nuclear hormone receptors and gene expression." Physiol Rev **81**(3): 1269-304.

Arbeitman, M. N., et al. (2002). "Gene expression during the life cycle of Drosophila melanogaster." Science **297**(5590): 2270-5.

Arber, S., et al. (1999). "Requirement for the homeobox gene Hb9 in the consolidation of motor neuron identity." Neuron **23**(4): 659-74.

Aronow, B. J., et al. (1992). "Functional analysis of the human adenosine deaminase gene thymic regulatory region and its ability to generate position-independent transgene expression." Mol Cell Biol **12**(9): 4170-85.

Atchison, M. L. (1988). "Enhancers: mechanisms of action and cell specificity." Annu Rev Cell Biol **4**: 127-53.

Babin, P. J., et al. (1997). "Both apolipoprotein E and A-I genes are present in a nonmammalian vertebrate and are highly expressed during embryonic development." Proc Natl Acad Sci U S A **94**(16): 8622-7.

Baek, D., et al. (2007). "Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters." Genome Res **17**(2): 145-55.

Bailey, T. L., et al. (1995). "The value of prior knowledge in discovering motifs with MEME." Proc Int Conf Intell Syst Mol Biol **3**: 21-9.

Bajic, V. B., et al. (2002). "Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters." Bioinformatics **18**(1): 198-9.

Bajic, V. B., et al. (2004). "Promoter prediction analysis on the whole human genome." Nat Biotechnol **22**(11): 1467-73.

Baldwin, A. S., Jr. (1996). "The NF-kappa B and I kappa B proteins: new discoveries and insights." Annu Rev Immunol **14**: 649-83.

Banerji, J., et al. (1981). "Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences." Cell **27**(2 Pt 1): 299-308.

Bang, P. I., et al. (2002). "High-throughput behavioral screening method for detecting auditory response defects in zebrafish." J Neurosci Methods **118**(2): 177-87.

Banine, F., et al. (2005). "SWI/SNF chromatin-remodeling factors induce changes in DNA methylation to promote transcriptional activation." Cancer Res **65**(9): 3542-7.

Barbazuk, W. B., et al. (2000). "The syntenic relationship of the zebrafish and human genomes." Genome Res **10**(9): 1351-8.

Bartfai, R., et al. (2004). "TBP2, a vertebrate-specific member of the TBP family, is required in embryonic development of zebrafish." Curr Biol **14**(7): 593-8.

Basehoar, A. D., et al. (2004). "Identification and distinct regulation of yeast TATA box-containing genes." Cell **116**(5): 699-709.

Beato, M. (1987). "Induction of transcription by steroid hormones." Biochim Biophys Acta **910**(2): 95-102.

Behra, M., et al. (2004). "The use of zebrafish mutants to identify secondary target effects of acetylcholine esterase inhibitors." Toxicol Sci **77**(2): 325-33.

Bejerano, G., et al. (2006). "A distal enhancer and an ultraconserved exon are derived from a novel retroposon." Nature **441**(7089): 87-90.

Bejerano, G., et al. (2004). "Ultraconserved elements in the human genome." Science **304**(5675): 1321-5.

Bell, A. C., et al. (1999). "The protein CTCF is required for the enhancer blocking activity of vertebrate insulators." Cell **98**(3): 387-96.

Belloni, E., et al. (1996). "Identification of Sonic hedgehog as a candidate gene responsible for holoprosencephaly." Nat Genet **14**(3): 353-6.

Bergman, C. M., et al. (2001). "Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences." Genome Res **11**(8): 1335-45.

Bernstein, B. E., et al. (2006). "A bivalent chromatin structure marks key developmental genes in embryonic stem cells." Cell **125**(2): 315-26.

Birney, E., et al. (2006). "Ensembl 2006." Nucleic Acids Res **34**(Database issue): D556-61.

Bishop, C. E., et al. (2000). "A transgenic insertion upstream of sox9 is associated with dominant XX sex reversal in the mouse." Nat Genet **26**(4): 490-4.

Blackwood, E. M., et al. (1998). "Going the distance: a current view of enhancer action." Science **281**(5373): 60-3.

Bourbon, H. M., et al. (1988). "Sequence and structure of the nucleolin promoter in rodents: characterization of a strikingly conserved CpG island." Gene **68**(1): 73-84.

Brand, A. H., et al. (1985). "Characterization of a "silencer" in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer." Cell **41**(1): 41-8.

Brand, M., et al. (1996). "Mutations in zebrafish genes affecting the formation of the boundary between midbrain and hindbrain." Development **123**: 179-90.

Brand, M., et al. (1999). "Identification of TATA-binding protein-free TAFII-containing complex subunits suggests a role in nucleosome acetylation and signal transduction." J Biol Chem **274**(26): 18285-9.

Breathnach, R., et al. (1981). "Organization and expression of eucaryotic split genes coding for proteins." Annu Rev Biochem **50**: 349-83.

Brenner, S., et al. (1993). "Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome." Nature **366**(6452): 265-8.

Brent, R., et al. (1985). "A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor." Cell **43**(3 Pt 2): 729-36.

Brooks, A. R., et al. (1994). "Sequences containing the second-intron enhancer are essential for transcription of the human apolipoprotein B gene in the livers of transgenic mice." Mol Cell Biol **14**(4): 2243-56.

Brudno, M., et al. (2003a). "Fast and sensitive multiple alignment of large genomic sequences." BMC Bioinformatics **4**: 66.

Brudno, M., et al. (2003b). "LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA." Genome Res **13**(4): 721-31.

Bucher, P. (1990). "Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences." J Mol Biol **212**(4): 563-78.

Buckland, P. R., et al. (2005). "Strong bias in the location of functional promoter polymorphisms." Hum Mutat **26**(3): 214-23.

Bulun, S. E., et al. (2007). "Aromatase excess in cancers of breast, endometrium and ovary." J Steroid Biochem Mol Biol **106**(1-5): 81-96.

Buratowski, S., et al. (1988). "Function of a yeast TATA element-binding protein in a mammalian transcription system." Nature **334**(6177): 37-42.

Burke, T. W., et al. (1996). "Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters." Genes Dev **10**(6): 711-24.

Burke, T. W., et al. (1997). "The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila." Genes Dev **11**(22): 3020-31.

Burket, C. T., et al. (2008). "Generation and characterization of transgenic zebrafish lines using different ubiquitous promoters." Transgenic Res **17**(2): 265-79.

Burkhart, J. G. (2000). "Fishing for mutations." Nat Biotechnol **18**(1): 21-2.

Burns, C. G., et al. (2005). "High-throughput assay for small molecules that modulate zebrafish embryonic heart rate." Nat Chem Biol **1**(5): 263-4.

Bussmann, J., et al. (2008). "Zebrafish VEGF receptors: a guideline to nomenclature." PLoS Genet **4**(5): e1000064.

Butler, J. E., et al. (2001). "Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs." Genes Dev **15**(19): 2515-9.

Butler, J. E., et al. (2002). "The RNA polymerase II core promoter: a key component in the regulation of gene expression." Genes Dev **16**(20): 2583-92.

Buttgereit, D. (1993). "Redundant enhancer elements guide beta 1 tubulin gene expression in apodemes during Drosophila embryogenesis." J Cell Sci **105 (Pt 3)**: 721-7.

Cai, J., et al. (2003). "Increased risk for developmental delay in Saethre-Chotzen syndrome is associated with TWIST deletions: an improved strategy for TWIST mutation screening." Hum Genet **114**(1): 68-76.

Calin, G. A., et al. (2007). "Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas." Cancer Cell **12**(3): 215-29.

Callaerts, P., et al. (1999). "Isolation and expression of a Pax-6 gene in the regenerating and intact Planarian Dugesia(G)tigrina." Proc Natl Acad Sci U S A **96**(2): 558-63.

Cammas, L., et al. (2007). "Expression of the murine retinol dehydrogenase 10 (Rdh10) gene correlates with many sites of retinoid signalling during embryogenesis and organ differentiation." Dev Dyn **236**(10): 2899-908.

Camp, S., et al. (2008). "Acetylcholinesterase expression in muscle is specifically controlled by a promoter-selective enhancesome in the first intron." J Neurosci **28**(10): 2459-70.

Carninci, P., et al. (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." Nat Genet **38**(6): 626-35.

Carroll, J. S., et al. (2006). "Genome-wide analysis of estrogen receptor binding sites." Nat Genet **38**(11): 1289-97.

Carroll, S. B. (2000). "Endless forms: the evolution of gene regulation and morphological diversity." Cell **101**(6): 577-80.

Cavalier-Smith, T. (1985). "Selfish DNA and the origin of introns." Nature **315**(6017): 283-4.

Chalkley, G. E., et al. (1999). "DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator." Embo J **18**(17): 4835-45.

Chang, B. E., et al. (1997). "Axial (HNF3beta) and retinoic acid receptors are regulators of the zebrafish sonic hedgehog promoter." Embo J **16**(13): 3955-64.

Chen, J. C., et al. (2004). "The core enhancer is essential for proper timing of MyoD activation in limb buds and branchial arches." Dev Biol **265**(2): 502-12.

