Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Wissenschaftliche Berichte
FZKA 7341

Phylogenomic and Functional Analyses of Enhancer Evolution of *Sonic Hedgehog* Paralogs

Y. Hadzhiev

Institut für Toxikologie und Genetik

September 2007

Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Wissenschaftliche Berichte

FZKA 7341

Phylogenomic and functional analyses of enhancer evolution of *sonic hedgehog* paralogs

Yavor Hadzhiev

Institut für Toxikologie und Genetik

Von der Fakultät für Chemie und Biowissenschaften
der Universität Karlsruhe (TH) genehmigte Dissertation

Forschungszentrum Karlsruhe GmbH, Karlsruhe
2007

Forschungszentrum Karlsruhe GmbH
Postfach 3640, 76021 Karlsruhe

Mitglied der Hermann von Helmholtz-Gemeinschaft
Deutscher Forschungszentren (HGF)

ISSN 0947-8620

urn:nbn:de:0005-073410

Phylogenomic and functional analyses of enhancer evolution of *sonic hedgehog* paralogs

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

(Dr. rer. nat.)

von der Fakultät für Chemie und Biowissenschaften der

Universität Karlsruhe (TH) genemigte

DISSERTATION

von

Yavor Hadzhiev

aus Smolyan, Bulgarien

Dekan: Prof. Dr. Holger Puchta
Referent: Prof. Dr. Andrew Cato
Korreferent: Prof. Dr. Martin Bastmeyer
Tag der mündlichen Prüfung: 23.-27. April 2007

Abstract

Background

***Cis*-regulatory modules (CRM) of developmental genes are targets of evolutionary changes that underlie morphological diversity of animals (Carroll et al. 2000). Due to our lack of knowledge of the "grammar" of transcription factor/CRM interactions, still little is known about the molecular mechanisms that underlie the changes that take place in the CRMs of genes, particularly after gene and genome duplications.**

To this end, we investigated the *ar-C* midline enhancer of *sonic hedgehog (shh)* orthologs and paralogs of distantly related vertebrate lineages from fish to human including the basal vertebrate *Latimeria*.

Results

We demonstrate that the *shh* paralog *tiggy winkle hh* genes of fishes have a modified *ar-C* enhancer which specifies a diverged function at the embryonic midline. We have identified several conserved motifs indicative of putative transcription factor binding sites by a local alignment of *ar-C* enhancers of numerous vertebrate sequences. To trace the evolutionary changes among paralog enhancers phylogenomic reconstruction was carried out and lineage-specific motif changes were identified. The relevance of the motif composition to observed developmental differences was studied through transgenic functional analyses. Altering and exchanging motifs between paralog enhancers resulted in the reversal of enhancer specificity in the floor plate and notochord.

Conclusions

By functional analysis of *shh* paralog enhancers a reconstruction model for enhancer divergence during vertebrate evolution was developed. Our model suggests that the identified motifs of the *ar-C* enhancer function as binary switches responsible for specific activity between midline tissues and that these motifs are adjusted during functional diversification of paralogs. The unravelled motif changes can also account for the complex interpretation of activator and repressor input signals within a single enhancer.

Phylogenomische und funktionelle Analyse der Enhancerevolution von *sonic hedgehog* Paralogen

Zusammenfassung

Hintergrund

***Cis*-regulatorische Module (CRM) entwicklungsspezifischer Gene sind Ziel evolutionärer Veränderungen, die der morphologischen Vielfalt von Tieren zugrunde liegen (Carroll et al. 2000). Aufgrund des fehlenden Wissen über die „Grammatik" der Interaktion von Transcriptionsfaktoren mit CRMs ist immer noch wenig über die molekularen Mechanismen bekannt, denen diese Veränderungen in CRMs unterliegen, besonders im Hinblick auf die Duplikation von Genen bzw. Genomen.**

Wir haben daher den *ar-C* Mittellinie Enhancer von *sonic hedgehog (shh)* Orthologen und Paralogen entfernter Abstammungslinien vom Fisch bis zum Mensch inklusive des ursprünglichen Vertebraten *Latimeria* auf diese Mechanismen hin untersucht.

Ergebnisse

Wir konnten zeigen, dass das *shh* paralog *tiggy winkle hedgehog* in Fischen einen modifizierten *ar-C* Enhancer besitzt, der eine abweichende Funktion in der embryonalen Mittellinie aufweist. Weiterhin konnten wir durch lokale Sequenz-Alignments von *ar-C* Enhancern verschiedener vertebrater Spezies mehrere konservierte Motive identifizieren, die auf Bindestellen für Transkriptionsfaktoren hinweisen. Um die evolutionären Veränderungen innerhalb paraloger Enhancer zu verfolgen, wurden phylogenomische Rekonstruktionen durchgeführt und abstammungsspezifische Veränderungen in den Motiven identifiziert. Die Relevanz der Motivzusammensetzung im Vergleich zur beobachteten Veränderung in der Entwicklung, wurde anhand funktioneller Analysen in transgenen Zebrafischen studiert. Die Veränderung und der Austausch von paralogen Enhancern führte zu einer Umkehrung der Spezifität des Enhancers in der Bodenplatte und des Notochords.

Schlussfolgerung

Durch die funktionelle Analyse von paralogen shh Enhancern konnte ein rekonstruierendes Modell zur Divergenz von Enhancern in der Evolution der Vertebraten entwickelt werden. Unser Modell legt den Schluss nahe, dass die identifizierten Motive des *ar-C* Enhancers als binäre Schalter fungieren, verantwortlich für die spezifische Aktivität zwischen Mittellinie Geweben und dass diese Motive während der funktionellen Diversifizierung der Paraloge angepasst wurden. Die aufgedeckten Motivänderungen könnten auch auf die komplexe Interpretation von Aktivator und Repressor Eingangssignalen innerhalb eines einzigen Enhancers hinweisen.

Contents

Abbreviations

To be consistent with the accepted nomenclature, regarding the font format of the gene and protein names, certain rules has been followed:

- The gene name, including gene enhancers and Latin names, are always in lower-case and italic.
- The names of the fish proteins are in sentence case and regular font.
- The names of proteins of other organisms are in upper-case and regular font

ar	*Activation region*
BOC	brother of CDO
BOI	Brother of IHOG
CAM	Calmodulin
CDO	CAM-related 1 down-regulated by oncogenes
CI	Cubitus interruptus
CIA	Cubitus interruptus activator form
CIR	Cubitus interruptus repressor form
CK1	Casein Kinase I
COS-2	Costal-2
CRM	Conserved regulatory module
Cyc	Cyclops
DDC-model	Duplication-degeneration-complementation model
dhh	*Desert hedgehog*
DV	Dorso-ventral
ehh	*Echidna hedgehog*

fp	Floor plate
FU	Fused
Gli	Glioma-associated oncogene homologue
GSK-3	Glycogen Synthase Kinase-3
Hip	Hedgehog interacting protein
HNF3β	Hepatocyte nuclear factor 3 beta
hyp	Hypothalamus
ihh	*Indian hedgehog*
Ihog	Interference hedgehog
LR	Left-right
MAR	matrix attachment region
nt	Notochord
ntn1	*netrin1*
PKA	Protein kinase A
PTC, Ptch	Patched
qhh	Qiqihar hedgehog
SBE	*Sonic brain enhancer*
SFPE	*Sonic floor plate enhancer*
shh	*Sonic hedgehog*
Smo	Smoothened
Su(fu)	Suppressor of fused
TF	Transcription factor
TFB	Transcription factor bindig site
twhh	*Tiggy winkle hedgehog*
zli	*Zona limitans intrathalamica*

Introduction

Zebrafish as a model organism

Due to several advantageous properties, the zebrafish (*Danio rerio*) has become a well-established model organism for studies in development and genetics. The embryos develop externally and are therefore amenable for experimental manipulations and microscopic observations. These are facilitated by the transparency of the embryos. Zebrafish is highly fertile (one fish-pair can lay appr. 150-200 eggs at optimal conditions) and has a short generation cycle of approximately three months. The development of the embryo is fast: within 48 hours a free-swimming larvae has grown-up from the fertilised egg, facilitating genetic analysis. Furthermore, forward genetic screens have provided several thousand mutations affecting genes involved in many developmental processes (Driever et al. 1996; Haffter et al. 1996). In the last years high efficient systems have been established for transgenesis and enhancer traps, based on retroviral (Amsterdam and Becker 2005; Gaiano et al. 1996a) and transposon (Kawakami et al. 1998; Kawakami et al. 2004; Parinov et al. 2004) insertions. With the "TILLING" technique (Wienholds et al. 2003) based on the PCR screening of a large number of mutated alleles and the retrovirus based insertional mutagenesis (Amsterdam 2006; Amsterdam and Hopkins 2004; Gaiano et al. 1996b), the researchers can now commercially obtain mutants specifically in a given gene of interest. In addition, knock-down techniques using morpholino oligonucleotides are widely used, allowing the specific inactivation of the studied genes (Nasevicius and Ekker 2000) or gain/lost of function approaches by mRNA microinjection. The advantages mentioned above make zebrafish a unique vertebrate model system for fast, large scale, *in vivo* promoter and enhancer screens using fluorescent proteins as reporter genes. The genetic analyses 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.

Embryonic development of zebrafish

Development proceeds upon fertilisation, after which the extra-embryonic chorion lifts from the zygote and yolk-free cytoplasm streams to the animal pole of the cell to form the blastodisc. The blastodisc sits on top of a yolk sac. The subsequent cleavages are meroblastic, e.g. they only take place at the animal pole of the embryo. These first cleavages are synchronous, producing about 512 cells within 2 3/4 hours, forming a blastula. At that stage begins the midblastula transition (MBT). This process is characterised by activation of zygotic transcription, cell cycle lengthening, loss of cell synchrony, and appearance of cell motility (Kane and Kimmel 1993). Subsequently, gastrulation starts at about 4 hpf with the epiboly, a process of cell migration around the yolk sac. At about 6 hpf, involution takes place, a process by

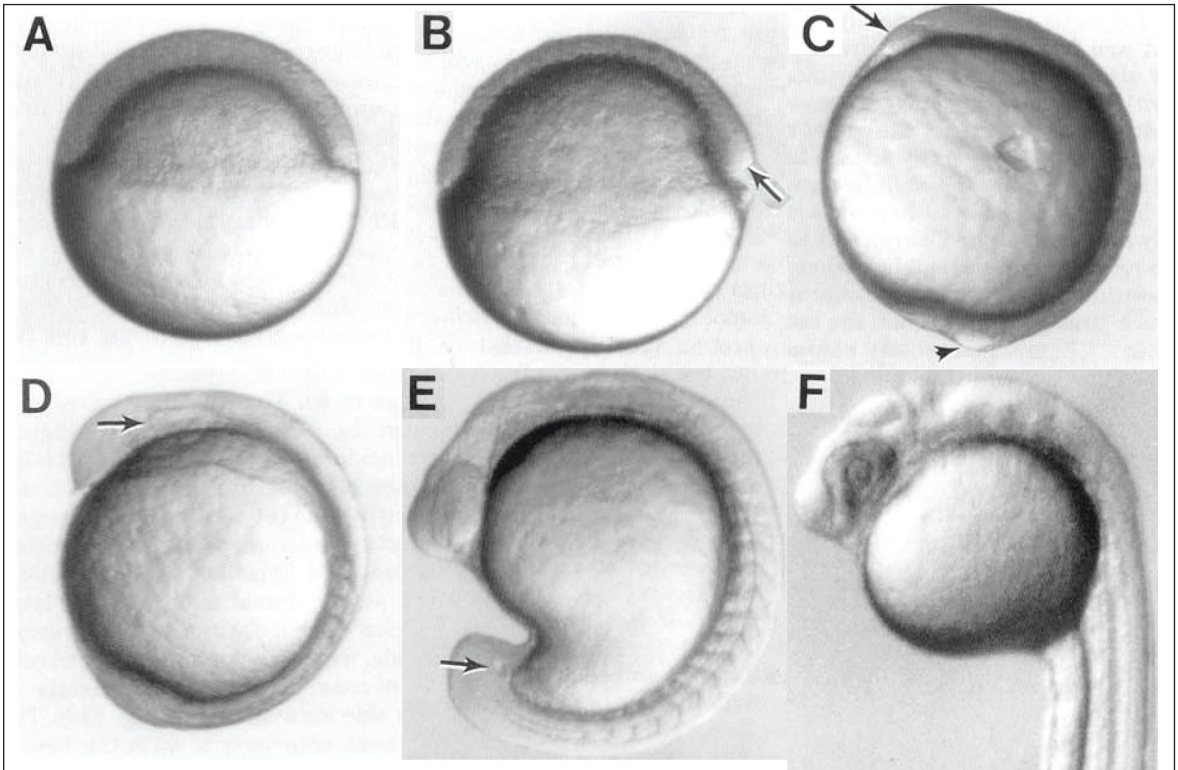

Figure 1. **Selected stages from zebrafish embryonic development.** A, 50 % epiboly (5,25 hpf). B, Shield stage (6 hpf), the arrow points at the embryonic shield, the dorsal organiser. C, Bud stage (10 hpf), the arrow shows the polster, and the arrowhead shows the tail bud. D, 4-somite (11,3 hpf), the optic primordium is marked by arrow. E, 15-somite stage (16.5 h). The arrow shows Kupffer's vesicle, ciliated organ controlling the asymmetry. F, Prim-5 stage (24 h). The arrow points to the notochord, the arrowhead to floor plate. Modified form Kimmel et al. 1995.

which cells at the future dorsal side of the embryo, or shield, start to migrate underneath the overlying cells. In addition, cells converge towards the midline and extend anteriorly. The gastrulation process will finally give rise to the 3 different germ layers of the embryo and is finished by reaching the tailbud stage at 10 hpf. During this gastrulation period, the main body axes are specified. After the tailbud stage, segmentation starts with the formation of the somites, mesodermal blocks of tissue, which will give, rise to several organs, such as muscle, spleen and blood. Furthermore, the process of neurulation subdivides the ectodermal neural plate into a regionalised neural tube. At the end of the segmentation period, at about 24 hpf, most of the organ primordia are specified and the first functioning organ, the beating heart, becomes visible. The embryo now enters the pharyngula period and hatches at about 48 hpf. An overview of selected developmental stages of zebrafish is shown on Fig. 1.

Specification and function of the embryonic midline structures: notochord and floor plate

Notochord

The notochord is an embryonic midline structure common to all members of the Chordata. The notochord is positioned centrally in the embryo with respect to both the dorsal-ventral (DV) and left-right (LR) axes (Fig. 2) (Cleaver and Krieg 2001; Stemple 2004; Stemple 2005). In some vertebrate clades, such as the agnathans (lampreys), and in primitive bony fish, such as sturgeons, the notochord persists throughout life. In higher vertebrates, however, the notochord becomes ossified in regions of forming vertebrae and contributes to the centre of the interver-

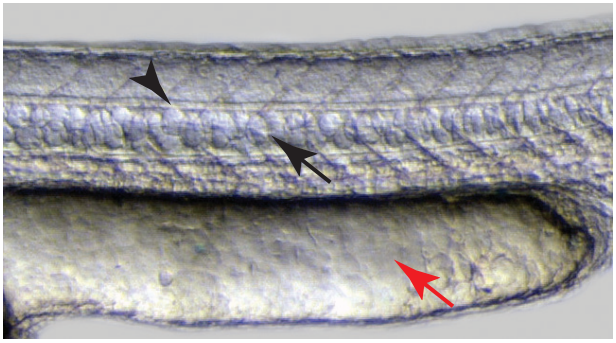

Figure 2. High magnification of the trunk (on the level of the yolk extension, indicated by red arrow) of 24 h old zebrafish embryo. The notochord and the floor plate are clearly visible as pointed by the black arrow and arrowhead respectively.

tebral discs in a structure called the nucleus pulposis (Linsenmayer et al. 1986; Smits and Lefebvre 2003; Swiderski and Solursh 1992). The notochord formation starts in early gastrula stages. It arises from the dorsal organiser. Originally identified by Spemann and Mangold in amphibians, the dorsal organiser is a region of a vertebrate gastrula that, when transplanted into prospective lateral or ventral regions of a host embryo, induces the formation of a second embryonic axis (Harland and Gerhart 1997; Spemann and Mangold 1924). In amphibians, this region is the dorsal lip of the blastopore. In other species, homologous structures have been found: the embryonic shield of teleost fish, Hensen's node in the chick and the node of mouse embryos all possess essentially the same activities as Spemann and Mangold's dorsal organiser (Beddington 1994; Oppenheimer 1936; Waddington 1930). The functions and activities of the dorsal organiser are complex and have been discussed in (Harland and Gerhart 1997).

The notochord has at least two important functions. First, it plays an important structural role. As a tissue, it is most closely related to cartilage and is likely to represent a primitive form of cartilage and serves as the axial skeleton of the embryo until other elements, such as the vertebrae, form. Second, it produces secreted factors that signal to all surrounding tissues, providing position and fate information. In this role, the notochord is important for specifying ventral fates in the central nervous system, controlling aspects of LR asymmetry, patterning of the underlying endoderm, including pancreas development, the arterial versus venous identity of the major axial blood vessels and specifying a variety of cell types in forming somites (Christ et al. 2004; Danos and Yost 1995; Fouquet et al. 1997; Goldstein and Fishman 1998; Lohr et al. 1997; Munsterberg and Lassar 1995; Pourquie et al. 1993; Yamada et al. 1993; Yamada et al. 1991). The best characterised is the role of the notochord in patterning of the neural tube. Among the signals secreted by the notochord are the Hedgehog (Hh) proteins. Sonic hedgehog (Shh), in particular, induces a range of ventral spinal cord fates in a graded fashion while simultaneously suppressing the expression of dorsal genes (Placzek et al. 1993; Yamada et al. 1993;

Yamada et al. 1991). In addition to Shh, studies in different organisms provide a long list of growth factors and secreted signalling molecules expressed in the notochord, including members of the Bone morphogenic protein (Bmp) family, Transforming growth factor beta (TGF-β) family: TGF-, Fibroblast growth factor (Fgf) family, Nodal-related 2 (Ndr2), Follistatin, Noggin, Activin, Chordin, and Hedgehog interacting protein (Hip) (Cheng et al. 2000; Chuang and McMahon 1999; Dale et al. 1999; Dudley and Robertson 1997; Harland et al. 1992; Hemmati-Brivanlou et al. 1994; Isaacs et al. 1995; Joseph and Melton 1997; Kondaiah et al. 2000; Marti 2000; Rebagliati et al. 1998; Sasai et al. 1994; Shamim et al. 1999; Smith and Harland 1992). Although the precise roles of these potent signalling molecules secreted from the notochord are not completely understood, it seems that at least some will be important for the development of the adjacent tissues.

Floor Plate

The floor plate (FP) is a transient embryonic glial structure, located at the most ventral midline of the neural tube (Fig. 2). It functions as an organiser, which controls the development of the vertebrate central nervous system (Dodd et al. 1998; Placzek and Briscoe 2005; Strahle et al. 2004). The zebrafish FP is three-cell-wide and can be subdivided into the one-cell-wide medial floor plate (mFP) and the lateral floor plate (lFP) (Odenthal et al. 2000; Strahle et al. 1996), which differ in gene expression. The one-cell-wide mFP represents the neural tube organiser as it expresses the patterning signals *sonic hedgehog* (*shh*), its paralog, *tiggy winkle hedgehog* (*twhh*) (Ekker et al. 1995) and *netrin1* (*ntn1*) (Korzh 2001), signalling protein directing the axonal trajectories of commissural interneurons and certain motor neurons at the ventral midline of the neural tube (Colamarino and Tessier-Lavigne 1995). Expression of the winged-helix transcription factor *foxa2* (Korzh 2001) (also known as *axial/HNF3β*) is detectable in both mFP and lFP, whereas the homeobox transcription factor *nkx2.2* is expressed in lFP cells exclusively (Barth and Wilson 1995; Schafer et al. 2005). The FP in mouse is broader than that of zebrafish but also can be subdivided

into a medial and a lateral domain. The medial domain expresses *shh*, whereas *foxa2* transcripts are present in medial and lateral regions (Marti et al. 1995). A slightly different picture is presented in chicken neural tube. In contrast to the neural tube of zebrafish and mouse, *shh*, *ntr1* and *foxa2* are co-expressed in the FP. At later stages, *foxa2* becomes restricted to the mFP, whereas expression of shh extends into adjacent cells expressing *nkx2.2* (Charrier et al. 2002).

The origin and specification of the FP is not completely understood. Ectopic FP cells were induced in neural tissues when pieces of the and prechordal plate were grafted adjacent to the dorsolateral region of the neural tube in chicken embryo, or combined with neural plate explants *in vitro* (Placzek et al. 2000). Application of recombinant SHH to neural tissue had a similar effect, indicating that the mediator of FP induction by the notochord is SHH (Marti et al. 1995; Roelink et al. 1995). Moreover, *shh* gene knockout in mouse abolished FP differentiation (Chiang et al. 1996). These observations led to the proposal of a model for FP induction in which SHH secreted from the notochord and prechordal plate induces the FP in the overlying neuroectoderm (Dodd et al. 1998; Placzek et al. 2000). However, this induction model does not correlate with findings from genetic analyses in zebrafish FP. In this species, Shh plays a minor role in the initial specification of the FP (Briscoe et al. 2001; Etheridge et al. 2001; Neumann et al. 1999; Varga et al. 2001) and most likely the Cyclops (Nodal) signalling in zebrafish plays similar role as Shh in mouse (Dodd et al. 1998; Feldman et al. 1998; Gritsman et al. 1999; Pogoda et al. 2000; Sirotkin et al. 2000; Zhang et al. 1998). In addition, has been shown that recombinant Nodal protein can act synergistically with Shh to induce FP in chicken neural plate explants, suggesting a function of Nodals in the FP induction in amniotes as well. The induction model was also challenged by observations in chick and quail chimeric embryos suggesting that notochord and FP are derived from the same precursor cells located in Hensen's node, the chicken organiser, equivalent of the embryonic shield in zebrafish.(Catala et al. 1996; Le Douarin and Halpern 2000). Taken together, these studies suggest that induction of the FP in zebrafish embryos starts to differentiate

by early gastrulation already, before the notochord has formed entity. This process, which is independent of Shh, requires Nodal signaling. Shh plays a role in lateral expansion of the FP and appears to have a maintenance function once the FP has formed. This is inconsistent with the previously proposed floor induction model in amniotes and suggests a more complex mechanism in fishes.

Transcriptional regulation in eukaryotes

Only some of the genes in a eukaryotic cell are expressed at any given moment. The proportion and composition of transcribed genes changes during the life cycle, among cell types, and in response to fluctuating physiological and environmental conditions (Arbeitman et al. 2002; Iyer et al. 2001; White et al. 1999). This differential gene expression requires complex specific interactions of macromolecules. Eukaryotes use diverse mechanisms to regulate gene expression, including chromatin condensation, DNA methylation, transcriptional initiation, alternative splicing of RNA, mRNA stability, translational controls, several forms of post-translational modification, intracellular trafficking, and protein degradation (Alberts 2002; Lewin 2000). The most common point of control is the rate of transcriptional initiation (Latchman 1997; Lemon and Tjian 2000; White 2001). Transcriptional regulators are classified in two major groups: cis-regulators and trans-regulators. The first are specific sequences in the DNA (usually in the proximity of the regulated gene), which are recognised and bound by proteins called transcription factors (TFs). The TFs are the trans-regulators and most of them belong to relatively small protein families. There are approximately 12 to 15 structurally distinct DNA-binding domains known from eukaryotic TF (Locker 2001). In zebrafish they are approximately 2000 TFs. Far less is known about the diversity and evolutionary history of transcription cofactors, proteins that bind to TFs but not to DNA. Transcription cofactors, by definition, lack a DNA-binding domain, but they typically contain domains that mediate a specific protein-protein association with a TFs and directly or indirectly interact with effector complexes. The binding of TFs to specific DNA sequences is achieved through

their DNA-binding domain, which is a short amino-acid motif, often highly conserved evolutionary, that usually inserts into the major groove of double-stranded DNA (Choo and Klug 1997; Jones et al. 1999; Locker 2001). Sequence-specific protein-DNA contacts rarely extend across more than 5 bp, and for some motifs, such as Zn-fingers, they extend only 3 bp. The extent of this physical interaction is not sufficient to provide much sequence specificity, as a given 5-bp sequence can occur on every 1,024 bp. Three structural features can increase DNA binding specificity (Wray et al. 2003): multiple DNA binding domains can exist within a single TFs (most Pax family members contain both paired-box and homeodomain DNA binding domains, whereas all Zn-finger TF contain multiple Zn-fingers); additional structural features can bind nearby nucleotides through minor groove contacts (many homeodomain and GATA factors); and binding to DNA may require homodimerisation or heterodimerisation (Myc/Mad/Max, Fos/Jun, and most nuclear receptor family members). All three structural features effectively increase the number of specific nucleotides required for efficient binding. However, most TFs, although with different affinity, bind a range of motifs rather than a single one. The extent of this binding site matrix differs considerably among TFs. Binding specificity may be strongly influenced by cofactors (Berthelsen et al. 1998; Knoepfler and Kamps 1995).

The binding of the TFs to the cis-regulatory sequences (transcription factor binding sites, TFBS) results in a variety of molecular interactions. A TF bound to DNA can interact with components of the basal transcriptional machinery. It may alter chromatin structure or stabilise the bending or looping of DNA. In addition, it may physically inhibit binding of different TFs to a nearby site. All these interactions, which lead to inhibition or initiation of transcription, might be direct or mediated by transcriptional cofactors that do not bind to DNA.

The central cis-regulatory module (CRM), required for the transcriptional initiation is the core (basal) promoter. This region is located approximately 50-100 bp up- or downstream of the transcriptional start site, serves as a “docking station” for the assembly of the transcriptional machinery and positions the start of transcription

relative to coding sequences (Butler and Kadonaga 2002). Although necessary for transcription, the core promoter is not a common point of regulation, and it cannot by itself generate functionally significant levels of mRNA (Kuras and Struhl 1999; Lee and Young 2000; Lemon and Tjian 2000). For precise temporal and spatial regulation of transcription and the amount of transcripts, additional CRMs (collection of diverse TFBS) are required. Some of these modules are located upstream, immediately after the core promoter to form the so-called proximal promoter region. This region is a few hundred bp long and contains several TFBS. Other CRMs like enhancers (Wray et al. 2003) regulate a discrete aspect of the activity of basal promoters in specific cell types and at particular time points. These elements are typically up to 300 bp long; contain a series of TFBS. Enhancers are traditionally defined by their ability to recapitulate an aspect of the endogenous gene activity when linked to a reporter gene in a position- and orientation independent manner (Arnone and Davidson 1997). They may reside far away from the proximal promoter region in both directions (usually several kilobases), but in some cases even to several megabases (Mb) away from the locus they regulate (Lettice et al. 2003). Some CRMs may function as 'silencers' that negatively modulate transcription activity (Ogbourne and Antalis 1998).

Boundary and insulator elements are another class of cis-regulatory elements. They are able to inhibit distant enhancer effects on core promoter regions or block the spread of the non-transcribed heterochromatin (Burgess-Beusse et al. 2002). Insulator function has also been attributed to some matrix attachment regions (MARs) (Laemmli et al. 1992). The complex expression patterns of the genes are regulated by a multiplicity of scattered CRMs, alternative core promoters and variety of interaction networks between transcription factors and cofactors.

Evolution of transcriptional regulation and morphological diversity

Species diverge from common ancestors through changes in their DNA. One of the questions in biology, then, is which are the genes that affect morphology and which changes in DNA are responsible for the evolution of morphological diversity? One of the most surprising biological discoveries of the past years is that most animals, no matter how different in appearance they are, share several families of genes that regulate major aspects of the body pattern. The discovery of this common genetic "toolkit" for animal development, containing many families of transcription factors and most signalling pathways, has provided the basis to study the genetic of animal diversity by enabling comparisons of how the number, regulation, or function of genes within the toolkit has changed in the course of animal evolution (Carroll et al. 2005; Davidson 2001). Comparisons of developmental gene regulation between morphologically divergent animals, analyses of intraspecific variation, and the response of organisms and genes to selection support the claim that regulatory DNA is the predominant source of the genetic diversity that underlies morphological variation and evolution (Belting et al. 1998; Carroll 2000; Carroll 2005; Gompel et al. 2005; Shapiro et al. 2004; Stern 1998; Sucena et al. 2003; Sucena and Stern 2000; Tautz 2000; Wang et al. 1999; Wittkopp 2006; Wittkopp et al. 2004; Wittkopp et al. 2002). There are many factors contributing to the importance of cis-regulatory DNA in evolution (Carroll 2000). 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. In contrast, the products of most of the genes involved in morphology patterning, as well as many TFs have pleiotropic function, e.g., they influence multiple phenotypic traits or regulate the expression of many different genes. Thus, 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 (as opposed to coding sequences) which allows greater varieties of mutations. Regulatory ele-

ments 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's are sufficiently relaxed such that the affinity and number of sites for each factor can evolve at a significant rate, even in functionally conserved elements.

Despite the predominant contribution of the regulatory DNA to the morphological evolution, the role of coding sequences should also be considered. There are several clear examples of functional sequence changes in proteins that affect form. For instance, several arthropod Hox proteins have changed in ways that are associated with shifts in form or developmental mechanisms (Damen and Tautz 1998; Stauber et al. 2002). In this case, selection against coding changes might have been relaxed because of functional redundancy among Hox paralogs. It also has been shown that morphological variation in dog breeds is associated with variation in the length of repeated amino acid sequences in the coding regions of a variety of developmentally important transcription factors (Fondon and Garner 2004) However, this variations may have led to deleterious, pleiotropic effects, which is manageable under domestication, but would limit its contribution to evolution under natural selection. Thus, it remains still unclear how often and to what extend do coding sequences of regulatory molecules functionally evolve and contribute to the morphological diversity.