Chen, Q., et al. (2005). "Multiple Promoter Targeting Sequences exist in Abdominal-B to regulate long-range gene activation." Dev Biol **286**(2): 629-36.

Chen, W., et al. (2002). "High-throughput selection of retrovirus producer cell lines leads to markedly improved efficiency of germ line-transmissible insertions in zebra fish." J Virol **76**(5): 2192-8.

Chen, Y. H., et al. (2007). "Multiple upstream modules regulate zebrafish myf5 expression." BMC Dev Biol **7**: 1.

Cheng, K. C., et al. (2003). "Functional genomic dissection of multimeric protein families in zebrafish." Dev Dyn **228**(3): 555-67.

Chiang, C., et al. (1996). "Cyclopia and defective axial patterning in mice lacking Sonic hedgehog gene function." Nature **383**(6599): 407-13.

Choi, J., et al. (2007). "FoxH1 negatively modulates flk1 gene expression and vascular formation in zebrafish." Dev Biol **304**(2): 735-44.

Clark, R. M., et al. (2000). "A novel candidate gene for mouse and human preaxial polydactyly with altered expression in limbs of Hemimelic extra-toes mutant mice." Genomics **67**(1): 19-27.

Clarke, M., et al. (1977). "Nonmuscle contractile proteins: the role of actin and myosin in cell motility and shape determination." Annu Rev Biochem **46**: 797-822.

Cooke, J., et al. (1997). "Evolutionary origins and maintenance of redundant gene expression during metazoan development." Trends Genet **13**(9): 360-4.

Cosma, M. P. (2002). "Ordered recruitment: gene-specific mechanism of transcription activation." Mol Cell **10**(2): 227-36.

Covassin, L. D., et al. (2006). "Distinct genetic interactions between multiple Vegf receptors are required for development of different blood vessel types in zebrafish." Proc Natl Acad Sci U S A **103**(17): 6554-9.

Crawford, G. E., et al. (2004). "Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites." Proc Natl Acad Sci U S A **101**(4): 992-7.

Cretekos, C. J., et al. (2008). "Regulatory divergence modifies limb length between mammals." Genes Dev **22**(2): 141-51.

Crossley, M., et al. (1990). "Disruption of a C/EBP binding site in the factor IX promoter is associated with haemophilia B." Nature **345**(6274): 444-6.

Crowley, T. E., et al. (1993). "A new factor related to TATA-binding protein has highly restricted expression patterns in Drosophila." Nature **361**(6412): 557-61.

Curran, T., et al. (1988). "Fos and Jun: the AP-1 connection." Cell **55**(3): 395-7.

Dang, Q., et al. (1995). "Structure of the hepatic control region of the human apolipoprotein E/C-I gene locus." J Biol Chem **270**(38): 22577-85.

Dantonel, J. C., et al. (2000). "TBP-like factor is required for embryonic RNA polymerase II transcription in C. elegans." Mol Cell **6**(3): 715-22.

Davidson, A. E., et al. (2003). "Efficient gene delivery and gene expression in zebrafish using the Sleeping Beauty transposon." Dev Biol **263**(2): 191-202.

Davies, A. F., et al. (1999). "An interstitial deletion of 6p24-p25 proximal to the FKHL7 locus and including AP-2alpha that affects anterior eye chamber development." J Med Genet **36**(9): 708-10.

Davuluri, R. V., et al. (2001). "Computational identification of promoters and first exons in the human genome." Nat Genet **29**(4): 412-7.

Davuluri, R. V., et al. (2008). "The functional consequences of alternative promoter use in mammalian genomes." Trends Genet **24**(4): 167-177.

De Baere, E., et al. (2001). "Spectrum of FOXL2 gene mutations in blepharophimosis-ptosis-epicanthus inversus (BPES) families demonstrates a genotype--phenotype correlation." Hum Mol Genet **10**(15): 1591-600.

de Kok, Y. J., et al. (1996). "Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene POU3F4." Hum Mol Genet **5**(9): 1229-35.

de la Calle-Mustienes, E., et al. (2005). "A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts." Genome Res **15**(8): 1061-72.

de Laat, W., et al. (2003). "Spatial organization of gene expression: the active chromatin hub." Chromosome Res **11**(5): 447-59.

Dee, C. T., et al. (2005). "A novel family of mitochondrial proteins is represented by the Drosophila genes slmo, preli-like and real-time." Dev Genes Evol **215**(5): 248-54.

Deng, W., et al. (2005). "A core promoter element downstream of the TATA box that is recognized by TFIIB." Genes Dev **19**(20): 2418-23.

Diaz, P., et al. (1994). "A locus control region in the T cell receptor alpha/delta locus." Immunity **1**(3): 207-17.

Dillon, N., et al. (2000). "Functional gene expression domains: defining the functional unit of eukaryotic gene regulation." Bioessays **22**(7): 657-65.

Distel, M., et al. (2006). "Multicolor in vivo time-lapse imaging at cellular resolution by stereomicroscopy." Dev Dyn **235**(4): 1100-06.

Dooley, K., et al. (2000). "Zebrafish: a model system for the study of human disease." Curr Opin Genet Dev **10**(3): 252-6.

Dressler, G. R., et al. (1990). "Pax2, a new murine paired-box-containing gene and its expression in the developing excretory system." Development **109**(4): 787-95.

Driever, W., et al. (1996). "A genetic screen for mutations affecting embryogenesis in zebrafish." Development **123**: 37-46.

Echelard, Y., et al. (1993). "Sonic hedgehog, a member of a family of putative signaling molecules, is implicated in the regulation of CNS polarity." Cell **75**(7): 1417-30.

Eddy, S. R. (1999). "Noncoding RNA genes." Curr Opin Genet Dev **9**(6): 695-9.

Ekker, M., et al. (1992). "Coordinate embryonic expression of three zebrafish engrailed genes." Development **116**(4): 1001-10.

Elgar, G., et al. (2008). "Tuning in to the signals: noncoding sequence conservation in vertebrate genomes." Trends Genet **24**(7): 344-52.

Ellies, D. L., et al. (1997). "Relationship between the genomic organization and the overlapping embryonic expression patterns of the zebrafish dlx genes." Genomics **45**(3): 580-90.

Ellingsen, S., et al. (2005). "Large-scale enhancer detection in the zebrafish genome." Development **132**(17): 3799-811.

Emelyanov, A., et al. (2006). "Trans-kingdom transposition of the maize dissociation element." Genetics **174**(3): 1095-104.

Engstrom, P. G., et al. (2008). "Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes." Genome Biol **9**(2): R34.

Ericson, J., et al. (1992). "Early stages of motor neuron differentiation revealed by expression of homeobox gene Islet-1." Science **256**(5063): 1555-60.

Ernst, P., et al. (1996). "A potential role for Elf-1 in terminal transferase gene regulation." Mol Cell Biol **16**(11): 6121-31.

Ettwiller, L., et al. (2005). "The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates." Genome Biol **6**(12): R104.

Fang, J., et al. (2000). "Mutations in FOXC2 (MFH-1), a forkhead family transcription factor, are responsible for the hereditary lymphedema-distichiasis syndrome." Am J Hum Genet **67**(6): 1382-8.

Fantes, J., et al. (1995). "Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype." Hum Mol Genet **4**(3): 415-22.

Feldman, B., et al. (1998). "Zebrafish organizer development and germ-layer formation require nodal-related signals." Nature **395**(6698): 181-5.

Feng, J., et al. (2006). "The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator." Genes Dev **20**(11): 1470-84.

Ferg, M., et al. (2007). "The TATA-binding protein regulates maternal mRNA degradation and differential zygotic transcription in zebrafish." Embo J **26**(17): 3945-56.

Fickett, J. W., et al. (1997). "Eukaryotic promoter recognition." Genome Res **7**(9): 861-78.

Filippova, G. N., et al. (1996). "An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes." Mol Cell Biol **16**(6): 2802-13.

Finbow, M. E., et al. (1997). "The vacuolar H+-ATPase: a universal proton pump of eukaryotes." Biochem J **324 (Pt 3)**: 697-712.

Fisher, S., et al. (2006a). "Conservation of RET regulatory function from human to zebrafish without sequence similarity." Science **312**(5771): 276-9.

Fisher, S., et al. (2006b). "Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish." Nat Protoc **1**(3): 1297-305.

FitzGerald, P. C., et al. (2004). "Clustering of DNA sequences in human promoters." Genome Res **14**(8): 1562-74.

Fjose, A., et al. (1992). "Structure and early embryonic expression of the zebrafish engrailed-2 gene." Mech Dev **39**(1-2): 51-62.

Flodby, P., et al. (2007). "Conserved elements within first intron of aquaporin-5 (Aqp5) function as transcriptional enhancers." Biochem Biophys Res Commun **356**(1): 26-31.

Flomen, R. H., et al. (1998). "Construction and analysis of a sequence-ready map in 4q25: Rieger syndrome can be caused by haploinsufficiency of RIEG, but also by chromosome breaks approximately 90 kb upstream of this gene." Genomics **47**(3): 409-13.