Identification and characterisation of cis-regulatory elements

The main question of the post-genomic era is how to decipher the sequence information of the already sequenced genomes, which number is rapidly increasing in the last few years. At the moment several vertebrate genomes are available, including many mammalian and other tetrapod species, such as human, mouse, rat, chick, frog etc. as well as teleost genomes like green spotted pufferfish (*Tetraodon*

nigroviridis), fugu (*Takifugu rubripes*), zebrafish (*Danio rerio*), medaka (*Oryzias latipes*) and stickleback (*Gasterosteus aculeatus*). For the evolutionary relationship between vertebrates see Fig. 10. There are reasonably good tools to predict regions of the genome that correspond to protein coding genes, but it is less clear which parts of the genome are being transcribed. Also, very little is known about which regions are involved in the regulation of transcription or which regions fulfil structural functions in the chromosome. Cis-regulatory elements do not have stringent directional, positional and compositional constraints, such as those seen in the case of coding exons, which makes their automated detection by bioinformatics tools considerably more difficult than that of coding sequences or even mRNA splice sites. The simplest approach to predict cis-regulatory elements is to look for individual TFBS by using known motif models. This strategy is computationally straightforward to implement, but relies critically on the availability and quality of models of binding sites. The binding sites can be represented as consensus strings or as more informative position-specific score (weight) matrices, which include information about the frequencies of different nucleotides in different positions of the binding site (Stormo 2000). Information about consensus sequences and weight matrices of TFBS can be found in databases, which are either commercially available like TRANSFAC (Wingender et al. 1996) or provide open access JASPAR (Sandelin et al. 2004). Using known motif models to scan for putative TFBS works reasonably well for short individual regulatory regions, in which more or less is known, what to look for. In the absence of this information, such searchers for TFBS result in to many false positives, which preclude their application in genome wide analysis. In the last years many motif discovery algorithms have been developed, which do not rely on pre-existing models and can in principle find novel, previously unknown motifs (e. g. they can be used for *de novo* identification of TFBS without knowledge about the TFs). The most well-known of these algorithms have been compared by Tompa and co-workers. It has been shown that the predictive value of the motif search algorithms can be significantly increased by introducing sequence conservation (sequence similarity

between loci of different species) as additional criteria (Berman et al. 2004). Numerous algorithms that evaluate the conservation of TFBS are available. For example, TraFaC (Jegga et al. 2002) identifies conserved sites by scanning regions of conserved sequence similarity with a 200 bp window to detect co-occurrence of binding sites, whereas rVista (Loots and Ovcharenko 2004; Loots et al. 2002) and ConSite (Lenhard et al. 2003) score aligned binding sites in conserved regions. CONREAL (Berezikov et al. 2004) uses binding site predictions as anchors for sequence alignment, and performs better than other sequence alignment programs when aligning sequences from distant species.

As mentioned above the TFBS are usually organised in CRMs. Several recently developed tools, such as MSCAN (Alkema et al. 2004; Johansson et al. 2003 and EMCMODULE {Gupta, 2005 #562) are supposed to perform discovery of CRMs by detecting clusters of co-occurring TFBS. A major advance in identifying CRMs can be the usage of cross-species sequence comparison, which is the basement of a comparative genomics method, called phylogenetic footprinting, first introduced by Tagle and co-workers (Tagle et al. 1988), who investigated primate γ- and ε-globin genes. The basic assumption of phylogenetic footprinting is that regulatory elements in non-coding regions are under a higher selective pressure during evolution than non-functional regions. In other words, sequence comparison between species with enough evolutionary distance, like mammals and fish, for example, can reveal conserved sequence blocks with a potential function. The search for enhancers has been the main application of phylogenetic footprinting, because of the relatively well developed tools for their functional analysis, compared with other conserved non-coding sequences. A major problem regarding the use of phylogenetic footprinting for identification of CRMs is the insufficient knowledge available about such regulatory elements for their accurate prediction. The regulatory organisation of the genes shows huge variety and complexity and has only been deciphered systematically and mechanistically for a very small number of genes. As mentioned above cis-regulatory function may not be the only reason for the sequence conservation

of the non-coding DNA. Chromatin-structure defining elements like MARs etc. can be conserved as well. For instance, it has been reported that up to 11% of non-coding conserved sequences between mouse and human contain MARs (Glazko et al. 2003) and more than half of all predicted MAR sites occur within conserved regions. In addition, a role in the regulation of alternative splicing has been suggested for intronic sequences that are highly conserved between mouse and human and that are close to alternatively spliced exons (Sorek and Ast 2003). Another proportion of conserved non-coding sequences may consist of non-coding RNAs, which is speculated to be as numerous as protein-coding RNAs (Cawley et al. 2004). Conserved non-coding RNAs include antisense regulatory RNAs, microRNA sequences (Lim et al. 2003) that regulate gene expression. The cis-regulatory elements may coincide with such structures, which makes it even more difficult to predict them on base of sequence conservation analysis. Another caveat in the utilisation of the phylogenetic footprinting is the potential divergence in sequence and function of CRMs over large evolutionary distances (Dickmeis and Muller 2005; Dickmeis et al. 2004; Ludwig 2002; Ludwig et al. 2005; Ludwig et al. 1998).

Genome duplications in vertebrate evolution

It is widely accepted that gene duplication is a major source for the evolution of novel gene function resulting ultimately in the increase in organismal complexity and speciation (Mazet and Shimeld 2002; Meyer 2003; O'Brien et al. 1999; Taylor et al. 2001a). Three full genome duplication events have occurred during vertebrate evolution. The first two have been before the origin of jawed fishes and the split between the actinopterygians (ray-finned fishes) and sarcopterygians (land vertebrates), respectively 500 and 450 million years (mya) ago (Holland et al. 1994). The third happened early (~360 mya ago) in the evolution of actinopterygians (Fig. 10) The first evidence for the actinopterygian-specific full genome duplication has been provided by a studies describing that the ray-finned fishes have seven *hox*-gene clusters, in contrast to the land vertebrates, which have only four (Holland 1997;

Meyer and Malaga-Trillo 1999; Meyer and Schartl 1999; Postlethwait et al. 1998). The suggested genome duplication has recently been confirmed by the sequencing of the *Fugu* and *Tetraodon* genomes (Christoffels et al. 2004; Jaillon et al. 2004). This duplication event has led to a large number of duplicated copies of non-allelic genes found in different groups of teleosts (Amores et al. 1998; J. Wittbrodt 1998; Taylor et al. 2003; Taylor et al. 2001b) and is meant to be the main reason behind the high biodiversity of the teleost (half of the existing vertebrate species belong to the teleost fishes) (Volff 2005).

Retention of duplicated genes

All vertebrate animals, despite their generally diploid state, carry large numbers of duplicated genes, revealing that there is frequent evolutionary conservation of genes that arise through local, regional or global DNA duplication events. Classical models predict two potential fates for these duplicate gene pairs (Fisher 1935; Haldane 1933). The most likely fate is that one of the pair will degenerate to a pseudogene or be lost from the genome due to the vagaries of chromosomal remodelling, locus deletion or point mutation, in a process known as nonfunctionalisation. A less frequently expected model is that a population acquires a new, advantageous allele as the result of alterations in coding or regulatory sequences. Mutations that lead to such novel gene functions (a process called neofunctionalisation) are assumed to be extremely rare, so the classical model predicts that few duplicates should be retained in the genome over the long term and fails to explain the existence of the many duplicated genes found in vertebrate genomes.

It has been speculated that the mechanism by which duplicated genes are retained, is the evolution of new spacial or temporal expression domains through changes in their regulatory control elements (Cooke et al. 1997; Gompel et al. 2005; Jeong et al. 2006a; Marcellini and Simpson 2006; Prud'homme et al. 2006). An elaborate alternative model, called Duplication-degeneration-complementation (DDC) has been proposed by Force and co-workers to explain the retention of duplicated

paralogs during evolution (Force et al. 1999) (Fig. 3) Their model is based on the (often) multifunctional nature of genes, which is reflected by the multitude of regulatory elements specific to a particular expression domain. Mutations in subsets of regulatory elements in either one of the duplicated paralog may result in post-duplication spatial and temporal partitioning of expression patterns (subfunctionalisation) between them. As a result, both paralogs can fulfil only a subset of complementing functions of the ancestral gene, and will thus be retained by selection and not be lost secondarily (reviewed in (Prince and Pickett 2002).

The diversity of possible mechanisms of subfunctionalisation at the level of regulatory elements, however, are still poorly understood due to the lack of thorough

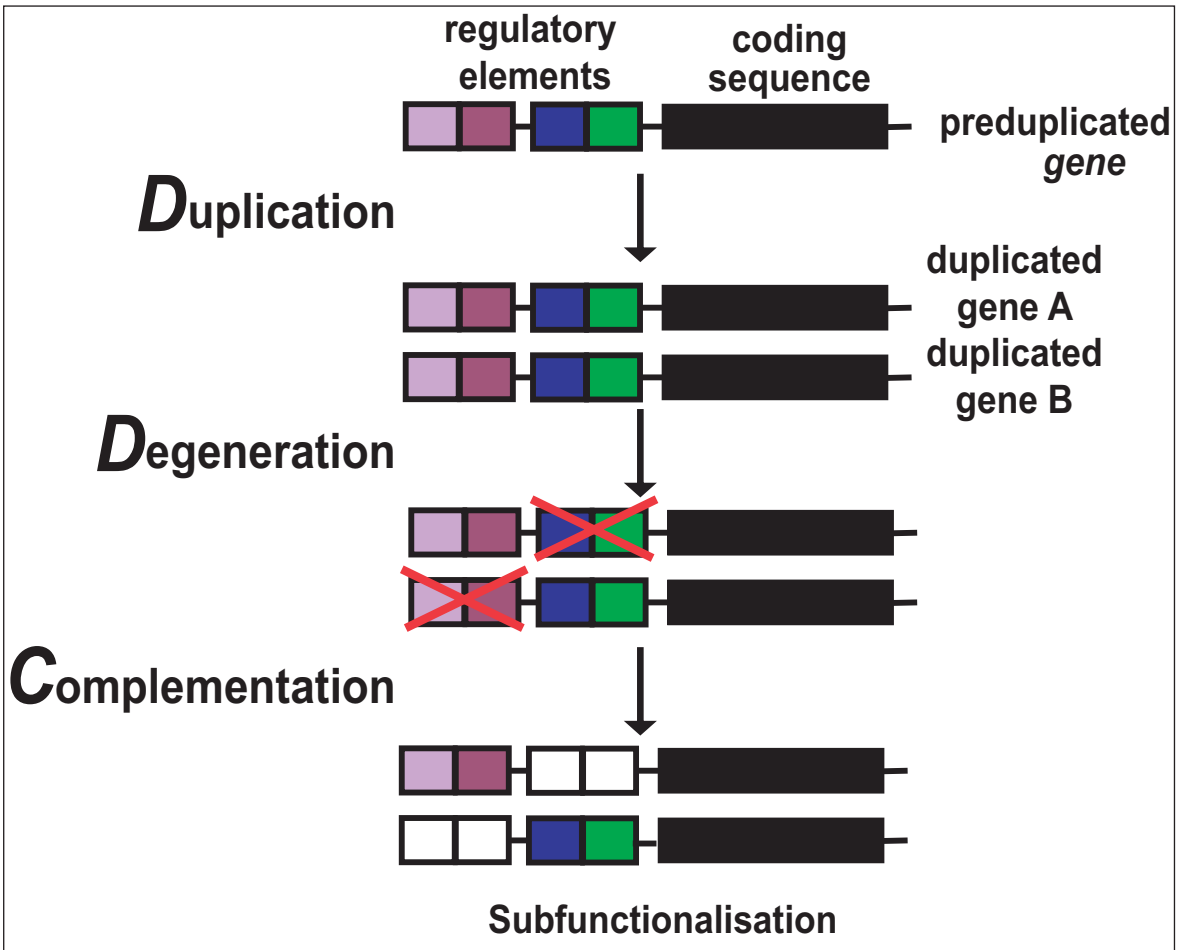

Figure 3. The duplication–degeneration–complementation model. The black boxes indicates coding sequence, the green, blue, purple and dark pink indicate cis-regulatory elements. This model relies on complementary degenerative changes in the cis-regulatory elements of the duplicated genes in such way that the duplicates together retain the original functions of their single ancestor. Modified from Force et al. 1999.

comparative molecular evolutionary studies on cis-acting elements (Ludwig 2002) backed by experimental verification of their function. Despite numerous presumed examples of subfunctionalisation of gene expression patterns between paralogs, there have only been two very recent reports that include the necessary experimental verification of the hypothesis of subfunctionalisation due to changes in cis-regulatory modules (Tumpel et al. 2006; Tvrdik and Capecchi 2006). Several studies, however, implicated specific mutations in enhancers of parologous gene copies to be the likely source of subfunctionalisation in duplicated *hox2b*, *hoxb3a* and *hob4a* enhancers in fish (Hadrys et al. 2004; Hadrys et al. 2006; Scemama et al. 2002).

The hedgehog gene family

Although they are absent from nematode worms, *hedgehog* (*hh*) family genes are widely distributed throughout the animal kingdom. They have been first discovered in *Drosophila melanogaster*, (Nusslein-Volhard and Wieschaus 1980) in which mutation of the single *hedgehog* (*hh*) gene that is present in this species gives rise to an embryo that is covered in cuticular processes called denticles, thus giving the name "*hedgehog*". In vertebrates, genome duplication has given rise to multiple *hh* genes. Comparative studies on the evolution of the vertebrate *hh* gene family (Zardoya et al. 1996a; Zardoya et al. 1996b) showed that two rounds of duplication led to the evolution of three copies from a single ancestral hedgehog gene: *sonic hedgehog (shh*), *indian hedgehog* (*ihh*) and *desert hedgehog* (*dhh*). In addition the extra genome duplication in the ray-finned fish lineages resulted in the *shh* paralog *tiggy-winkle hedgehog* (*twhh*) genes (Zardoya et al. 1996a; Zardoya et al. 1996b) as well as duplication of *ihh* (Avaron et al. 2006) and probably also *dhh* genes (There isn't any evidence for *dhh* duplicates in teleost so far.). For more details about the phylogenetic relationship of vertebrate *hedgehog* genes see Fig. 4.

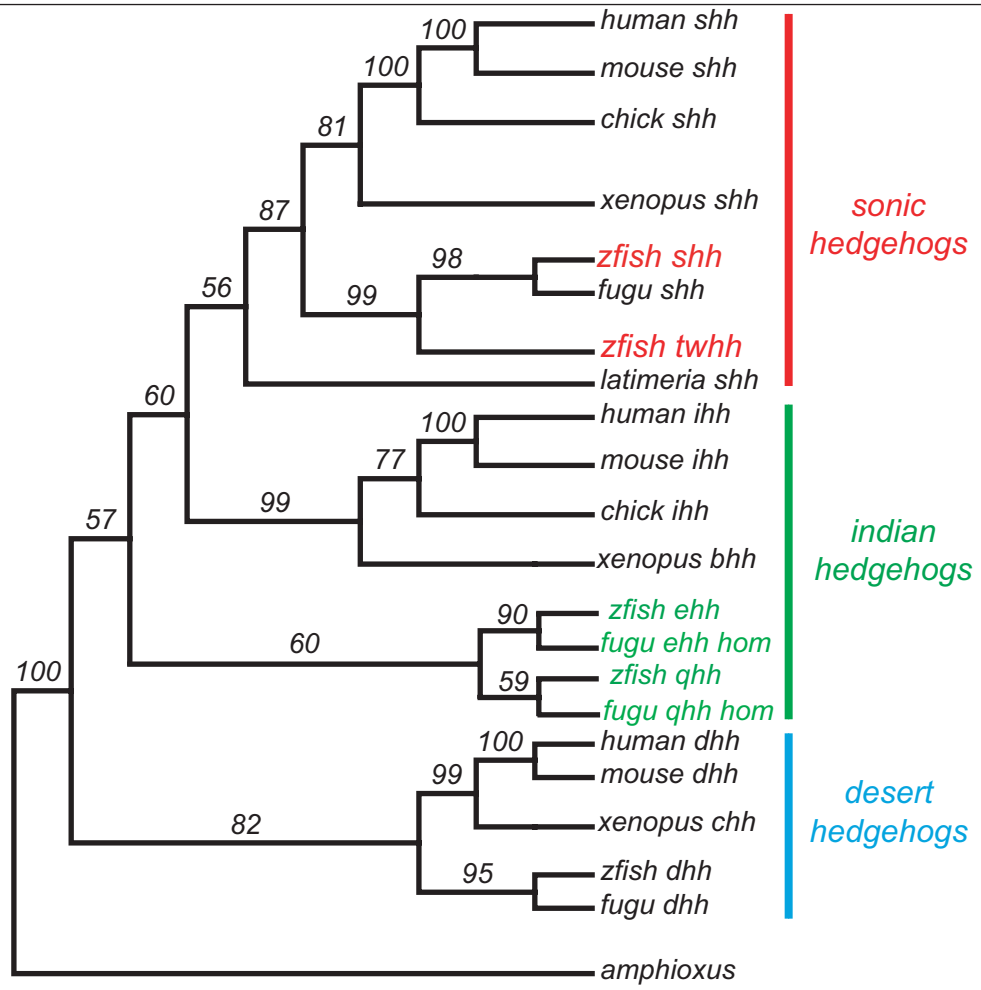

Figure 4. Phylogenetic tree of the *hedgehog* gene family in vertebrates. Due to the extra genome duplication the ray-finned fishes often have more paralogous genes than land vertebrates and this is also true for the *hedgehog* gene family. The duplicated *hedgehog* genes in teleosts are highlighted in colour. On the top of each branch are shown the bootstrap values, which indicate the possibility in percentage the respective genes to be joined into a group.

Mechanism of action and signal transduction pathway of the hedgehog proteins

Hedgehog proteins are secreted signalling molecules, which act as a morphogens to regulate a variety of developmental processes. Morphogens are signalling molecules, produced by a localised source of cells and then establish a gradient of concentration over a field of cells. They act directly on the cells at a distance without any relay mechanisms. (Entchev and Gonzalez-Gaitan 2002). The morphogen-concept provides an attractive explanation of how cells in the embryo acquire positional information and how a relatively low number of signalling molecules can

coordinate the multitude of cellular differences.

The core components of the hedgehog-signalling pathway are highly conserved through evolution and were mainly discovered by genetic analyses of *Drosophila melanogaster*. An overview of the Hedgehog pathway, which has been extensively reviewed (Hooper and Scott 2005; Ingham and McMahon 2001; Ingham and Placzek 2006; Lum and Beachy 2004), is presented on Fig. 5. One key feature of the HH proteins is their lipid (cholesterol) modification, which has an important effect on the signalling activity by facilitating the movement of hedgehog molecules between cells (Gallet et al. 2006; Lewis et al. 2001; Li et al. 2006). The Hh molecules require the activity of a highly conserved transporter-like protein, named Dispatched, for their efficient secretion from cells (Burke et al. 1999; Ma et al. 2002; Nakano et al. 2004). The intercellular transport of the secreted HH protein is modulated by interaction with proteoglycans (Han et al. 2004).

Recent studies have revealed that HH proteins bind directly to two related and conserved transmembrane proteins, known as Interference hedgehog (IHOG) and Brother of IHOG (BOI) in *D. melanogaster* and CAM-related 1 down-regulated by oncogenes (CDO) and brother of CDO (BOC) in vertebrates (Tenzen et al. 2006; Yao et al. 2006; Zhang et al. 2006) This binding facilitates the interaction between hedgehog proteins and the transmembrane protein Patched (Ptc in flies and fish, PTCH in mammals), an interaction that leads to the activation of another transmembrane protein Smoothened (SMO), the universal transducer of hedgehog activity (Alcedo et al. 1996; Stone et al. 1996; van den Heuvel and Ingham 1996). SMO activation initiates a signalling cascade that results in transcription of hedgehog target genes. Whereas in *D. melanogaster*, this is mediated by a single member of the Glioma-associated oncogene homologue (GLI) family of transcription factors (Alexandre et al. 1996). In vertebrates, three distinct GLI proteins are involved in the transcriptional response to HH proteins (Bai et al. 2004). Two of these, GLI2 and GLI3, are structurally similar to the D. melanogaster Cubitus interruptus (CI) protein, possessing both repressor and activator domains that flank a DNA-binding domain. In the

absence of HH signalling, GLI3 and GLI2, like CI, undergoes proteolytic cleavage to yield a truncated protein that functions as a repressor (CIR) of HH target gene transcription. (Wang et al. 2000). Activation of the pathway results in the suppression of cleavage flowed by nuclear import of full-length GLI2 and GLI3 proteins. These activate the expression of target genes. By contrast, the GLI1 protein lacks the repressor domain and seems to enhance the activating function of GLI2. Unlike the other two *gli* genes, *gli1* is itself a target of SHH signalling as a part of a positive feedback loop (Bai et al. 2002).

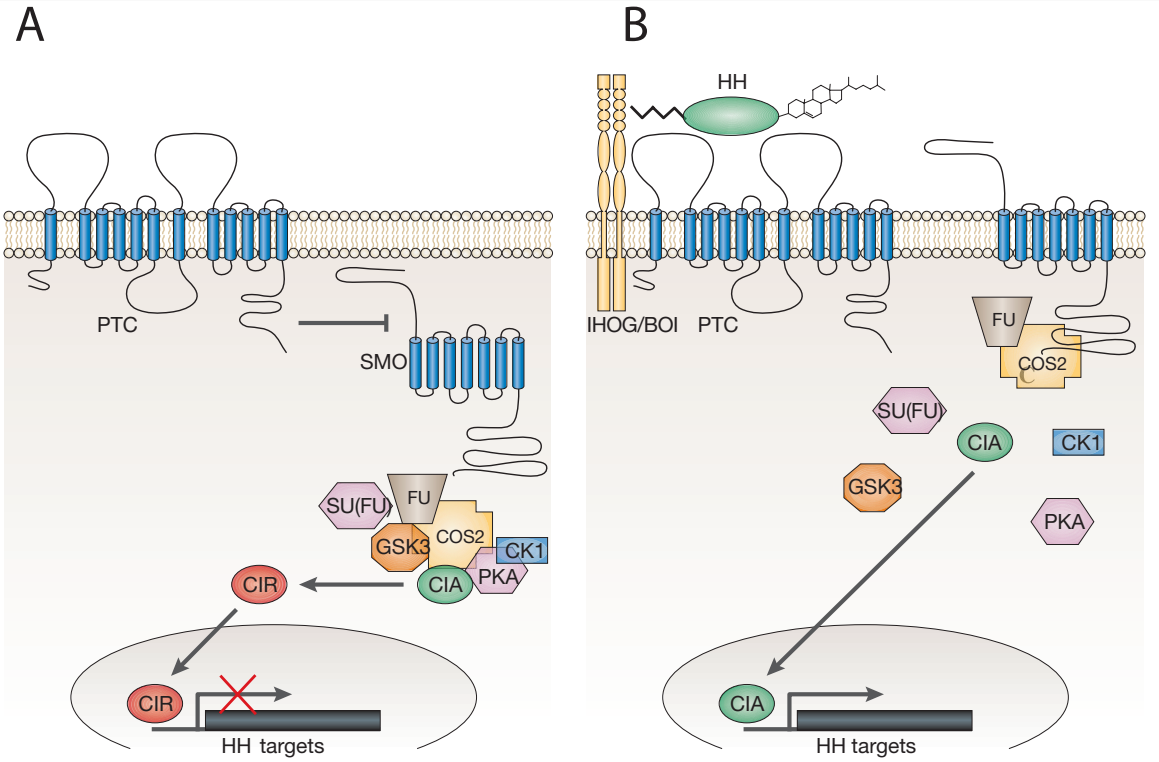

Figure 5. The hedgehog signalling pathway in *Drosophila*. A, In absecence of HH signals PTC represses the activity of SMO and the pathway is inactive. B, The binding of HH molecule to PTC relieves the repressive effect of PTC on the SMO. Once SMO becomes activated the cleavage of CI to a repressor form (CIR) is blocked. A full length activator protein (CIA) is released form a protein complex, to which is bound in absence of HH signals, becomes transported into the nucleus, where activates the transcription of target genes.. Modified from Ingham et al. 2006.

Expression pattern and biological role of the sonic hedgehog genes

In zebrafish as in higher vertebrates, expression of *shh* is highly restricted to regions with organiser activity (Krauss et al. 1993; Scholpp et al. 2006; Strahle et al. 1996). In the zebrafish embryo, *shh* is expressed initially in the embryonic shield, and subsequently in the notochord, prechordal plate and the floor plate (Fig. 6). In the brain, *shh* expression is detected in the ventral midbrain, the hypothalamus, the *zona limitans intrathalamica* (*zli*) and in a small patch of cells in the telencephalon (Krauss et al. 1993). In the two-day old embryo, strong expression of *shh* is also found in the endoderm and its derivatives (Strahle et al. 1996). The *shh* paralog, *twhh,* is also expressed in the midline and parts of the ventral brain of the zebrafish embryo (Fig. 6). There are however, distinct differences, regard to timing and tissue restriction of expression between the two paralogous genes (Ekker et al. 1995), which may have important implication in their cooperative function. First expression of *twhh* is first detected at about 50 % epiboly and its expression during gastrulation in the dorsal mesoderm; this appears to precede the expression of *shh,* which is first detected at about 60 % epiboly. In addition, Etheridge et al, have shown that *shh* is expressed in notochord precursors and *twhh* is exclusively expressed in the overlying floor plate cells during gastrulation (Etheridge et al. 2001). Later (24-36 hpf), *shh* is expressed both in the notochord and floor plate, while *twhh* remains restricted to the floor plate (Fig. 6). Another notable difference at later these stages is differential rostro-caudal restriction within the diencephalon and midbrain such that the later domain of *twhh* expression appears to constitute a subset of the *shh* domain (Ekker et al. 1995). The protein activity of *twhh* is very similar to that of *shh* (Lauderdale et al. 1998). It is likely, that the concerted action of *shh* and *twhh* is regulated quantitatively by their partially overlapping and tightly controlled level of expression.

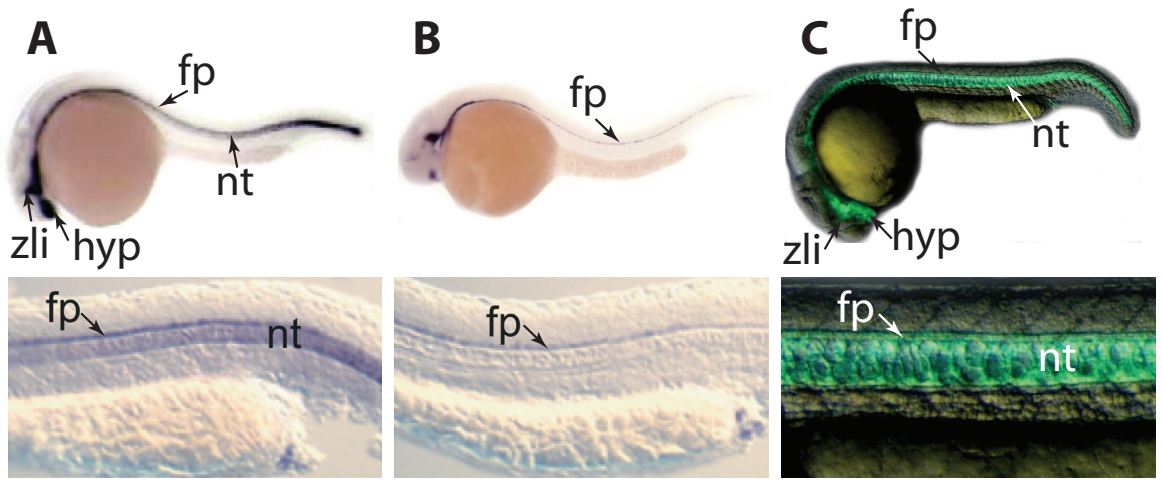

Figure 6. Expression of *shh* and *twhh* genes at 24 hpf in zebrafish. A-B Expression of *shh* (A) and *twhh* (B) mRNAs, detected by *in situ* hybridisation. On the top panel a whole embryo is shown, on the bottom high magnification of the trunk on the level of the yolk extension. C, Transgenic zebrafish embryo carrying *gfp-reporter* construct containing the zebrafish *shh* promoter region and four of the gene's enhancers (*ar-A, ar-B, ar-C and ar-D*; for details see "*Cis-regulatory modules of the vertebrate sonic hedgehog gene*" and Fig. 7). As can be seen on the pictures the expression of the transgene fully recapitulates the expression pattern of the endogenous *shh* gene. Abbreviations: fp, floor plate; nt, notochord; hyp, hypothalamus; z*li zona limitans intrathalamica*.

The biological role of Shh has been extensively studied. It controls a multitude of different differentiation processes during vertebrate embryogenesis. In the neural tube, it acts as a morphogen that drives the differentiation of specific neurons in a concentration-dependent manner and is important for neurogenesis (Ingham and McMahon 2001; Ingham and Placzek 2006; Ruiz i Altaba 1994). Other functions include patterning of the endoderm, the somites and the paired appendages and Shh was also implicated in hair and tooth development (Ingham and McMahon 2001; Ingham and Placzek 2006; Jessell 2000) Moreover, misregulation of the Shh signalling pathway can cause a variety of tumours in humans including basal cell carcinoma and medulloblastoma (Bale and Yu 2001; Ingham 1998; Marino 2005; Ruiz i Altaba 1994). So far, only *shh* function was studied in genetic mutants (Schauerte et al. 1998). Nevertheless, morpholino knock down and gene expression analyses identified several functions of the *twhh* gene. *Twhh* was shown to cooperate with *shh* in the midline to specify branchiomotor neurons, but is also required in the *zli* and was

implicated in eye morphogenesis (Bingham et al. 2001; Chandrasekhar et al. 1998; Nasevicius and Ekker 2000; Scholpp et al. 2006; Yamamoto et al. 2004).