Force, A., et al. (1999). "Preservation of duplicate genes by complementary, degenerative mutations." Genetics **151**(4): 1531-45.

Fujisawa-Sehara, A., et al. (1987). "Characterization of xenobiotic responsive elements upstream from the drug-metabolizing cytochrome P-450c gene: a similarity to glucocorticoid regulatory elements." Nucleic Acids Res **15**(10): 4179-91.

Furlong, E. E., et al. (2001). "Automated sorting of live transgenic embryos." Nat Biotechnol **19**(2): 153-6.

Gahtan, E., et al. (2004). "Of lasers, mutants, and see-through brains: functional neuroanatomy in zebrafish." J Neurobiol **59**(1): 147-61.

Gall, J. G. (1956). "On the submicroscopic structure of chromosomes." Brookhaven Symp Biol(8): 17-32.

Gaszner, M., et al. (2006). "Insulators: exploiting transcriptional and epigenetic mechanisms." Nat Rev Genet **7**(9): 703-13.

Gershenzon, N. I., et al. (2005). "Synergy of human Pol II core promoter elements revealed by statistical sequence analysis." Bioinformatics **21**(8): 1295-300.

Gilbert, W., et al. (1986). "On the antiquity of introns." Cell **46**(2): 151-3.

Giraldez, A. J., et al. (2006). "Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs." Science **312**(5770): 75-9.

Glover, D. M., et al. (1995). "Mutations in aurora prevent centrosome separation leading to the formation of monopolar spindles." Cell **81**(1): 95-105.

Gompel, N., et al. (2005). "Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila." Nature **433**(7025): 481-7.

Goode, D. K., et al. (2005). "Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3." Genomics **86**(2): 172-81.

Goping, I. S., et al. (1995). "A gene-type-specific enhancer regulates the carbamyl phosphate synthetase I promoter by cooperating with the proximal GAG activating element." Nucleic Acids Res **23**(10): 1717-21.

Griffin, K. J., et al. (1998). "Molecular identification of spadetail: regulation of zebrafish trunk and tail mesoderm formation by T-box genes." Development **125**(17): 3379-88.

Gross, P., et al. (2006). "Core promoter-selective RNA polymerase II transcription." Biochem Soc Symp(73): 225-36.

Grosschedl, R., et al. (1981). "Point mutation in the TATA box curtails expression of sea urchin H2A histone gene in vivo." Nature **294**(5837): 178-80.

Grosveld, F., et al. (1987). "Position-independent, high-level expression of the human beta-globin gene in transgenic mice." Cell **51**(6): 975-85.

Gu, H., et al. (1994). "Deletion of a DNA polymerase beta gene segment in T cells using cell type-specific gene targeting." Science **265**(5168): 103-6.

Guo, S., et al. (1999). "Mutations in the zebrafish unmask shared regulatory pathways controlling the development of catecholaminergic neurons." Dev Biol **208**(2): 473-87.

Guyot, B., et al. (2004). "Deletion of the major GATA1 enhancer HS 1 does not affect eosinophil GATA1 expression and eosinophil differentiation." Blood **104**(1): 89-91.

Habeck, H., et al. (2002). "Analysis of a zebrafish VEGF receptor mutant reveals specific disruption of angiogenesis." Curr Biol **12**(16): 1405-12.

Hadrys, T., et al. (2004). "Comparative genomic analysis of vertebrate Hox3 and Hox4 genes." J Exp Zoolog B Mol Dev Evol **302**(2): 147-64.

Hadrys, T., et al. (2006). "Conserved co-regulation and promoter sharing of hoxb3a and hoxb4a in zebrafish." Dev Biol **297**(1): 26-43.

Haecker, S. A., et al. (1995). "Repression of the ovalbumin gene involves multiple negative elements including a ubiquitous transcriptional silencer." Mol Endocrinol **9**(9): 1113-26.

Haffter, P., et al. (1996). "The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio." Development **123**: 1-36.

Hagos, E. G., et al. (2007). "Time-dependent patterning of the mesoderm and endoderm by Nodal signals in zebrafish." BMC Dev Biol **7**: 22.

Hahn, M. W., et al. (2002). "The g-value paradox." Evol Dev **4**(2): 73-5.

Hahn, S. (2004). "Structure and mechanism of the RNA polymerase II transcription machinery." Nat Struct Mol Biol **11**(5): 394-403.

Hahn, S., et al. (1989). "Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus DNA sequences." Proc Natl Acad Sci U S A **86**(15): 5718-22.

Halees, A. S., et al. (2003). "PromoSer: A large-scale mammalian promoter and transcription start site identification service." Nucleic Acids Res **31**(13): 3554-9.

Halfon, M. S., et al. (2002). "Exploring genetic regulatory networks in metazoan development: methods and models." Physiol Genomics **10**(3): 131-43.

Hardy, S., et al. (2002). "TATA-binding protein-free TAF-containing complex (TFTC) and p300 are both required for efficient transcriptional activation." J Biol Chem **277**(36): 32875-82.

Hartley, J. L., et al. (2000). "DNA cloning using in vitro site-specific recombination." Genome Res **10**(11): 1788-95.

Hedlund, E., et al. (2004). "Identification of a Hoxd10-regulated transcriptional network and combinatorial interactions with Hoxa10 during spinal cord development." J Neurosci Res **75**(3): 307-19.

Helms, A. W., et al. (2000). "Autoregulation and multiple enhancers control Math1 expression in the developing nervous system." Development **127**(6): 1185-96.

Henikoff, S., et al. (2004). "TILLING. Traditional mutagenesis meets functional genomics." Plant Physiol **135**(2): 630-6.

Henion, P. D., et al. (1996). "Screen for mutations affecting development of Zebrafish neural crest." Dev Genet **18**(1): 11-7.

Heuchel, R., et al. (1989). "Two closely spaced promoters are equally activated by a remote enhancer: evidence against a scanning model for enhancer action." Nucleic Acids Res **17**(22): 8931-47.

Hinits, Y., et al. (2007). "Mef2s are required for thick filament formation in nascent muscle fibres." Development **134**(13): 2511-9.

Ho, S. Y., et al. (2003). "Analysis of small molecule metabolism in zebrafish." Methods Enzymol **364**: 408-26.

Horikoshi, M., et al. (1989). "Purification of a yeast TATA box-binding protein that exhibits human transcription factor IID activity." Proc Natl Acad Sci U S A **86**(13): 4843-7.

Howell, M., et al. (1997). "XSmad2 directly activates the activin-inducible, dorsal mesoderm gene XFKH1 in Xenopus embryos." Embo J **16**(24): 7411-21.

Huang, C. J., et al. (2003). "Germ-line transmission of a myocardium-specific GFP transgene reveals critical regulatory elements in the cardiac myosin light chain 2 promoter of zebrafish." Dev Dyn **228**(1): 30-40.

Hural, J. A., et al. (2000). "An intron transcriptional enhancer element regulates IL-4 gene locus accessibility in mast cells." J Immunol **165**(6): 3239-49.

Imboden, M., et al. (1997). "Cytokeratin 8 is a suitable epidermal marker during zebrafish development." C R Acad Sci III **320**(9): 689-700.

Ingolia, T. D., et al. (1982). "Drosophila gene related to the major heat shock-induced gene is transcribed at normal temperatures and not induced by heat shock." Proc Natl Acad Sci U S A **79**(2): 525-9.

Inoue, A., et al. (1994). "Developmental regulation of islet-1 mRNA expression during neuronal differentiation in embryonic zebrafish." Dev Dyn **199**(1): 1-11.

Ivics, Z., et al. (1997). "Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells." Cell **91**(4): 501-10.

Iwama, H., et al. (2004). "Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network." Proc Natl Acad Sci U S A **101**(49): 17156-61.

Jackman, W. R., et al. (2000). "islet reveals segmentation in the Amphioxus hindbrain homolog." Dev Biol **220**(1): 16-26.

Jamieson, R. V., et al. (2002). "Domain disruption and mutation of the bZIP transcription factor, MAF, associated with cataract, ocular anterior segment dysgenesis and coloboma." Hum Mol Genet **11**(1): 33-42.

Jeong, Y., et al. (2006). "A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers." Development **133**(4): 761-72.

Jessen, J. R., et al. (1999). "Artificial chromosome transgenesis reveals long-distance negative regulation of rag1 in zebrafish." Nat Genet **23**(1): 15-6.

Jones, B. K., et al. (1995). "The human growth hormone gene is regulated by a multicomponent locus control region." Mol Cell Biol **15**(12): 7010-21.

Jongens, T. A., et al. (1988). "Functional redundancy in the tissue-specific enhancer of the Drosophila Sgs-4 gene." Embo J **7**(8): 2559-67.

Joyner, A. L., et al. (1985). "Expression during embryogenesis of a mouse gene with sequence homology to the Drosophila engrailed gene." Cell **43**(1): 29-37.

Juneau, K., et al. (2006). "Introns regulate RNA and protein abundance in yeast." Genetics **174**(1): 511-8.