Cis-regulatory modules of the vertebrate sonic hedgehog gene

The genomic locus of the zebrafish *sonic hedgehog* gene is well characterised and a substantial amount of data on the functionality of its cis-acting elements exist (Ertzer et al. 2007; Muller et al. 2000; Muller et al. 1999). Enhancers that drive expression in the ventral neural tube and notochord of the developing embryo reside in the two introns and upstream sequences (Fig 7). *Activating regions A* and *B* (*ar-A*, *ar-B*) reside in the first intron of the gene and drive expression in the notochord and FP respectively. The *ar-A* region is conserved trough most vertebrate *shh* genes, but so far no function has been applied to it in mouse. Corresponding conserved region to *ar-B* has not been found in tetrapods, but is well conserved between fish species. The activation region *ar-D* is located upstream (appr. 1 kb) from the transcriptional start site of the gene and corresponds to the mouse FP enhancer *SFPE2* (Epstein et al. 1999; Ertzer et al. 2007; Jeong and Epstein 2003). The activity of this enhancer in zebrafish is restricted to the anterior FP, in contrast to the mouse, where it drives FP expression throughout its full length (Fig 7). A fourth region, *ar-C*, in the second intron, directed expression in the notochord and weakly in the posterior FP. (Ertzer et al. 2007; Muller et al. 1999). This zebrafish enhancer also functions in the midline of mouse embryos (Muller et al. 1999) suggesting, that the cis-regulatory mechanisms involved in regulating shh expression are at least in part conserved between zebrafish and mouse. However, the mouse enhancer, *SFPE2*, which shows sequence similarity with *ar-C* of zebrafish, is FP-specific (Epstein et al. 1999; Jeong and Epstein 2003) (Fig. 7 and Fig. 8). This difference of enhancer activity emphasises the importance of addressing the mechanisms of divergence in enhancer function between distantly related vertebrates. Given the observations on the *ar-C* enhancer in fish and mouse, we postulated, that this enhancer may have been a target of enhancer divergence between *twhh* and *shh* paralogs in zebrafish during evolu-

tion. Recent studies by Ertzer et al. (Ertzer et al. 2007) have shown, that precipice tuning of spatio-temporal expression of shh in the zebrafish midline and ventral forebrain is controlled by synergistic cooperation between the different enhancers of the gene. For instance, the *ar-B* can synergistically act with *ar-C* to initiate FP expression in early stages (3-somites). Combination of *ar-A* and *ar-C* is required for enhancement of the notochord expression in early and later stages. This synergistic action of both enhancers is also needed to maintain the expression in hypothalamus at 24 hpf, however maintaining the expression at later stages (32 hpf) requires additional activity of the *ar-B*. Additional studies have contributed to identification of long distance *shh* enhancers in mouse. For instance, Lettice and co-workers (Lettice et al. 2003) have described an enhancer important for *shh* expression in the zone of polarising activity (ZPA) of the developing limb bunds. Using BAC reporter assays, Jeong et al. (Jeong et al. 2006b) have uncovered six enhancers distributed over 400 kb along the mouse *shh* locus, directing expression in different domains of the ventral forebrain and spinal cord as well as in the notochord (Fig. 7).

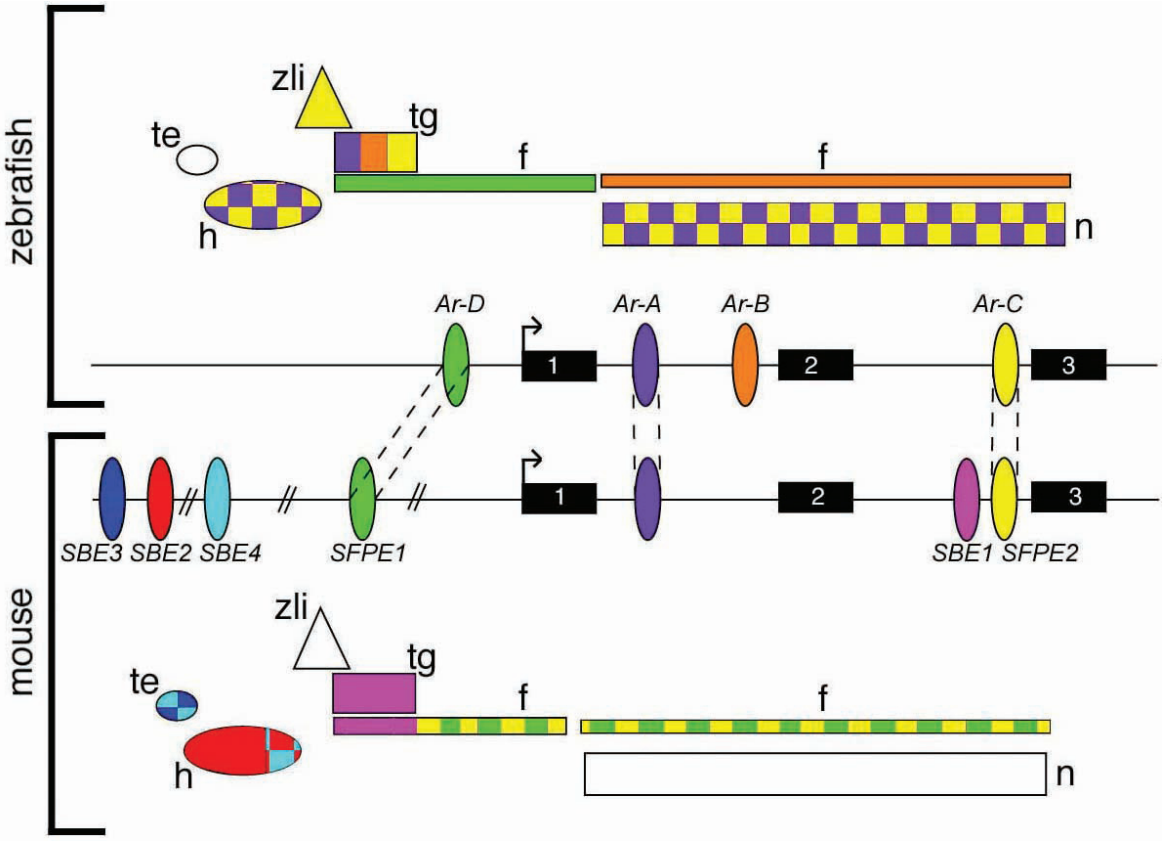

Figure 7. Comparison of the *shh* enhancer regions between zebrafish and mouse. Each enhancer is labelled whit different colour and the same colour is used to mark the region where the respective enhancer is active on the schematic representation of the zebrafish (top) and mouse (bottom) midline and central nervous system. The black boxes represent the exons. Abbreviations: *SBE*, sonic brain enhancer; *SFPE,* sonic floor plate enhancer; f, floor plate; h hypothalamus; n, notochord; tg, tegmentum, *zli, zona limitans intrathalamica*; te, telencephalon. Modified form Ertzer et al. 2007

The ar-C/SFPE2 cis-regulatory paradigm

As mentioned above, despite the high degree of sequence similarity (Fig. 12) there is a clear functional divergence between zebrafish *ar-C* and the corresponding mouse enhancer *SFPE2*. In zebrafish *ar-C* is manly active in the notochord and only weekly in the floor plate, however the corresponding mouse enhancer, *SFPE2* is a floor plate enhancer and only shows notochord activity in a multimerised and truncated form (Jeong and Epstein 2003) (Fig.7, Fig.8). This divergence can be explained with the presence and complementing function of a *shh* duplicate paralog gene, *twhh* in zebrafish while in the mouse only one *shh* copy is present. The two paralogous genes in zebrafish have different expression patterns. As described above *shh* is expressed

in both, the floor plate and the notochord, while *twhh* is expressed only in the floor plate. These diverged expression patterns suggest enhancer divergence resulting in diverged tissue specificities after the gene duplication, and the predicted enhancer divergence may have led to possible subfunctionalisation between the two paralogs. (Fig. 8). If the subfunctionalisation model is valid, the functional divergence between *ar-C* and *SFPE2* implies that the *twhh* paralog may also have a functional *a-C* enhancer which is predicted to carry floor plate specific activity. Additional support to this hypothesis is provided by the fact, that although weekly detectable, *ar-C* is the only one *shh* enhancer which shows some sequence similarity to the positionally conserved region in the second intron of the *twhh* gene.(Fig. 15).

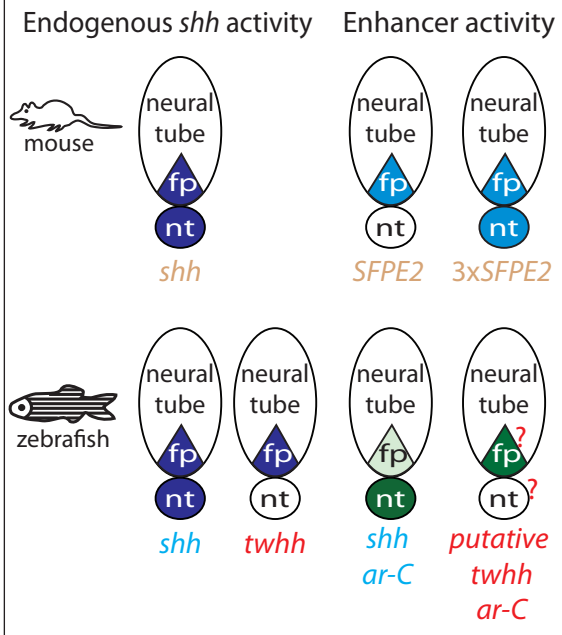

Figure 8. Functional divergence of *ar-C/SFPE2* between zebrafish and mouse. The expression pattern of the endogenous *shh* gene (s) (left panel) and the *ar-C/SFPE2* (right) is schematically represented on the cross-section of the neural tube, where coloured triangle indicates floor plate expression and coloured circle below the neural tube a notochord expression. The question marks indicate that this is expected activity of a putative *twhh* ar-C, objective of verification in this study. nt, notochord, fp, floor plate.

Aims and objectives

A vexing problem in understanding the evolution of cis-regulatory modules is that only relatively small number of enhancers and other CRMs have been characterised in sufficient detail to develop generally applicable rules about their conserved structures and evolutionarily permitted modifications. Therefore, it is paramount also to functionally investigate the molecular mechanisms that underlie the divergence of CRMs. Duplicated genes can provide suitable model for such studies. The advantage of using duplicated gene is that they allow comparison of cis-regulatory elements, which have evolved parallel for the same amount of time in the same environment. Moreover, the paralogous elements can be compared to non- or pre-duplicated (ancestral) orthologous gene elements, which had similar time to evolve but in different environment. Such comparisons may allow for the prediction of evolutionary events shaping CRMs. The validation of such predictions can then be carried out by molecular and functional cross-species analyses of conserved CRMs (enhancers) from paralogous and orthologous genes. The cross-species analysis can be extended by a mutational analysis of the TFBS in reconstruction of the ancestral archetype of the enhancer, which can provide insights into the mechanisms involved in enhancer divergence after gene duplication and overall mechanism of cis-regulatory element's evolution.

To find out more about these mechanisms we have chosen the vertebrate sonic hedgehog genes as model system. There are several reasons for this choice: First, the regulatory elements of the *shh* genes have been well studied in both zebrafish and mouse; Second, due to the extra genome duplication in the actinopterygian lineage two paralogous genes (*shh* and *twhh*) with partially non-overlapping expression pattern exist in zebrafish, thus supporting the hypothesis for subfunctionalisation of these gene after the gene duplication. To prove this hypothesis we focused our study on the ar-C enhancer located in the second intron of the gene, because of its functional divergence between mouse and zebrafish and the indication of pos-

sible existence of putative paralogs to the *ar-C* enhancer in the twhh gene (see *"The ar-C/SFPE2 cis-regulatory paradigm"* and Fig. 8) Another question we aim to answer in this study is the mechanism for the subfunctionalisation of the *ar-C* enhancer. For instance, models can be drawn, that the subfunctionalisation occurred by changes of binding sites within the *ar-C* enhancer modifying its function or, alternatively, multiple enhancers are involved, e.g the subfunctionalisation happened on the level of modulating whole enhancers.

To answer the above listed questions our aims were:

- To isolate *ar-C* enhancers (intron 2 sequences) form *shh* and *twhh* genes from vertebrate species on important time points of evolution. (Fig. 10).
- To analyse the activity of the isolated enhancers by transgenic reporter assays in zebrafish embryos.
- To identify potential binding sites by sequence comparison (phylogenetic footprinting), and the changes occurred in them, leading to subfunctionalisation.
- To analyse the identified putative binding sites by mutations and functional test in transient transgenic zebrafish embryos to be able to reconstruct the evolutionary events responsible for the enhancer divergence after gene duplication.
- To generate a molecular model for enhancer divergence in duplicated *ar-C* enhancers.

Materials and methods

Chemicals, enzymes and kits

Acetic acid	Merck, Darmstadt
Agarose	Sigma, Taufkirchen
Ammonium acetate	Merck, Darmstadt
Ampicillin	Roche, Mannheim
Bacto-Agar	Roth, Karlsruhe
Bacto-Trypton	Roth, Karlsruhe
Bacto-Yeast extract	Roth, Karlsruhe
Boric acid	Roth, Karlsruhe
BSA	Serva, Heidelberg
Calcium chloride	Merck, Darmstadt
Calf intestine alkaline phosphatase	Promega, Mannheim
Disodium hydrogen phosphate	Roth, Karlsruhe
DNA-Ladder (1 kb)	New England Biolabs, Frankfurt a.M.
DNA-Ladder (100 bp)	New England Biolabs, Frankfurt a.M.
DNA-Ladder (Mix)	Peqlab, Erlangen
DNeasy Blood & Tissue Kit	Qiagen, Hilden

dNTPs	Promega, Mannheim
EDTA	Roth. Karlsruhe
Ethanol	Roth, Karlsruhe
Ethidium bromide	Roth, Karlsruhe
Gentamicin	Sigma, Taufkirchen
Isoamyl alcohol	Roth, Karlsruhe
Isopropanol	Merck, Darmstadt
Magnesium sulphate	Merck, Darmstadt
Methanol	Roth, Karlsruhe
Nuclease free water	Ambion, Huntigdon, UK
Ocean Sea Salt	Kölle-Zoo, Karlsruhe
Oligonucleotides	Metabion, Planegg
Pancreatic ribonuclease A	Sigma, Taufkirchen
PBS	Invitrogen, Karlsruhe
Phenol	Roth, Karlsruhe
Phenol red	Roth, Karlsruhe
Potassium acetate	Roth, Karlsruhe
Proteinase K	Sigma, Taufkirchen

PureYield Plasmid Midiprep System	Promega, Mannheim
QIAGEN Plasmid Maxi Kit	Qiagen, Hilden
QIAquick Gel Extraction Kit	Qiagen, Hilden
QIAquick PCR Purification Kit	Qiagen, Hilden
QuickLyse Miniprep Kit	Qiagen, Hilden
Restriction endonucleases	Promega, Mannheim, Invitrogen, Karlsruhe or New England Biolabs, Frankfurt a.M.
SDS	Roth, Karlsruhe
Sodium acetate	Roth, Karlsruhe
Sodium chloride	Roth, Karlsruhe
Sodium dihydrogen phosphate	Roth, Karlsruhe
Sodium hydrogen carbonate	Roth, Karlsruhe
Sodium hydroxide	Sigma, Taufkirchen
T4 DNA ligase	Promega, Mannheim
T4 DNA polymerase	Promega, Mannheim
GoTaq DNA polymerase	Promega, Mannheim
TOPO TA Cloning Kit	Invitrogen, Karlsruhe
Triple Master PCR System	Eppendorf, Hamburg
Tris-Base	Roth, Karlsruhe

Tris-HCl	Roth, Karlsruhe
Wizard Plus SV Minipreps DNA Purification System	Promega, Mannheim
Wizard SV Gel and PCR Clean-Up System	Promega, Mannheim

Equipment and materials

Bacteria incubators	Heraeus, Hanau
Borosilicate glass capillaries	Harvard Ltd., Kent, UK
Cool centrifuge J2-HS	Beckman, Stuttgart
Digital camera	DFC300 FX, Leica, Bensheim
Electrophorese chambers	Peqlab, Erlangen
Eppendorf microcentrifuge tubes	Eppendorf, Hamburg
Falkon tubes	Greiner, Nürtingen
FemtoJet microinjector	Eppendorf, Hamburg
Flaming-Brown Needle puller	Sutter Instruments, USA
Fluorescent stereomicroscope MZ FLIII	Leica, Bensheim
Gas microinjector	Tritech research inc., L.A., USA
Incubator for fish embryos	Heraeus, Hanau
Magnetic thermomixer	Heidolph, Rosenfeld
Microcentrifuge 5417 R and C	Eppendorf, Hamburg
Microcentrifuge Biofuge pico	Heraeus, Hanau
Microfiltration columns	Pall, Ann Arbor, USA

NanoDrop ND-1000	Peqlab, Erlangen
PCR-Thermocycler, MJ Research	Biozym, Oldendorf
Petri dishes	Greiner, Nürtingen
Pipette tips	Corning, Corning
Spectrophotometer	Eppendorf, Hamburg
Stereomicroscope SMZ645	Nikon, Düsseldorf
Sterile filters	Renner, Darmstadt
Thermomixer	Eppendorf, Hamburg
UV Transilluminator	Saur, Reutlingen
Vac-Man Vacuum manifold	Promega, Mannheim
Vortex	Bender & Hohbein, Karlsruhe
Water bath	Kötterman, Uetze-Hänigsen

Solutions and Buffers

If not specified otherwise, all solutions and buffers have been prepared in distilled water.

TAE Buffer

40 mM Tris-Base, 1 mM EDTA, 5 mM Acetic acid; pH=7,8

TBE-Buffer

90 mM Tris-Base, 1 mM EDTA, 44 mM Boric acid; pH=8,0

TE-Buffer

10 mM Tris-HCl (pH = 7,4), 1mM EDTA; (pH = 8)

Pancreatic ribonuclease A (RNase A) stock solution

20 mg/ml RNase A in 1 mM sodium acetate; pH=4,5

Proteinase K stock solution

10mg/ml in PBS

Lysis buffer for genomic DNA extraction

10 mM Tris-HCl, 0,1 M EDTA, 0,5% SDS; pH=8,0

LB-Agar

1,5% Bacto-Agar in LB-Medium

LB-Medium

1% Bacto-Trypton, 0,5% Yeast extract, 1% NaCl; pH=7,0

SOC–Medium

2% Bactotrypton, 0,5% Yeast extract, 10 mM NaCl, 25 mM KCl

Hank´s solution

0,14 M NaCl, 5,4 mM KCl, 0,25 mM Na_2HPO_4, 0,44 mM KH_2PO_4, 1,3 mM $CaCl_2$, 1 mM $MgSO_4$, 4,2 mM $NaHCO_3$.

System water in the fish facility

120 mg/l „Ocean Sea Salt“, 45 mg/l $NaHCO_3$ in desalted water

Phenol red solution (10x)

10% Phenolred, 0,2 M KCl; pH=7,5

Methylene blue solution (2000x)

0,1% methylene blue in distilled water

Isolation of plasmid DNA

Plasmid DNA was isolated using Quiagen or Promega kits for Mini, Midi and Maxi Plasmid preparations, following the manufacturer's instructions. For all types of plasmid preparation, an appropriate volume of over night bacterial culture in LB-medium was used. The kits from both manufacturers are based on the alkali-lysis method by which the plasmid DNA is separated from the genomic DNA and most of the proteins (they form white precipitate, which is removed by centrifugation or filtration). The remaining solution, containing the plasmid DNA, is subsequently purified on anion exchange or silica membrane column to ensure complete removal of the rest of the proteins, RNA and bacterial endotoxins. In the end, the DNA is resuspended or eluted in an appropriate volume of nuclease free water. The concentration was determined on a NanoDrop spectrophotometer device.

Isolation of genomic DNA

Genomic DNA was isolated according the standard protocol for *"Isolation of High-molecular-weight DNA from Mammalian Cells Using Proteinase K and Phenol"*, described by (Sambrook and Russell 2001). In brief, the tissue samples (~500 mg) are frozen in liquid nitrogen and subsequently pulverised. The tissue powder is spread slowly (to avoid building of clumps) over the surface of a 10 ml lysis buffer in a small beaker (25 ml). Pancreatic ribonuclease A in concentration 20 μg/ml is added to the lysis buffer before the spreading of the tissue powder. The lysate is incubated for 30 min. at 37 °C to ensure the degradation of the RNA. After the addition of proteinase K in concentration 100 μg/ml, the lysate is incubated further for 3 hours at 55 °C. To remove the peptides and remaining proteins, the cell lysate is extracted two times with equal volume of phenol and once with phenol/chloroform. The DNA is recovered from the remaining solution by precipitation with 0,2 volumes of 10 M ammonium acetate and 2 volumes of ethanol. White, clearly visible precipitate form genomic DNA is formed, which can be taken out with U-shaped Pasteur-pipette. The DNA is washed in 70 % ethanol and after drying, it is dissolved in an appropriate volume of TE buffer.

DNeasy Blood & Tissue Kit (Quiagen) was used for isolation of genomic DNA from small amount of tissue, according to the manufacturer's instructions.

Polymerase Chain reaction (PCR)

The amplification of DNA fragments from genomic DNA or plasmids was performed by PCR (Mullis and Faloona 1987; Rabinow 1996; Sambrook and Russell 2001). Two enzyme systems were used depending on the purpose. Ordinary Taq polymerase (GoTaq, Promega) was used when proof-reading activity was not needed (for example colony tests). Triple Master PCR System (Eppendorf) was utilised for the amplification of DNA fragments for cloning purposes. This is an enzyme mixture (Taq polymerase with proof-reading polymerases) which is optimised for amplification of long targets with relatively high speed and proof-reading activity.

The PCR was performed according to the user manuals provided with the enzymes, with adjustment of the annealing temperature and elongation time according to the used primers and the size of the amplified fragments. All PCRs were performed on MJ Research thermocycler.

A degenerate PCR approach was used to isolate shh and *twhh* intronic sequences form species for which sequence information about those gene loci was not available (tench and *Latimeria*). The degenerate PCR is identical in most respects to ordinary PCR, but with one major difference. Instead of specific PCR primers with a given sequence, mixed PCR primers are used. In other words, if the sequence of the fragment to be amplified is not exactly known, "wobbles"are inserted in the PCR primers where there is more than one possibility. For instance, if just a protein motif is available, it can be back-translated to the corresponding nucleotide motif. However, the genetic code is degenerate (in most cases given amino-acid is encoded by more than one codon), which results in more than one possibility for the identity of some nucleotides the motif. In our case, degenerative primers were designed on the base of well-conserved protein motifs in the second and third exons (flanking the second intron) of *shh* and *twhh* genes from different vertebrate species.

Site-directed mutagenesis by PCR

A PCR based approach ("Higuchi Method", described in (Higuchi et al. 1988) was utilised to generated mutation and insertion in zebrafish *shh* and *twhh ar-C* enhancers. This method allows mutation, deletion and insertion of sequences at any position in the DNA fragment. The method is based on two PCR rounds (see Fig. 9). In the first round, two primary PCRs produce two overlapping DNA fragments, both bearing the same mutation introduced via primer mismatch in the region of overlap. In the second round, the products of the first two reactions are mixed (after gel purification) and used as a template in a second PCR. The overlap in sequence allows the two fragments to anneal after their denaturation and renaturation and produce a structure with recessed 3' OH ends that can be extended by a DNA polymerase to

produce a complete duplex fragment. These extended segments can then serve as a template for the secondary reamplification of the combined sequences using only the outermost two of the four primers used to produce the primary fragments.

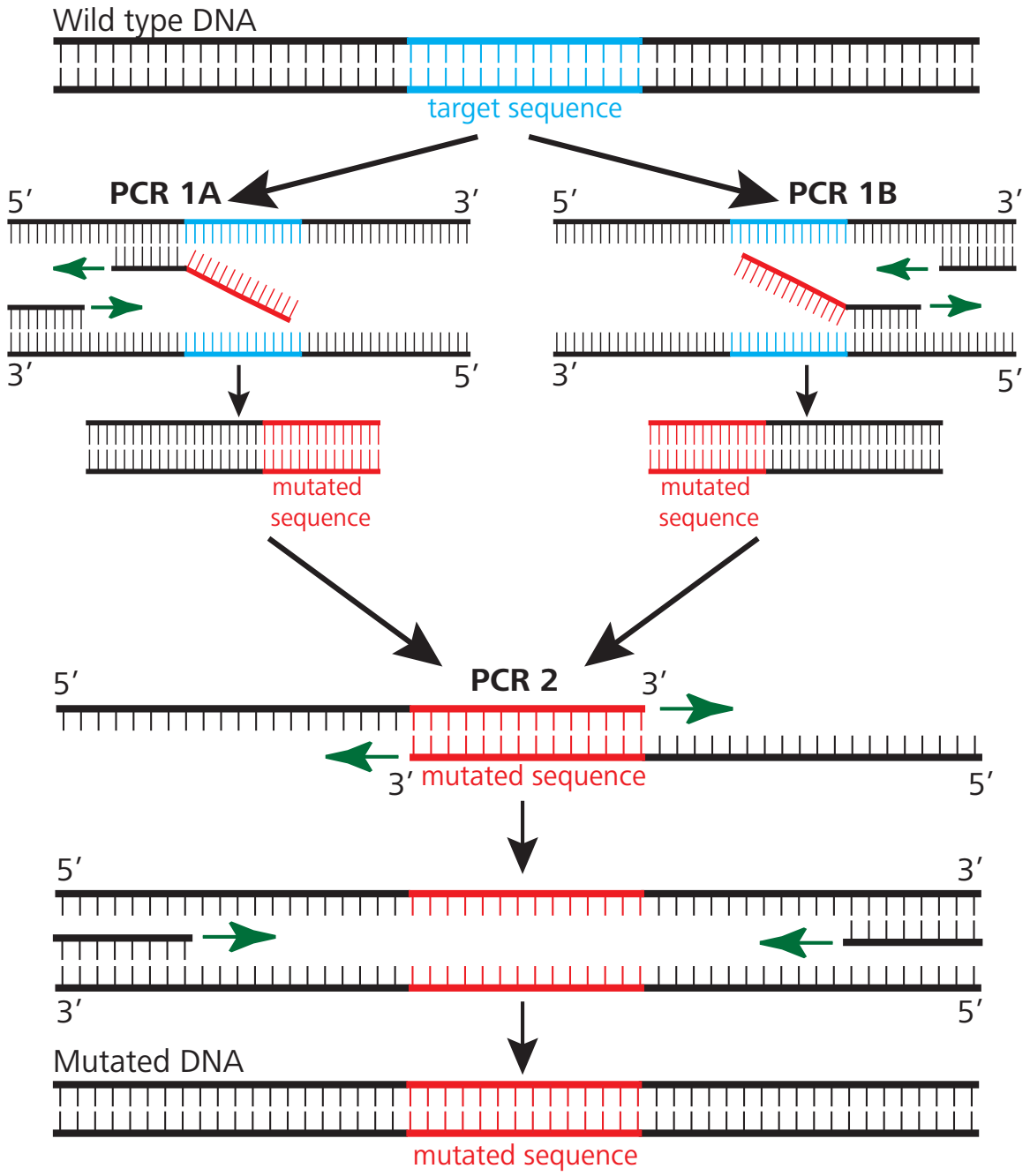

Figure 9. Schematic representation of the "Higuchi Method" method, used for site-directed mutagenesis by PCR.

Agarose Gel Electrophoresis

The size and quality check, as well as the separation of DNA fragments were performed by agarose gel electrophoresis. The electrophoresis was carried out in TAE or TBE electrophoretic buffers with 3-5 V/cm intensity of the electric field. For visualisation of the DNA on UV-transilluminator, ethidium bromide in final concentration 0,5 μg/ml was added into the agarose gel. The percentage of the agarose in the gel varied, depending on the size of the analysed DNA fragments: 0,7-0,8% for genomic DNA and fragments bigger than 10 kb, 1% for fragments between 1 and 10 kb and 1,5% for fragments smaller than 1 kb. Before loading, the stemples were supplemented with loading buffer, containing glycerol (to ensure the sinking of the sample to the bottom of the pocket) and two dyes (xylene cyanol, migrates approximately with 4000 kb DNA and bromphenol blue, migrates with 500 bp DNA) for controlling the migration of the samples. An appropriate DNA marker (DNA ladder) was loaded in parallel for determining the size and approximate quantity of the DNA samples.

Restriction digest of DNA

The digestion of DNA with restriction endonucleases was performed according to the instructions of the enzyme supplier. Approximately one unit of enzyme per 1μg DNA in appropriate buffered digestion reaction was used. If not otherwise specified by the manufacturer, the reaction was incubated for 1-4 hours on 37 °C, depending on the amount of DNA.

Extraction of DNA from agarose gel

For cloning purposes or microinjections, the DNA fragments were separated by agarose gel electrophoresis. The band containing the desired DNA fragment was cut out from the gel and the DNA was extracted with the SV Gel and PCR Clean-Up System or with the QIAquick Gel Extraction kit according to the manufacturer's instructions.

Dephosphorylation, blunting and ligation of DNA fragments

When needed for the cloning purposes, the DNA fragments were dephosphorylated (removing the phosphate on the 5′ end of the DNA). Normally, this step is required to prevent the self-ligation of a vector digested only with one or two restriction endonucleases, producing compatible cohesive ends. Calf intestine alkaline phosphatase was used for the dephosphorylation following the manufacturer's instructions. Before a ligation of DNA fragments with incompatible cohesive ends, a blunting (filling of a 5′-overhang or removing of a 3′-overhang) of the cohesive ends was performed, using T4 DNA polymerase, according to the supplier's instructions. The ligation of DNA fragments was performed with T4 DNA ligase as described in the user manual provided with the enzyme. Approximately 100 ng of vector DNA and 1-3 units of ligase were used in 20 μl ligation reaction. The amount of the insert was such, that the molar ratio of the free DNA ends between the vector and the inset be 1:3. In case of ligation of DNA fragments with cohesive ends, the reaction was incubated for 3 hours at room temperature and in case of fragments with blunt ends - for over night on 16°C.