Kamat, A., et al. (1999). "A 500-bp region, approximately 40 kb upstream of the human CYP19 (aromatase) gene, mediates placenta-specific expression in transgenic mice." Proc Natl Acad Sci U S A **96**(8): 4575-80.

Kaneko, M., et al. (2005). "Light-dependent development of circadian gene expression in transgenic zebrafish." PLoS Biol **3**(2): e34.

Karlstrom, R. O., et al. (1999). "Comparative synteny cloning of zebrafish you-too: mutations in the Hedgehog target gli2 affect ventral forebrain patterning." Genes Dev **13**(4): 388-93.

Katayama, S., et al. (2005). "Antisense transcription in the mammalian transcriptome." Science **309**(5740): 1564-6.

Kaufmann, J., et al. (1994). "Direct recognition of initiator elements by a component of the transcription factor IID complex." Genes Dev **8**(7): 821-9.

Kawahara, A., et al. (2000). "Expression of the Kruppel-like zinc finger gene biklf during zebrafish development." Mech Dev **97**(1-2): 173-6.

Kawaji, H., et al. (2006). "Dynamic usage of transcription start sites within core promoters." Genome Biol **7**(12): R118.

Kawakami, K. (2004). "Transgenesis and gene trap methods in zebrafish by using the Tol2 transposable element." Methods Cell Biol **77**: 201-22.

Kawakami, K., et al. (1998). "Excision of the tol2 transposable element of the medaka fish, Oryzias latipes, in zebrafish, Danio rerio." Gene **225**(1-2): 17-22.

Keplinger, B. L., et al. (2001). "Complex organization of promoter and enhancer elements regulate the tissue- and developmental stage-specific expression of the Drosophila melanogaster Gld gene." Genetics **157**(2): 699-716.

Kermekchiev, M., et al. (1991). "Every enhancer works with every promoter for all the combinations tested: could new regulatory pathways evolve by enhancer shuffling?" Gene Expr **1**(1): 71-81.

Khandekar, M., et al. (2007). "A Gata2 intronic enhancer confers its pan-endothelia-specific regulation." Development **134**(9): 1703-12.

Kikuta, H., et al. (2007). "Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates." Genome Res **17**(5): 545-55.

Kily, L. J., et al. (2008). "Gene expression changes in a zebrafish model of drug dependency suggest conservation of neuro-adaptation pathways." J Exp Biol **211**(Pt 10): 1623-34.

Kim, M. K., et al. (1996). "A soluble transcription factor, Oct-1, is also found in the insoluble nuclear matrix and possesses silencing activity in its alanine-rich domain." Mol Cell Biol **16**(8): 4366-77.

Kim, T. H., et al. (2005). "A high-resolution map of active promoters in the human genome." Nature **436**(7052): 876-80.

Kimura-Yoshida, C., et al. (2004). "Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification." Development **131**(1): 57-71.

Kimura, K., et al. (2006). "Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes." Genome Res **16**(1): 55-65.

King, D. C., et al. (2007). "Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data." Genome Res **17**(6): 775-86.

King, M. C., et al. (1975). "Evolution at two levels in humans and chimpanzees." Science **188**(4184): 107-16.

Kingsley, C., et al. (1992). "Cloning of GT box-binding proteins: a novel Sp1 multigene family regulating T-cell receptor gene expression." Mol Cell Biol **12**(10): 4251-61.

Kininis, M., et al. (2007). "Genomic analyses of transcription factor binding, histone acetylation, and gene expression reveal mechanistically distinct classes of estrogen-regulated promoters." Mol Cell Biol **27**(14): 5090-104.

Kininis, M., et al. (2008). "A global view of transcriptional regulation by nuclear receptors: gene expression, factor localization, and DNA sequence analysis." Nucl Recept Signal **6**: e005.

Kleinjan, D. A., et al. (2001). "Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6." Hum Mol Genet **10**(19): 2049-59.

Kleinjan, D. A., et al. (2005). "Long-range control of gene expression: emerging mechanisms and disruption in disease." Am J Hum Genet **76**(1): 8-32.

Klenova, E. M., et al. (1993). "CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms." Mol Cell Biol **13**(12): 7612-24.

Kmita, M., et al. (2002). "Serial deletions and duplications suggest a mechanism for the collinearity of Hoxd genes in limbs." Nature **420**(6912): 145-50.

Knudsen, S. (1999). "Promoter2.0: for the recognition of PolII promoter sequences." Bioinformatics **15**(5): 356-61.

Koivisto, U. M., et al. (1994). "A single-base substitution in the proximal Sp1 site of the human low density lipoprotein receptor promoter as a cause of heterozygous familial hypercholesterolemia." Proc Natl Acad Sci U S A **91**(22): 10526-30.

Kokel, D., et al. (2008). "Chemobehavioural phenomics and behaviour-based psychiatric drug discovery in the zebrafish." Brief Funct Genomic Proteomic.

Koster, R. W., et al. (2001). "Tracing transgene expression in living zebrafish embryos." Dev Biol **233**(2): 329-46.

Kothary, R., et al. (1989). "Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice." Development **105**(4): 707-14.

Krauss, S., et al. (1991). "Expression pattern of zebrafish pax genes suggests a role in early brain regionalization." Nature **353**(6341): 267-70.

Krichevsky, A. M., et al. (2003). "A microRNA array reveals extensive regulation of microRNAs during brain development." Rna **9**(10): 1274-81.

Kudoh, T., et al. (2001). "A gene expression screen in zebrafish embryogenesis." Genome Res **11**(12): 1979-87.

Kulozik, A. E., et al. (1991). "Thalassemia intermedia: moderate reduction of beta globin gene transcriptional activity by a novel mutation of the proximal CACCC promoter element." Blood **77**(9): 2054-8.

Kwon, Y. S., et al. (2007). "Sensitive ChIP-DSL technology reveals an extensive estrogen receptor alpha-binding program on human gene promoters." Proc Natl Acad Sci U S A **104**(12): 4852-7.

Lagrange, T., et al. (1998). "New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB." Genes Dev **12**(1): 34-44.

Lalioti, M. D., et al. (1997). "Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy." Nature **386**(6627): 847-51.

Lam, S. H., et al. (2008). "Zebrafish whole-adult-organism chemogenomics for large-scale predictive and discovery chemical biology." PLoS Genet **4**(7): e1000121.

Landry, J. R., et al. (2003). "Complex controls: the role of alternative promoters in mammalian genomes." Trends Genet **19**(11): 640-8.

Laplante, M., et al. (2006). "Enhancer detection in the zebrafish using pseudotyped murine retroviruses." Methods **39**(3): 189-98.

Lareau, L. F., et al. (2007). "Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements." Nature **446**(7138): 926-9.

Lee, D. H., et al. (2005). "Functional characterization of core promoter elements: the downstream core element is recognized by TAF1." Mol Cell Biol **25**(21): 9674-86.

Lee, T. I., et al. (2006). "Control of developmental regulators by Polycomb in human embryonic stem cells." Cell **125**(2): 301-13.

Lee, T. I., et al. (2000). "Transcription of eukaryotic protein-coding genes." Annu Rev Genet **34**: 77-137.

Lehmann, A. R. (2001). "The xeroderma pigmentosum group D (XPD) gene: one gene, two functions, three diseases." Genes Dev **15**(1): 15-23.

Lettice, L. A., et al. (2003). "A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly." Hum Mol Genet **12**(14): 1725-35.

Lettice, L. A., et al. (2002). "Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly." Proc Natl Acad Sci U S A **99**(11): 7548-53.

Levine, M., et al. (2003). "Transcription regulation and animal diversity." Nature **424**(6945): 147-51.

Lewis, B. A., et al. (2000). "A downstream element in the human beta-globin promoter: evidence of extended sequence-specific transcription factor IID contacts." Proc Natl Acad Sci U S A **97**(13): 7172-7.

Li, G., et al. (2007). "Detection of blob objects in microscopic zebrafish images based on gradient vector diffusion." Cytometry A **71**(10): 835-45.

Li, Q., et al. (2002). "Locus control regions." Blood **100**(9): 3077-86.

Li Song, D., et al. (2000). "Two Pax2/5/8-binding sites in Engrailed2 are required for proper initiation of endogenous mid-hindbrain expression." Mech Dev **90**(2): 155-65.

Li, X., et al. (1994). "Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the Drosophila embryo." Embo J **13**(2): 400-6.

Lim, C. Y., et al. (2004). "The MTE, a new core promoter element for transcription by RNA polymerase II." Genes Dev **18**(13): 1606-17.

Lin, Q., et al. (2007). "Promoter targeting sequence mediates enhancer interference in the Drosophila embryo." Proc Natl Acad Sci U S A **104**(9): 3237-42.

Liu, B., et al. (1996). "Repression of platelet-derived growth factor A-chain gene transcription by an upstream silencer element. Participation by sequence-specific single-stranded DNA-binding proteins." J Biol Chem **271**(42): 26281-90.

Liu, J., et al. (1990a). "Mouse U14 snRNA is encoded in an intron of the mouse cognate hsc70 heat shock gene." Nucleic Acids Res **18**(22): 6565-71.