TOPO-cloning

TOPO TA cloning kit (Invitrogen) was used for fast direct cloning of PCR-amplified fragments with T overhangs (fragments amplified with Taq polymerase or Taq polymerase based enzyme blends). When the PCR amplification resulted in one specific band 2-4 μl form the PCR were used directly (without any purification) for the cloning reaction; in the other cases the desired DNA fragment was purified from the PCR by agarose gel electrophoresis. Subsequently the purified fragment was adenylated (addition of an adenine on the 3′ end of the DNA fragment) before using it in the TOPO-cloning reaction. This adenylation step was necessary to increase the amount of the adenylated fragments (critical for the efficiency of the cloning reaction), which significantly decreases during the purification steps. The adenylation was performed by adding PCR buffer (to 1x final concentration), 1-2

units Taq polymerase and dATP to 250 μM into the solution, containing the purified fragment. The reaction was incubated for 15 min. at 72 °C and 2-4 μl were used for TOPO-cloning reaction. After 5-10 min. incubation at room temperature, the cloning reaction was transformed into TOP10 chemically competent cells (see below), provided with the kit.

Transformation of competent E. coli cells.

10-50 ng plasmid DNA or 10 μl of a ligation reaction (see Dephosphorylation, blunting and ligation of DNA fragments) were used for the transformation. The cells were incubated with the DNA for 10 min. on ice, heat shocked at 42°C for 45 sec. and placed again on ice for 2 min. When ampicillin resistance was used as a selection marker, the cells were immediately plated on LB-agar plates with ampicillin. In the case of other selection markers (kanamycin or chloramphenicol resistance), the cells were incubated for 1 hour in SOC medium without antibiotics and then plated on LB-agar plates with the respective antibiotic and incubated over a night on 37°C. The concentration of the used antibiotics was 100 μg/ml for ampicillin and 50 μg/ml for kanamycin and chloramphenicol.

Fish husbandry and care

The adult zebrafish stocks are maintained in the fish facility of the ITG, in an aquarium system build by Aquarienbau Schwarz (Göttingen). Approximately 15 pairs are 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 is set to 14 hours light and 10 hours dark. The fish are 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 is performed in one litre crossing cages, filled with system water. One fish pair is put in every cage in the evening. To avoid parental cannibalism the cage contains a sieve, which separates the eggs from the parents after the

laying. The laying starts 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.

The zebrafish embryos were kept in petri dishes with Hank's solution or system water. Antibiotics (ampicillin and gentamicin in concentration 50 μg/ml each) were added to the Hank's solution or methylene blue to the system water to prevent growing of bacteria and fungi. The embryos were incubated at 28-29°C until they reached the stage desired for the experiments or until day 4, if they were grown-up further to adulthood, then transferred to the main fish facility in specially prepared "baby-cages". They stayed there for approximately one month and then were moved to the ordinary fish tanks.

Preparation of injection solution

A circular plasmid DNA in concentration 10-20ng/μl was used for transient transgenic assays. For the generation of stable transgenic lines, a linear fragment (50-100 ng/μl) was used, containing the reporter gen region only (without any vector backbone). The injection solution was prepared by dilution of the DNA to the desired concentration in distilled water (final volume 10-20 μl) and addition of phenol red to final concentration of 1%. The phenol red serves as colour marker which makes the injected embryos distinguishable from the non-injected ones. The ready solution was filtered through a spin filter column (0,2 μm) and stored at -20 °C until needed.

DNA microinjections

The microinjection experiments were performed with FemtoJet (Eppendorf) or Gas microinjectors under Nikon SMZ645 stereomicroscope. The needles for the microinjection were prepared from borosilicate glass capillaries (0,7 mm inner and 1,0 mm outer diameter) on Flaming-Brown needle puller. Before the injections, the needles were filled with 1-3 μl injection solution (see Preparation of injection solution) using Eppendorf microloader pipette tips. The zebrafish eggs were collected

shortly after the fertilisation (zygote stage), transferred to a petri dish and the water was completely removed from the dish to facilitate the injection procedures. Each egg (zygote stage) was injected through the chorion into the cytoplasm with approximately 2-3 nl injection solution. System water supplemented with methylene blue was added into the dish after the injections and the embryos were kept at 28°C in incubator until the desired stage.

Expression analysis

The GFP expression was analysed on 24h old embryos using Leica MZ FLIII fluorescent stereomicroscope. The level of expression was quantified by counting the number of GFP positive cells in notochord and floor plate, as well as the number of ectopic GFP positive cells in tissues where *shh* and *twhh* are normally not expressed. Pictures were taken for the representative set of embryos, using maximum magnification and focusing on the trunk (the level of the yolk extension). The orientation of the embryos on the picture is anterior to the left and dorsal to the top.

Isolation of shh and twhh intron 2 sequences

The tench *shh* and *twhh* intron 2 fragments were isolated by using degenerative oligonucleotides, designed on the base of conserved protein blocks in the second and third exon of *shh* and *twhh* genes from several vertebrate species. The PCR products were directly cloned into *pCRII-TOPO* vector (Invitrogen) and the clone containing the right insert was identified by sequencing.

The *Latimeria* intron 2 was isolated by screening of genomic BAC library from *Latimeria. menadoensis* (Danke et al. 2004 and M. Lang et al., in preparation), kindly provided by Chris Amemiya. The positive BAC clone, containing the *shh* locus was shot gun sequenced and relevant genomic regions were secondarily amplified by gene specific primers. The correct PCR product was identified by sequencing. The mouse and chick intron 2 were directly amplified from genomic DNA with specific oligonucleotides containing NotI/KpnI restriction sites.

Plasmid construction

The *0.8shh:gfp* plasmid was constructed by cutting out the SalI/HindIII fragment from *2.4shh:gfp* plasmid (Ertzer et al. 2007) (described as *2.2shh:gfp* in (Chang et al. 1997; Muller et al. 1999); and subsequent, blunting and religating.

The *0.8shh:gfp:z-shh-I2*, *0.8shh:gfp:z-shh-arC* and *0.8shh:gfp z-twhh-I2* were created by subcloning the respective NotI/KpnI fragments from *2.4shh:gfp:C* (Ertzer et al. 2007), *2.4shh:gfp:ΔC* and *2.4shh:gfp:twhh C* (F. Müller, U. Strähle, and N. Fischer, unpublished) into *0.8shh:gfp* plasmid.

The plasmids *0.8shh:gfp:t-shh-I2* and *0.8shh:gfp:twhh-I2* were made by reamplifying the respective intron 2 fragments from *pCRII-TOPO:t-shh-I2*, *pCRII-TOPO:t-twhh-I2* and subcloning them in *0.8shh:gfp* using NotI/KpnI restriction sites.

The *0.8 shh:gfp:l-shh-I2* was constructed by reamplifying the intron 2 part from the correct PCR fragment isolated from the BAC clone and cloning it into *0.8shh:gfp* (NotI/KpnI).

The *0.8 shh:gfp:m-shh-I2* and *0.8 shh:gfp:c-shh-I2* were created by direct cloning of the PCR-amplified intron 2 sequences from mouse and chick into 0.8shh:gfp (NotI/KpnI.).

The *0.8 shh:gfp:z-twhh-non-cons.* and *0.8shh:gfp:z-twhh-arC* were made by cloning the PCR-amplified non-conserved 5′ part of z-*twhh I2* (1032 bp) and the 380 bp 3′ part containing the conserved region (*ar-C*) into *0.8shh:gfp* (NotI/KpnI).

All plasmids (*0.8shh:gfp:z-shh-arCΔC1*; *0.8shh:gfp:z-shh-arCΔC2*; *0.8shh:gfp:z-shh-arCΔC3*; *0.8shh:gfp:z-shh-arCΔC4*; *0.8shh:gfp:z-twhh-arC*ΔC1; *0.8shh:gfp:z-twhh-arCΔC3*), containing *z-shh-ar-C* or z-*twhh ar-C* carrying mutations in one of the conserved motifs (C1 to C4) were created by replacing the respective wild type sequence of each conserved block with random sequence using a PCR-based approach (see Site-directed mutagenesis by PCR).

The same method was used to introduce the C2 and C4 from z-*shh ar-C* or random sequence into z-*twhh ar-C* (*0.8shh:gfp:z-twhh-arC+C2*; *0.8shh:gfp:z-twhh-arC+C4*; *0.8shh:gfp:z-twhh-arC+C2rnd* and *0.8shh:gfp:z-twhh-arC+C4rnd*). The PCR products

were cloned into *0.8 shh:gfp* (NotI/KpnI) and verified by sequencing.

For more detailed information about the sequences, which have been mutated and introduced in *shh* and *twhh ar-Cs*, see Table 1.

The plasmid *2.7twhh:gfp* was constructed by replacing the *2.4shh* promoter fragment (SalI/XhoI) from *2,4shh:gfp* with the PCR-amplified 2.7 kb *twhh* promoter fragment (upstream from the translation start site).

The plasmid *2.7twhh:gfp:z-twhh-I1* and *2.7twhh:gfp:z-twhh-I2* were made by subcloning the *twhh I1* and *I2* from *2.4shh:gfp: twhh-I1* and *2.4 shh:gfp:twhh-I2* (F. Müller, U. Strähle and N.Fischer, unpublished) into *2.7twhh:gfp* (NotI/KpnI).

Electronic version of all intronic sequences can be found on the included with this thesis CD.

For sequence information on the used oligonucleotides, see Table 2.

Sequence alignments and analysis

Pairwise sequence aliments were performed using one of the global alignment algorithms: AVID (Bray et al. 2003), in the case of the intronic sequences or Shuffle-Lagan (Brudno et al. 2003) in case of the whole hh loci and visualised by Vista (Frazer et al. 2004; Mayor et al. 2000) http://genome.lbl.gov/vista/index.shtml.

The multiple alignments of the intronic sequences were made by using combination of two algorithms: CHAOS/DIALIGN (Brudno et al. 2004), http://dialign.gobics.de/chaos-dialign-submission and visualised by BioEdit (sequence alignment editor written by Tom Hall, Ibis Therapeutics, Carlsbad, CA92008: http://www.mbio.ncsu.edu/BioEdit/bioedit.html.

Construct name	WT sequence	Mutated/Introduced sequence
0.8shh:gfp:z-shh-arC C1	TGCACCTGAGCAAATA	GTACAAGTCTACCCGT
0.8shh:gfp:z-shh-arC C2	GAAGTGTCCTTTTCCAAGAGT	TCCTGTAAGCCCAAGCTCTAC
0.8shh:gfp:z-shh-arC C3	AATGACAATGTCC	CCGTCACCGTGAA
0.8shh:gfp:z-shh-arC C4	CTTTATTGGTTTTTAATTAGA	AGGGCGGTTGGGGGCAGGCGG
0.8shh:gfp:z-twhh-arC C1	TGCACCTGTGTAAACA	GTACAAGTCTACCCGT
0.8shh:gfp:z-twhh-arC C3	TTTAAATGACAATGTCT	GGCTCCGTCACCGTGAA
0.8shh:gfp:z-twhh-arC+C2	CAGGGAAAAGCACAGTCTGT	GAAGTGTCCTTTTCCAAGAGT
0.8shh:gfp:z-twhh-arC+C4	GACTTTGTGTAAATTCAGCAG	CTTTATTGGTTTTTAATTAGA
0.8shh:gfp:z-twhh arC+C2rnd	CAGGGAAAAGCACAGTCTGT	TCTCCAGGCTCAACCATGAGC
0.8shh:gfp:z-twhh-arC+C4rnd	GACTTTGTGTAAATTCAGCAG	AGAAAGCTCGCGCGACCATGA

Table 1. Sequences, which have been used to replace the wild type sequence in *shh* and *twhh ar-Cs* to generate the specified reporter constructs.

Sequence Name	Forward Primer	Reverse Primer
tench *shh intron 2*	GCIGGITTYGACTGGGTCTA (degenerative, used for isolation)	GAGTACCAGTGSAYICCIKC (degenerative, used for isolation)
	GTAAGACCATGGCAGGATG (specific, used for subcloning)	TCGAGATAATAGCAATGGGT (specific, used for subcloning)
tench *twhh intron 2*	GCIGGITTYGACTGGGTCTA (degenerative, used for isolation)	GAGTACCAGTGSAYICCIKC (degenerative, used for isolation)
	GTGAGAGCAATGTCACC (specific, used for subcloning)	GCGATAAAAGTAAAAAGAGAC (specific, used for subcloning)
latimeria *shh intron 2*	TCAAAGCAGGTAAGCAGACG	AAGCAACCCCCTGATTTTG
mouse *shh intron 2*	GTGGAAGCAGGTTTCGACTG	GAAAGACCAGGTGTTGAGTGC
chick *shh intron 2*	CGGCTTCGACTGGGTCTAC	GCTGCCACTGAGTTTTCTGC
zebrafish *twhh ar-C*	CCGAATAACAACAACTCGCAATC	CTGAGAAGATATACAAACACAA
zebrafish *twhh intron 2* non-conserved part	GTGAGCAAAAGCTGATATGC	GATTGCGAGTTGTTGTTATTCGG
2.7 kb zebrafish *twhh promoter*	CATCTAAATCAACTGCAAGAACG	GACGTTTGAATTATCTCTTCTGGTC

Table 2. Primer sequences used for the amplification of the specified fragments. In the degenerative oligonucleotides where the occurrence of all four nucleotides was equally possible, an inosine (I) was introduced to reduce degeneracy. On all specific primers, restriction enzyme sites were added

Results

Isolation of shh and twhh intron 2 sequences.

We aim to prove our main hypothesises for the mechanisms of enhancer divergence and subfunctionalisation between *shh* and *twhh* zebrafish in this study by performing a systematic sequence and functional analysis of *ar-C* enhancers from vertebrates species on important time points in evolution to understand how the *ar-C* enhancer has diverged on sequence and functional levels after the gene duplication. For these purposes we focused on isolation of *ar-C* enhancers (intron 2 sequences) form *shh* and *twhh* genes from vertebrate species representing divergence at important time points of evolution: preduplicted *shh* genes from species which did not undergo a genome duplication and have only one *shh* gene, e.g all sarcopterigyans (land vertebrates and lobe-finned fishes) species; form specices which went trough genome duplication and as a result have two *shh* copies (*shh* and *twhh*), e.g most of the actinopterigyans (ray-finned fishes) (Fig. 10).