Liu, J. K., et al. (1997). "Dlx genes encode DNA-binding proteins that are expressed in an overlapping and sequential pattern during basal ganglia differentiation." Dev Dyn **210**(4): 498-512.

Liu, Q. R., et al. (2005). "Human brain derived neurotrophic factor (BDNF) genes, splicing patterns, and assessments of associations with substance abuse and Parkinson's Disease." Am J Med Genet B Neuropsychiatr Genet **134B**(1): 93-103.

Liu, T., et al. (2008). "An automated method for cell detection in zebrafish." Neuroinformatics **6**(1): 5-21.

Liu, Z. J., et al. (1990b). "Functional analysis of elements affecting expression of the beta-actin gene of carp." Mol Cell Biol **10**(7): 3432-40.

Lo, J., et al. (2003). "15000 unique zebrafish EST clusters and their future use in microarray for profiling gene expression patterns during embryogenesis." Genome Res **13**(3): 455-66.

Logan, C., et al. (1993). "Two enhancer regions in the mouse En-2 locus direct expression to the mid/hindbrain region and mandibular myoblasts." Development **117**(3): 905-16.

Long, Q., et al. (1997). "GATA-1 expression pattern can be recapitulated in living transgenic zebrafish using GFP reporter gene." Development **124**(20): 4105-11.

Lu, J., et al. (2008). "Prediction for human transcription start site using diversity measure with quadratic discriminant." Bioinformation **2**(7): 316-21.

Ludlow, L. B., et al. (1996). "Identification of a mutation in a GATA binding site of the platelet glycoprotein Ibbeta promoter resulting in the Bernard-Soulier syndrome." J Biol Chem **271**(36): 22076-80.

Ludwig, M. Z., et al. (2000). "Evidence for stabilizing selection in a eukaryotic enhancer element." Nature **403**(6769): 564-7.

Lun, K., et al. (1998). "A series of no isthmus (noi) alleles of the zebrafish pax2.1 gene reveals multiple signaling events in development of the midbrain-hindbrain boundary." Development **125**(16): 3049-62.

Lynch, M., et al. (2000). "The evolutionary fate and consequences of duplicate genes." Science **290**(5494): 1151-5.

Mack, D. H., et al. (1993). "Specific repression of TATA-mediated but not initiator-mediated transcription by wild-type p53." Nature **363**(6426): 281-3.

Mackenzie, A., et al. (2004). "Is there a functional link between gene interdigitation and multi-species conservation of synteny blocks?" Bioessays **26**(11): 1217-24.

Makeyev, A. V., et al. (2002). "The poly(C)-binding proteins: a multiplicity of functions and a search for mechanisms." Rna **8**(3): 265-78.

Manco, L., et al. (2000). "A new PKLR gene mutation in the R-type promoter region affects the gene transcription causing pyruvate kinase deficiency." Br J Haematol **110**(4): 993-7.

Marcelle, C., et al. (1992). "Molecular cloning of a family of protein kinase genes expressed in the avian embryo." Oncogene **7**(12): 2479-87.

Marcu, K. B., et al. (1992). "myc function and regulation." Annu Rev Biochem **61**: 809-60.

Masotti, C., et al. (2005). "A functional SNP in the promoter region of TCOF1 is associated with reduced gene expression and YY1 DNA-protein interaction." Gene **359**: 44-52.

Maston, G. A., et al. (2006). "Transcriptional regulatory elements in the human genome." Annu Rev Genomics Hum Genet **7**: 29-59.

Mathavan, S., et al. (2005). "Transcriptome analysis of zebrafish embryogenesis using microarrays." PLoS Genet **1**(2): 260-76.

Matsuda, M., et al. (1992). "Delta-thalassemia caused by disruption of the site for an erythroid-specific transcription factor, GATA-1, in the delta-globin gene promoter." Blood **80**(5): 1347-51.

Mayor, C., et al. (2000). "VISTA : visualizing global DNA sequence alignments of arbitrary length." Bioinformatics **16**(11): 1046-7.

Mc Clintock, B. (1951). "Chromosome organization and genic expression." Cold Spring Harb Symp Quant Biol **16**: 13-47.

McEwen, G. K., et al. (2006). "Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis." Genome Res **16**(4): 451-65.

Meier, C. A. (1996). "Co-activators and co-repressors: mediators of gene activation by nuclear hormone receptors." Eur J Endocrinol **134**(2): 158-9.

Meijer, A. H., et al. (2005). "Transcriptome profiling of adult zebrafish at the late stage of chronic tuberculosis due to Mycobacterium marinum infection." Mol Immunol **42**(10): 1185-203.

Meng, A., et al. (1997). "Promoter analysis in living zebrafish embryos identifies a cis-acting motif required for neuronal expression of GATA-2." Proc Natl Acad Sci U S A **94**(12): 6267-72.

Meng, A., et al. (1999). "Positive and negative cis-acting elements are required for hematopoietic expression of zebrafish GATA-1." Blood **93**(2): 500-8.

Metz, R., et al. (1994). "c-Fos-induced activation of a TATA-box-containing promoter involves direct contact with TATA-box-binding protein." Mol Cell Biol **14**(9): 6021-9.

Mikkelsen, T. S., et al. (2007). "Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences." Nature **447**(7141): 167-77.

Milan, D. J., et al. (2003). "Drugs that induce repolarization abnormalities cause bradycardia in zebrafish." Circulation **107**(10): 1355-8.

Miles, C. G., et al. (2003). "Faithful expression of a tagged Fugu WT1 protein from a genomic transgene in zebrafish: efficient splicing of pufferfish genes in zebrafish but not mice." Nucleic Acids Res **31**(11): 2795-802.

Miyashita, T., et al. (2004). "PlexinA4 is necessary as a downstream target of Islet2 to mediate Slit signaling for promotion of sensory axon branching." Development **131**(15): 3705-15.

Moens, C. B., et al. (1996). "valentino: a zebrafish gene required for normal hindbrain segmentation." Development **122**(12): 3981-90.

Muller, F., et al. (2002). "Search for enhancers: teleost models in comparative genomic and transgenic analysis of cis regulatory elements." Bioessays **24**(6): 564-72.

Muller, F., et al. (1999). "Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord." Development **126**(10): 2103-16.

Muller, F., et al. (2001). "TBP is not universally required for zygotic RNA polymerase II transcription in zebrafish." Curr Biol **11**(4): 282-7.

Muller, F., et al. (2004). "The multicoloured world of promoter recognition complexes." Embo J **23**(1): 2-8.

Muller, F., et al. (1997). "Activator effect of coinjected enhancers on the muscle-specific expression of promoters in zebrafish embryos." Mol Reprod Dev **47**(4): 404-12.

Nakano, T., et al. (2005). "Identification of a conserved 125 base-pair Hb9 enhancer that specifies gene expression to spinal motor neurons." Dev Biol **283**(2): 474-85.

Nasevicius, A., et al. (2000). "Effective targeted gene 'knockdown' in zebrafish." Nat Genet **26**(2): 216-20.

Neznanov, N., et al. (1993). "Transcriptional insulation of the human keratin 18 gene in transgenic mice." Mol Cell Biol **13**(4): 2214-23.

Ni, J. Z., et al. (2007). "Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay." Genes Dev **21**(6): 708-18.

Nishi, T., et al. (2002). "The vacuolar (H+)-ATPases--nature's most versatile proton pumps." Nat Rev Mol Cell Biol **3**(2): 94-103.

Nishihara, H., et al. (2006). "Functional noncoding sequences derived from SINEs in the mammalian genome." Genome Res **16**(7): 864-74.

Nobrega, M. A., et al. (2003). "Scanning human gene deserts for long-range enhancers." Science **302**(5644): 413.

Nobrega, M. A., et al. (2004). "Megabase deletions of gene deserts result in viable mice." Nature **431**(7011): 988-93.

Nornes, H. O., et al. (1990). "Spatially and temporally restricted expression of Pax2 during murine neurogenesis." Development **109**(4): 797-809.

Nowak, M. A., et al. (1997). "Evolution of genetic redundancy." Nature **388**(6638): 167-71.

Nusslein-Volhard, C., et al. (1980). "Mutations affecting segment number and polarity in Drosophila." Nature **287**(5785): 795-801.

Ogbourne, S., et al. (1998). "Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes." Biochem J **331 (Pt 1)**: 1-14.

Ohno, S., et al. (1968). "Evolution from fish to mammals by gene duplication." Hereditas **59**(1): 169-87.

Ohtsuki, S., et al. (1998). "Different core promoters possess distinct regulatory activities in the Drosophila embryo." Genes Dev **12**(4): 547-56.

Osborne, C. S., et al. (2004). "Active genes dynamically colocalize to shared sites of ongoing transcription." Nat Genet **36**(10): 1065-71.

Ovcharenko, I., et al. (2005). "Evolution and functional classification of vertebrate gene deserts." Genome Res **15**(1): 137-45.

Pabo, C. O., et al. (1992). "Transcription factors: structural families and principles of DNA recognition." Annu Rev Biochem **61**: 1053-95.

Paige, A. J., et al. (2000). "A 700-kb physical map of a region of 16q23.2 homozygously deleted in multiple cancers and spanning the common fragile site FRA16D." Cancer Res **60**(6): 1690-7.

Paparidis, Z., et al. (2007). "Ultraconserved non-coding sequence element controls a subset of spatiotemporal GLI3 expression." Dev Growth Differ **49**(6): 543-53.

Parinov, S., et al. (2004). "Tol2 transposon-mediated enhancer trap to identify developmentally regulated zebrafish genes in vivo." Dev Dyn **231**(2): 449-59.

Park, C. S., et al. (1982). "Molecular mechanism of promoter selection in gene transcription. I. Development of a rapid mixing-photocrosslinking technique to study the kinetics of Escherichia coli RNA polymerase binding to T7 DNA." J Biol Chem **257**(12): 6944-9.

Parng, C., et al. (2002). "Zebrafish: a preclinical model for drug screening." Assay Drug Dev Technol **1**(1 Pt 1): 41-8.

Parsons, M. J., et al. (2004). "The -1438A/G polymorphism in the 5-hydroxytryptamine type 2A receptor gene affects promoter activity." Biol Psychiatry **56**(6): 406-10.

Pastinen, T., et al. (2006). "Influence of human genome polymorphism on gene expression." Hum Mol Genet **15 Spec No 1**: R9-16.

Patikoglou, G. A., et al. (1999). "TATA element recognition by the TATA box-binding protein has been conserved throughout evolution." Genes Dev **13**(24): 3217-30.

Pearson, W. R., et al. (1988). "Improved tools for biological sequence comparison." Proc Natl Acad Sci U S A **85**(8): 2444-8.

Peaston, A. E., et al. (2004). "Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos." Dev Cell **7**(4): 597-606.

Peltoketo, H., et al. (1994). "A point mutation in the putative TATA box, detected in nondiseased individuals and patients with hereditary breast cancer, decreases promoter activity of the 17 beta-hydroxysteroid dehydrogenase type 1 gene 2 (EDH17B2) in vitro." Genomics **23**(1): 250-2.

Pennacchio, L. A., et al. (2006). "In vivo enhancer analysis of human conserved non-coding sequences." Nature **444**(7118): 499-502.

Pfeffer, P., et al. (2002). "The activation and maintenance of Pax2 expression at the mid-hindbrain boundary is controlled by separate enhancers." Development **129**(2): 307-18.

Pfeffer, P. L., et al. (1998). "Characterization of three novel members of the zebrafish Pax2/5/8 family: dependency of Pax5 and Pax8 expression on the Pax2.1 (noi) function " Development **125**(16): 3063-3074.

Pfeifer, D., et al. (1999). "Campomelic dysplasia translocation breakpoints are scattered over 1 Mb proximal to SOX9: evidence for an extended control region." Am J Hum Genet **65**(1): 111-24.

Pichler, F. B., et al. (2003). "Chemical discovery and global gene expression analysis in zebrafish." Nat Biotechnol **21**(8): 879-83.

Picker A, et al. (2002). "A novel positive transcriptional feedback loop in midbrain-hindbrain boundary development is revealed through analysis of the zebrafish pax2.1 promoter in transgenic lines." Development **129**(13): 3227-39.

Plessy, C., et al. (2005). "Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes." Trends Genet **21**(4): 207-10.

Pollet, N., et al. (2001). "Expression profiling by systematic high-throughput in situ hybridization to whole-mount embryos." Methods Mol Biol **175**: 309-21.

Ponjavic, J., et al. (2006). "Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters." Genome Biol **7**(8): R78.

Popescu, N. C., et al. (2002). "Chromosome-mediated alterations of the MYC gene in human cancer." J Cell Mol Med **6**(2): 151-9.

Postlethwait, J., et al. (2004). "Subfunction partitioning, the teleost radiation and the annotation of the human genome." Trends Genet **20**(10): 481-90.

Poulin, F., et al. (2005). "In vivo characterization of a vertebrate ultraconserved enhancer." Genomics **85**(6): 774-81.

Pownall, M. E., et al. (2002). "Myogenic regulatory factors and the specification of muscle progenitors in vertebrate embryos." Annu Rev Cell Dev Biol **18**: 747-83.

Prabhakar, S., et al. (2008). "Human-specific gain of function in a developmental enhancer." Science **321**(5894): 1346-50.

Prince, V. E., et al. (1998). "Zebrafish hox genes: expression in the hindbrain region of wild-type and mutants of the segmentation gene, valentino." Development **125**(3): 393-406.

Prince, V. E., et al. (2002). "Splitting pairs: the diverging fates of duplicated genes." Nat Rev Genet **3**(11): 827-37.

Prud'homme, B., et al. (2006). "Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene." Nature **440**(7087): 1050-3.

Ptashne, M., et al. (1997). "Transcriptional activation by recruitment." Nature **386**(6625): 569-77.

Rauch, G. J., et al. (2003). "Submission and Curation of Gene Expression Data " Direct data submission (http://zfin.org).

Riddle, R. D., et al. (1993). "Sonic hedgehog mediates the polarizing activity of the ZPA." Cell **75**(7): 1401-16.

Rinchik, E. M., et al. (1990). "A strategy for fine-structure functional analysis of a 6- to 11-centimorgan region of mouse chromosome 7 by high-efficiency mutagenesis." Proc Natl Acad Sci U S A **87**(3): 896-900.

Rippe, K. (2001). "Making contacts on a nucleic acid polymer." Trends Biochem Sci **26**(12): 733-40.

Rossi, F. M., et al. (2000). "Transcriptional control: rheostat converted to on/off switch." Mol Cell **6**(3): 723-8.

Roure, A., et al. (2007). "A multicassette Gateway vector set for high throughput and comparative analyses in ciona and vertebrate embryos." PLoS ONE **2**(9): e916.

Roy, A. L., et al. (1991). "Cooperative interaction of an initiator-binding transcription initiation factor and the helix-loop-helix activator USF." Nature **354**(6350): 245-8.

Russell, L. B., et al. (1982). "Analysis of the albino-locus region of the mouse: IV. Characterization of 34 deficiencies." Genetics **100**(3): 427-53.

Sagai, T., et al. (2005). "Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb." Development **132**(4): 797-803.

Sakata-Takatani, K., et al. (2004). "Identification of a functional CBF-binding CCAAT-like motif in the core promoter of the mouse pro-alpha 1(V) collagen gene (Col5a1)." Matrix Biol **23**(2): 87-99.

Sandelin, A., et al. (2004). "Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes." BMC Genomics **5**(1): 99.

Sandelin, A., et al. (2007). "Mammalian RNA polymerase II core promoters: insights from genome-wide studies." Nat Rev Genet **8**(6): 424-36.

Sanges, R., et al. (2006). "Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage." Genome Biol **7**(7): R56.

Santoro, C., et al. (1988). "A family of human CCAAT-box-binding proteins active in transcription and DNA replication: cloning and expression of multiple cDNAs." Nature **334**(6179): 218-24.

Scemama, J. L., et al. (2002). "Evolutionary divergence of vertebrate Hoxb2 expression patterns and transcriptional regulatory loci." J Exp Zool **294**(3): 285-99.

Scheer, N., et al. (1999). "Use of the Gal4-UAS technique for targeted gene expression in the zebrafish." Mech Dev **80**(2): 153-8.

Scheer, N., et al. (2002). "A quantitative analysis of the kinetics of Gal4 activator and effector gene expression in the zebrafish." Mech Dev **112**(1-2): 9-14.

Scherer, S. W., et al. (1994). "Physical mapping of the split hand/split foot locus on chromosome 7 and implication in syndromic ectrodactyly." Hum Mol Genet **3**(8): 1345-54.

Scherf, M., et al. (2000). "Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach." J Mol Biol **297**(3): 599-606.

Schmid, C. D., et al. (2006). "EPD in its twentieth year: towards complete promoter coverage of selected model organisms." Nucleic Acids Res **34**(Database issue): D82-5.

Schulte-Merker, S., et al. (1994). "no tail (ntl) is the zebrafish homologue of the mouse T (Brachyury) gene." Development **120**(4): 1009-15.

Schwarz, M., et al. (2000). "Spatial specification of mammalian eye territories by reciprocal transcriptional repression of Pax2 and Pax6." Development **127**(20): 4325-34.

Shalaby, F., et al. (1995). "Failure of blood-island formation and vasculogenesis in Flk-1-deficient mice." Nature **376**(6535): 62-6.

Shang, Y., et al. (2002). "Formation of the androgen receptor transcription complex." Mol Cell **9**(3): 601-10.

Shapiro, M. D., et al. (2004). "Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks." Nature **428**(6984): 717-23.

Sharov, A. A., et al. (2005). "Genome-wide assembly and analysis of alternative transcripts in mouse." Genome Res **15**(5): 748-54.

Sharpe, J., et al. (1998). "Selectivity, sharing and competitive interactions in the regulation of Hoxb genes." Embo J **17**(6): 1788-98.

Sherr, C. J. (1993). "Mammalian G1 cyclins." Cell **73**(6): 1059-65.

Simon, M. C., et al. (1988). "Definition of multiple, functionally distinct TATA elements, one of which is a target in the hsp70 promoter for E1A regulation." Cell **52**(5): 723-9.

Sivak, L. E., et al. (1999). "A novel intron element operates posttranscriptionally To regulate human N-myc expression." Mol Cell Biol **19**(1): 155-63.

Smale, S. T., et al. (1989). "The "initiator" as a transcription control element." Cell **57**(1): 103-13.

Solis, C., et al. (2001). "Uroporphyrinogen III synthase erythroid promoter mutations in adjacent GATA1 and CP2 elements cause congenital erythropoietic porphyria." J Clin Invest **107**(6): 753-62.

Song, D. L., et al. (1996). "Two Pax-binding sites are required for early embryonic brain expression of an Engrailed-2 transgene." Development **122**(2): 627-35.

Spicuglia, S., et al. (2002). "Promoter activation by enhancer-dependent and -independent loading of activator and coactivator complexes." Mol Cell **10**(6): 1479-87.

Spilianakis, C. G., et al. (2005). "Interchromosomal associations between alternatively expressed loci." Nature **435**(7042): 637-45.

Spitz, F., et al. (2003). "A global control region defines a chromosomal regulatory landscape containing the HoxD cluster." Cell **113**(3): 405-17.

Stern, D. L. (1998). "A role of Ultrabithorax in morphological differences between Drosophila species." Nature **396**(6710): 463-6.

Stickney, H. L., et al. (2002). "Rapid mapping of zebrafish mutations with SNPs and oligonucleotide microarrays." Genome Res **12**(12): 1929-34.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.

Strahle, U., et al. (1993). "Axial, a zebrafish gene expressed along the developing body axis, shows altered expression in cyclops mutant embryos." Genes Dev **7**(7B): 1436-46.

Strahle, U., et al. (1996). "Expression of axial and sonic hedgehog in wildtype and midline defective zebrafish embryos." Int J Dev Biol **40**(5): 929-40.

Stuart, G. W., et al. (1988). "Replication, integration and stable germ-line transmission of foreign sequences injected into early zebrafish embryos." Development **103**(2): 403-12.

Stuart, G. W., et al. (1985). "Identification of multiple metal regulatory elements in mouse metallothionein-I promoter by assaying synthetic sequences." Nature **317**(6040): 828-31.

Stuart, G. W., et al. (1990). "Stable lines of transgenic zebrafish exhibit reproducible patterns of transgene expression." Development **109**(3): 577-84.

Sucena, E., et al. (2000). "Divergence of larval morphology between Drosophila sechellia and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby." Proc Natl Acad Sci U S A **97**(9): 4530-4.

Sumanas, S., et al. (2005). "Identification of novel vascular endothelial-specific genes by the microarray analysis of the zebrafish cloche mutants." Blood **106**(2): 534-41.

Sutherland, H. G., et al. (1997). "A globin enhancer acts by increasing the proportion of erythrocytes expressing a linked transgene." Mol Cell Biol **17**(3): 1607-14.

Suzuki, Y., et al. (2001). "Identification and characterization of the potential promoter regions of 1031 kinds of human genes." Genome Res **11**(5): 677-84.

Suzuki, Y., et al. (2002). "DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs." Nucleic Acids Res **30**(1): 328-31.

Swanson, M. S., et al. (1988). "Classification and purification of proteins of heterogeneous nuclear ribonucleoprotein particles by RNA-binding specificities." Mol Cell Biol **8**(5): 2237-41.

Szutorisz, H., et al. (2005). "The role of enhancers as centres for general transcription factor recruitment." Trends Biochem Sci **30**(11): 593-9.

Tabach, Y., et al. (2007). "Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site." PLoS ONE **2**(8): e807.

Taft, R. J., et al. (2007). "The relationship between non-protein-coding DNA and eukaryotic complexity." Bioessays **29**(3): 288-99.

Takihara, Y., et al. (1986). "A novel mutation in the TATA box in a Japanese patient with beta +-thalassemia." Blood **67**(2): 547-50.

Talbot, W. S., et al. (1995). "A homeobox gene essential for zebrafish notochord development." Nature **378**(6553): 150-7.

Tanabe, Y., et al. (1998). "Specification of motor neuron identity by the MNR2 homeodomain protein." Cell **95**(1): 67-80.

Tebb, G., et al. (1989). "The Xenopus laevis U2 gene distal sequence element (enhancer) is composed of four subdomains that can act independently and are partly functionally redundant." Mol Cell Biol **9**(4): 1682-90.

Thermes, V., et al. (2002). "I-SceI meganuclease mediates highly efficient transgenesis in fish." Mech Dev **118**(1-2): 91-8.

Thisse, B., Pflumio, S., Fürthauer, M., Loppin, B., Heyer, V., Degrave, A., Woehl, R., Lux, A., Steffan, T., Charbonnier, X.Q. and Thisse, C. (2001). "Expression of the zebrafish genome during embryogenesis." ZFIN Direct Data Submission: http://zfin.org.

Thisse, B., Thisse, C (2004). "Fast Release Clones: A High Throughput Expression Analysis." ZFIN Direct Data Submission.

Thisse, C., et al. (2008). "High-resolution in situ hybridization to whole-mount zebrafish embryos." Nat Protoc **3**(1): 59-69.

Thor, S., et al. (1997). "The Drosophila islet gene governs axon pathfinding and neurotransmitter identity." Neuron **18**(3): 397-409.

Thummel, R., et al. (2005). "Cre-mediated site-specific recombination in zebrafish embryos." Dev Dyn **233**(4): 1366-77.

Tokusumi, Y., et al. (2007). "The new core promoter element XCPE1 (X Core Promoter Element 1) directs activator-, mediator-, and TATA-binding protein-dependent but TFIID-independent RNA polymerase II transcription from TATA-less promoters." Mol Cell Biol **27**(5): 1844-58.

Ton, C., et al. (2002). "Construction of a zebrafish cDNA microarray: gene expression profiling of the zebrafish during development." Biochem Biophys Res Commun **296**(5): 1134-42.

Tora, L. (2002). "A unified nomenclature for TATA box binding protein (TBP)-associated factors (TAFs) involved in RNA polymerase II transcription." Genes Dev **16**(6): 673-5.

Torres, M., et al. (1996). "Pax2 contributes to inner ear patterning and optic nerve trajectory." Development **122**(11): 3381-91.

Tran, T. C., et al. (2007). "Automated, quantitative screening assay for antiangiogenic compounds using transgenic zebrafish." Cancer Res **67**(23): 11386-92.

Trinklein, N. D., et al. (2004). "An abundance of bidirectional promoters in the human genome." Genome Res **14**(1): 62-6.

Trinklein, N. D., et al. (2003). "Identification and functional analysis of human transcriptional promoters." Genome Res **13**(2): 308-12.

Tuan, D., et al. (1992). "Transcription of the hypersensitive site HS2 enhancer in erythroid cells." Proc Natl Acad Sci U S A **89**(23): 11219-23.

Tufarelli, C., et al. (2003). "Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease." Nat Genet **34**(2): 157-65.

Tuggle, C. K., et al. (1990). "Region-specific enhancers near two mammalian homeo box genes define adjacent rostrocaudal domains in the central nervous system." Genes Dev **4**(2): 180-9.

Uchikawa, M., et al. (2004). "Efficient identification of regulatory sequences in the chicken genome by a powerful combination of embryo electroporation and genome comparison." Mech Dev **121**(9): 1145-58.

Uemura, O., et al. (2005). "Comparative functional genomics revealed conservation and diversification of three enhancers of the isl1 gene for motor and sensory neuron-specific expression." Dev Biol **278**(2): 587-606.

van Boxtel, A. L., et al. (2008). "Microarray analysis reveals a mechanism of phenolic polybrominated diphenylether toxicity in zebrafish." Environ Sci Technol **42**(5): 1773-9.

van Eeden, F. J., et al. (1999). "Developmental mutant screens in the zebrafish." Methods Cell Biol **60**: 21-41.

van Heyningen, V., et al. (2002). "PAX6 in sensory development." Hum Mol Genet **11**(10): 1161-7.

Van Leeuwen, C. J., et al. (1990). "Fish embryos as teratogenicity screens: a comparison of embryotoxicity between fish and birds." Ecotoxicol Environ Saf **20**(1): 42-52.

Vavouri, T., et al. (2005). "Prediction of cis-regulatory elements using binding site matrices--the successes, the failures and the reasons for both." Curr Opin Genet Dev **15**(4): 395-402.

Vavouri, T., et al. (2006). "Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key." Trends Genet **22**(1): 5-10.

Veenstra, G. J., et al. (2000). "Distinct roles for TBP and TBP-like factor in early embryonic gene transcription in Xenopus." Science **290**(5500): 2312-5.

Venkatesh, B., et al. (2006). "Ancient noncoding elements conserved in the human genome." Science **314**(5807): 1892.

Venter, J. C., et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.

Vilar, J. M., et al. (2005). "DNA looping in gene regulation: from the assembly of macromolecular complexes to the control of transcriptional noise." Curr Opin Genet Dev **15**(2): 136-44.

Visel, A., et al. (2007a). "Regulatory pathway analysis by high-throughput in situ hybridization." PLoS Genet **3**(10): 1867-83.

Visel, A., et al. (2007b). "VISTA Enhancer Browser--a database of tissue-specific human enhancers." Nucleic Acids Res **35**(Database issue): D88-92.

Visel, A., et al. (2008). "Ultraconservation identifies a small subset of extremely constrained developmental enhancers." Nat Genet **40**(2): 158-60.

Voigt, S., et al. (1996). "Signal sequence-dependent function of the TRAM protein during early phases of protein transport across the endoplasmic reticulum membrane." J Cell Biol **134**(1): 25-35.

Vong, L. H., et al. (2005). "Generation of conditional Mef2cloxP/loxP mice for temporal- and tissue-specific analyses." Genesis **43**(1): 43-8.

Walhout, A. J., et al. (2000). "GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes." Methods Enzymol **328**: 575-92.

Wallis, D. E., et al. (1999). "Mutations in the homeodomain of the human SIX3 gene cause holoprosencephaly." Nat Genet **22**(2): 196-8.

Walther, C., et al. (1991). "Pax: a murine multigene family of paired box-containing genes." Genomics **11**(2): 424-34.

Walton, R. Z., et al. (2006). "Fog1 is required for cardiac looping in zebrafish." Dev Biol **289**(2): 482-93.

Wang, H. L., et al. (2000). "Point mutation associated with X-linked dominant Charcot-Marie-Tooth disease impairs the P2 promoter activity of human connexin-32 gene." Brain Res Mol Brain Res **78**(1-2): 146-53.

Wang, J. C., et al. (1988). "Action at a distance along a DNA." Science **240**(4850): 300-4.

Wang, W., et al. (2007). "A fully automated robotic system for microinjection of zebrafish embryos." PLoS ONE **2**(9): e862.

Wang, X., et al. (2004). "Evolution of regulatory elements producing a conserved gene expression pattern in Caenorhabditis." Evol Dev **6**(4): 237-45.

Wardle, F. C., et al. (2006). "Zebrafish promoter microarrays identify actively transcribed embryonic genes." Genome Biol **7**(8): R71.

Wasserman, W. W., et al. (2000). "Human-mouse genome comparisons to locate regulatory sites." Nat Genet **26**(2): 225-8.

Wasserman, W. W., et al. (2004). "Applied bioinformatics for the identification of regulatory elements." Nat Rev Genet **5**(4): 276-87.

Wefald, F. C., et al. (1990). "Functional heterogeneity of mammalian TATA-box sequences revealed by interaction with a cell-specific enhancer." Nature **344**(6263): 260-2.

Wei, W., et al. (2001). "The gypsy insulator can act as a promoter-specific transcriptional stimulator." Mol Cell Biol **21**(22): 7714-20.

Weinberg, E. S., et al. (1996). "Developmental regulation of zebrafish MyoD in wild-type, no tail and spadetail embryos." Development **122**(1): 271-80.

Weis, L., et al. (1997). "Accurate positioning of RNA polymerase II on a natural TATA-less promoter is independent of TATA-binding-protein-associated factors and initiator-binding proteins." Mol Cell Biol **17**(6): 2973-84.

West, A. G., et al. (2005). "Remote control of gene transcription." Hum Mol Genet **14 Spec No 1**: R101-11.

Westerfield, M. (1993). "THE ZEBRAFISH BOOK, A guide for the laboratory use of zebrafish (Danio rerio)." Book.

Westerfield, M., et al. (1992). "Specific activation of mammalian Hox promoters in mosaic transgenic zebrafish." Genes Dev **6**(4): 591-8.

White, K. P., et al. (1999). "Microarray analysis of Drosophila development during metamorphosis." Science **286**(5447): 2179-84.

Wienholds, E., et al. (2005). "MicroRNA expression in zebrafish embryonic development." Science **309**(5732): 310-1.

Wienholds, E., et al. (2002). "Target-selected inactivation of the zebrafish rag1 gene." Science **297**(5578): 99-102.

Wild, A., et al. (1997). "Point mutations in human GLI3 cause Greig syndrome." Hum Mol Genet **6**(11): 1979-84.

Wilson, S. W., et al. (1990). "The development of a simple scaffold of axon tracts in the brain of the embryonic zebrafish, Brachydanio rerio." Development **108**(1): 121-45.

Winkler, C., et al. (1992). "Analysis of heterologous and homologous promoters and enhancers in vitro and in vivo by gene transfer into Japanese medaka (Oryzias latipes) Xiphophorus." Mol Mar Biol Biotechnol **1**(4-5): 326-37.

Winkler, C., et al. (1991). "Transient expression of foreign DNA during embryonic and larval development of the medaka fish (Oryzias latipes)." Mol Gen Genet **226**(1-2): 129-40.

Winkler, G. S., et al. (2001). "RNA polymerase II elongator holoenzyme is composed of two discrete subcomplexes." J Biol Chem **276**(35): 32743-9.

Wittkopp, P. J., et al. (2003). "Drosophila pigmentation evolution: divergent genotypes underlying convergent phenotypes." Proc Natl Acad Sci U S A **100**(4): 1808-13.

Woolfe, A., et al. (2007a). "Comparative genomics using Fugu reveals insights into regulatory subfunctionalization." Genome Biol **8**(4): R53.

Woolfe, A., et al. (2007b). "CONDOR: a database resource of developmentally associated conserved non-coding elements." BMC Dev Biol **7**: 100.

Woolfe, A., et al. (2005). "Highly conserved non-coding sequences are associated with vertebrate development." PLoS Biol **3**(1): e7.

Wray GA, et al. (2003). "The evolution of transcriptional regulation in eukaryotes." Mol Biol Evol. **20**(9): 1377-419. .

Wu, C. (1984). "Activating protein factor binds in vitro to upstream control sequences in heat shock gene chromatin." Nature **311**(5981): 81-4.

Xie, X., et al. (2006). "A family of conserved noncoding elements derived from an ancient transposable element." Proc Natl Acad Sci U S A **103**(31): 11659-64.

Xiong, N., et al. (2002). "Redundant and unique roles of two enhancer elements in the TCRgamma locus in gene regulation and gammadelta T cell development." Immunity **16**(3): 453-63.

Xu, J., et al. (2006). "Genomewide expression profiling in the zebrafish embryo identifies target genes regulated by Hedgehog signaling during vertebrate development." Genetics **174**(2): 735-52.

Yamamoto, A., et al. (1998). "Zebrafish paraxial protocadherin is a downstream target of spadetail involved in morphogenesis of gastrula mesoderm." Development **125**(17): 3389-97.

Yanagisawa, H., et al. (2003). "Targeted deletion of a branchial arch-specific enhancer reveals a role of dHAND in craniofacial development." Development **130**(6): 1069-78.

Yang, A., et al. (2002). "On the shoulders of giants: p63, p73 and the rise of p53." Trends Genet **18**(2): 90-5.

Yang, C., et al. (2007a). "Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters." Gene **389**(1): 52-65.

Yang, L., et al. (2007b). "Transcriptional profiling reveals barcode-like toxicogenomic responses in the zebrafish embryo." Genome Biol **8**(10): R227.

Yang, W. S., et al. (1995). "A mutation in the promoter of the lipoprotein lipase (LPL) gene in a patient with familial combined hyperlipidemia and low LPL activity." Proc Natl Acad Sci U S A **92**(10): 4462-6.

Yeh, F. L., et al. (2002). "Differential regulation of spontaneous and heat-induced HSP 70 expression in developing zebrafish (Danio rerio)." J Exp Zool **293**(4): 349-59.

Yelon, D., et al. (1999). "Restricted expression of cardiac myosin genes reveals regulated aspects of heart tube assembly in zebrafish." Dev Biol **214**(1): 23-37.

Zakany, J., et al. (1997). "Deletion of a HoxD enhancer induces transcriptional heterochrony leading to transposition of the sacrum." Embo J **16**(14): 4393-402.

Zanella, C., et al. (2007). "Segmentation of cells from 3-D confocal images of live zebrafish embryo." Conf Proc IEEE Eng Med Biol Soc **2007**: 6028-31.

Zavolan, M., et al. (2002). "Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome." Genome Res **12**(9): 1377-85.

Zerucha, T., et al. (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**(2): 709-21.

Zhou, J., et al. (1999). "A novel cis-regulatory element, the PTS, mediates an anti-insulator activity in the Drosophila embryo." Cell **99**(6): 567-75.

Zon, L. I. (1999). "Zebrafish: a new model for human disease." Genome Res **9**(2): 99-100.

Zuniga, A., et al. (2004). "Mouse limb deformity mutations disrupt a global control region within the large regulatory landscape required for Gremlin expression." Genes Dev **18**(13): 1553-64.