Selective divergence of twhh non-coding sequences from shh genes

First, to check if cis-regulatory elements, corresponding to the already described elements of *shh* genes, can identified on bases of sequence conservation in the zebrafish *twhh* gene as well, we have performed a sequence comparison of multiple vertebrate *shh* loci (zebrafish, fugu, chick, mouse and human) and the zebrafish *twhh* locus, extracted from ENSEMBL database (http://www.ensembl.org). A global alignment using shuffle Lagan algorithm and visualisation by VISTA plot clearly identifies all 3 exons of *shh* orthologs and paralogs throughout vertebrate evolution (Fig. 11). The cis-regulatory modules identified previously are conserved among *shh* genes (orange peaks) and the degree of their conservation agrees with the evolutionary distance between the species compared. In contrast, the zebrafish *twhh* gene shows no obvious conservation with the *shh ar-A, B, C* and *D* cis-regulatory modules. Besides Shuffle Lagan, Valis (Sanges et al. 2006) has also failed in detect-

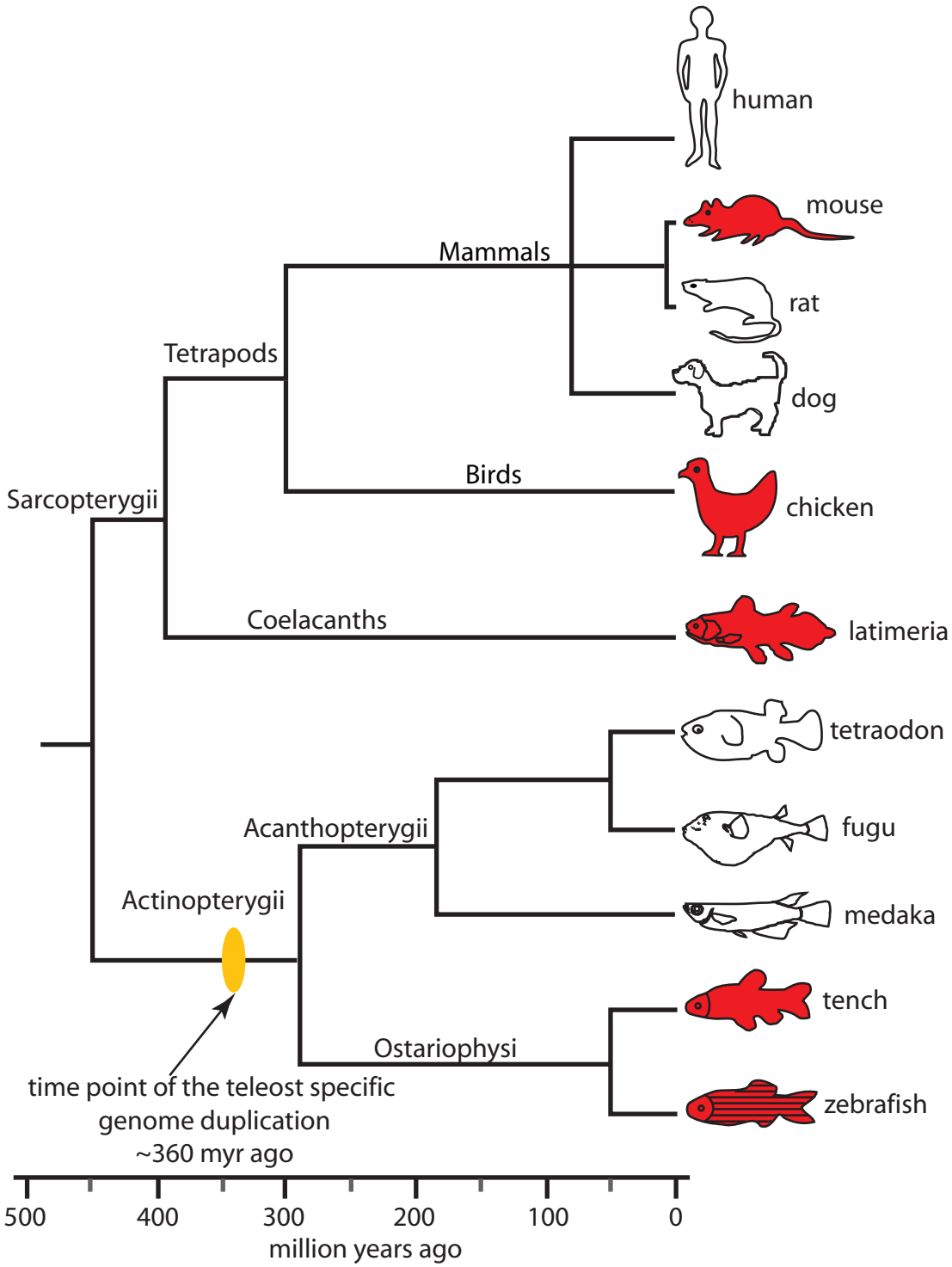

Figure 10. Evolutionary relationship of vertebrates. The phylogenetic tree represents the evolution of some main vertebrate groups in a time scale (see the time ruler on the bottom). The yellow eclipse marks the time point of the extra genome duplication in actinopterygians. The highlighted in red species are those which ar-C/intron 2 sequences have been isolated and functionally analysed in this study. The sequences from non-highlighted have been used in sequences comparisons, but have not been functionally analysed (see Fig. 20)

ing conserved putative cis-regulatory modules of *twhh*. Taken together, these results indicate that, although orthologous regulatory elements may exist between *twhh* and *shh*; however, they are much less conserved at the DNA sequence level than *shh* elements as detected by the applied alignment programmes.

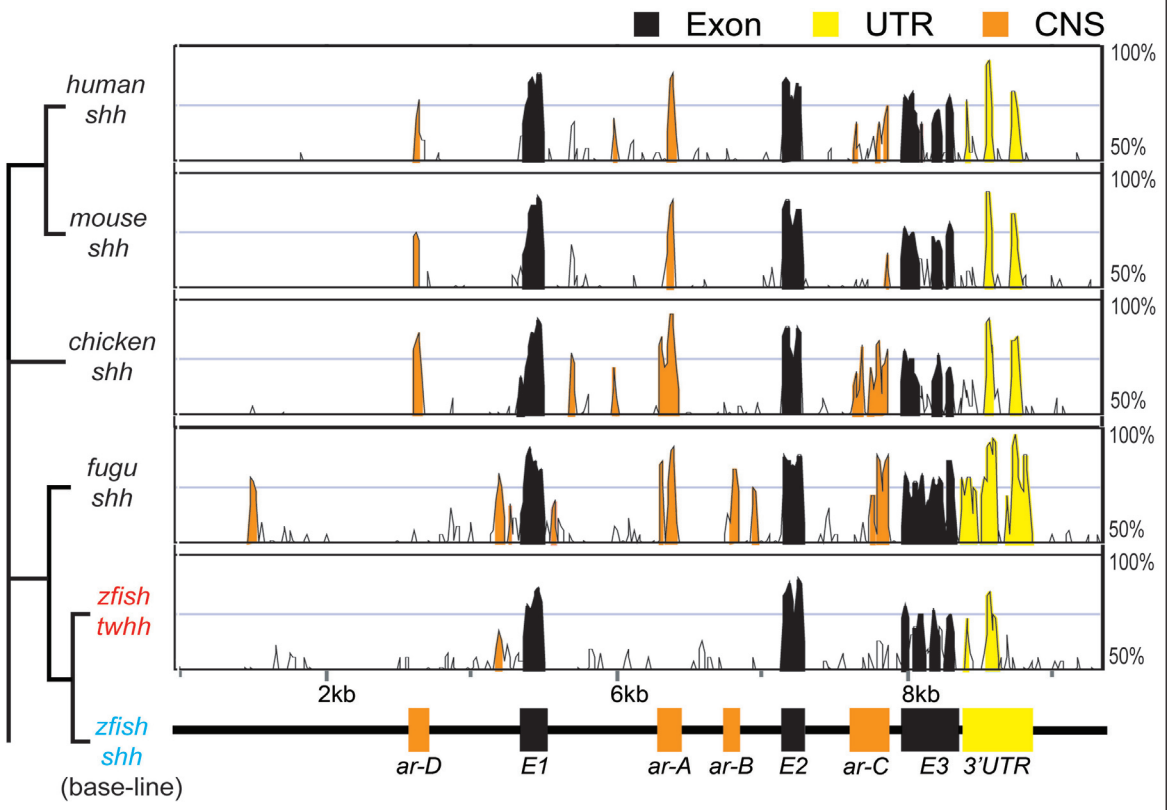

Figure 11. Selective divergence of *twhh* non-coding sequences from that of *shh* genes. Vista plot of Shuffle-Lagan alignment of *shh* and *twhh* gene loci from different vertebrate species. The zebrafish *shh* locus is the base sequence to which the other *hedgehog's* loci are compared. The peaks with more than 70% identity in a 50 base-pairs window are highlighted in colour (colour-legend on top). On the bottom of the plot a scheme of the zebrafish *shh* locus marks the position of the exons, known cis-regulatory elements and the *3'UTR*. The phylogenetic tree on the left side of the plot represents the evolutionary relationship of vertebrates. Abbreviations: zfish, zebrafish; E, exon; CNS, conserved non-coding sequence; UTR, untranslated region; ar, activation region

The ar-C enhancer is a highly conserved midline enhancer of vertebrate shh genes

To to learn more about the evolutionary processes acting on CRMs of *shh* genes, we focused on a single enhancer element *ar-C*, which is conserved between fish and mouse (*SFPE2*) and that has been analysed in considerable detail in both species (Epstein et al. 1999; Jeong and Epstein 2003; Muller et al. 1999). To this end, first we addressed, whether the *ar-C* enhancer or its mouse ortholog *SFPE2* is detectable across *shh* loci in various vertebrate species from different lineages that have diverged before and after the gene duplication event leading to the evolution of *shh* paralogs in zebrafish. Since the zebrafish *shh ar-C* enhancer is located in the second intron of *shh* and shows high sequence similarity to human and mouse counterparts, candidate *ar-C* containing intronic fragments of several vertebrate species were amplified by PCR with degenerate oligonucleotide primers. We cloned and sequenced the relevant genomic DNA fragments from several fish species that experienced the genome duplication such as the cyprinid tench (*Tinca tinca*), fugu, and, medaka (Hoegg et al. 2004). Besides actinopterygian fishes several species of sarcopterygians such as chick, mouse and the early sarcopterygian lineage *Latimeria menadoensis* were used in the analysis. All sarcopterygians diverged from the common ancestor with actinopterygians prior to the fish specific genome duplication in the ray-finned fish lineage. A sequence comparison of intron *SFPE2* sequences from the available vertebrate model systems revealed a high degree of sequence similarity in all species specifically in the region that spans the *ar-C* enhancer in zebrafish and the *SFPE2* enhancers of mouse (Fig. 12). This analysis also indicated that the orthologous *Latimeria* genomic region also contains a highly conserved stretch of sequence in the *ar-C* region, consistent with the hypothesis that *ar-C* is an ancestral enhancer of *shh* genes.

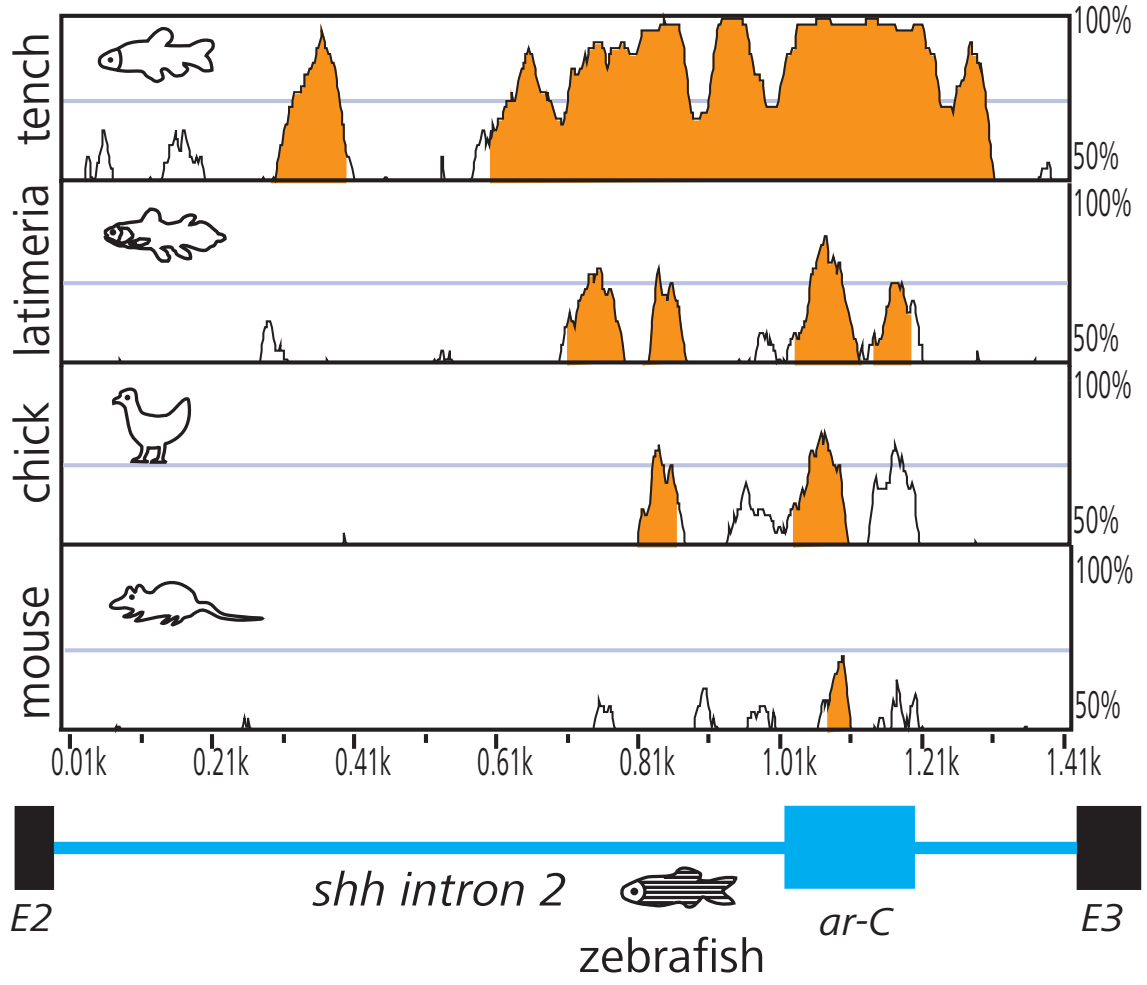

Figure 12. The *ar-C* enhancer is a highly conserved midline enhancer of vertebrate *shh* genes. Vista plot comparison (AVID-global sequence alignment algorithm) of *shh* intron 2 form zebrafish (base line), mouse, chick, *Latimeria* and tench (bottom to top order). The peaks showing more than 70% identity in a 50 base-pairs window are highlighted in orange. The scheme of the zebrafish *shh* intron 2 on the bottom marks the position of the zebrafish *ar- C* (blue rectangle) and the second and third exons (black rectangles). Abbreviations: E, exon; ar, activation region

Heterologous ar-C enhancers function in the notochord of zebrafish

To test whether the sequence similarity observed between *ar-C* enhancers of different lineages of vertebrates is also indicative of conserved tissue-specific enhancer function, we carried out transgenic analysis in microinjected zebrafish embryos of enhancers form different vertebrates. We utilised a minimal *shh* promoter construct, containing an 0.8 kb upstream sequence linked to *gfp* reporter, with an activity similar of a 563 bp promoter described in (Chang et al. 1997) . Transient mosaic expres-

sion of GFP was measured as read-out of reporter construct activity by counting fluorescence positive cells in the notochord and floor plate where the *ar-C* enhancer is active in the trunk of the 1-day-old embryo. As described previously, the zebrafish *ar-C* enhancer is primarily active in the notochord and only weakly in the floor plate (Fig. 13B). As expected for closely related species, intron 2 sequence of tench, gave strong enhancer activity in the notochord as well (Fig. 13C). Very similar was the ac-

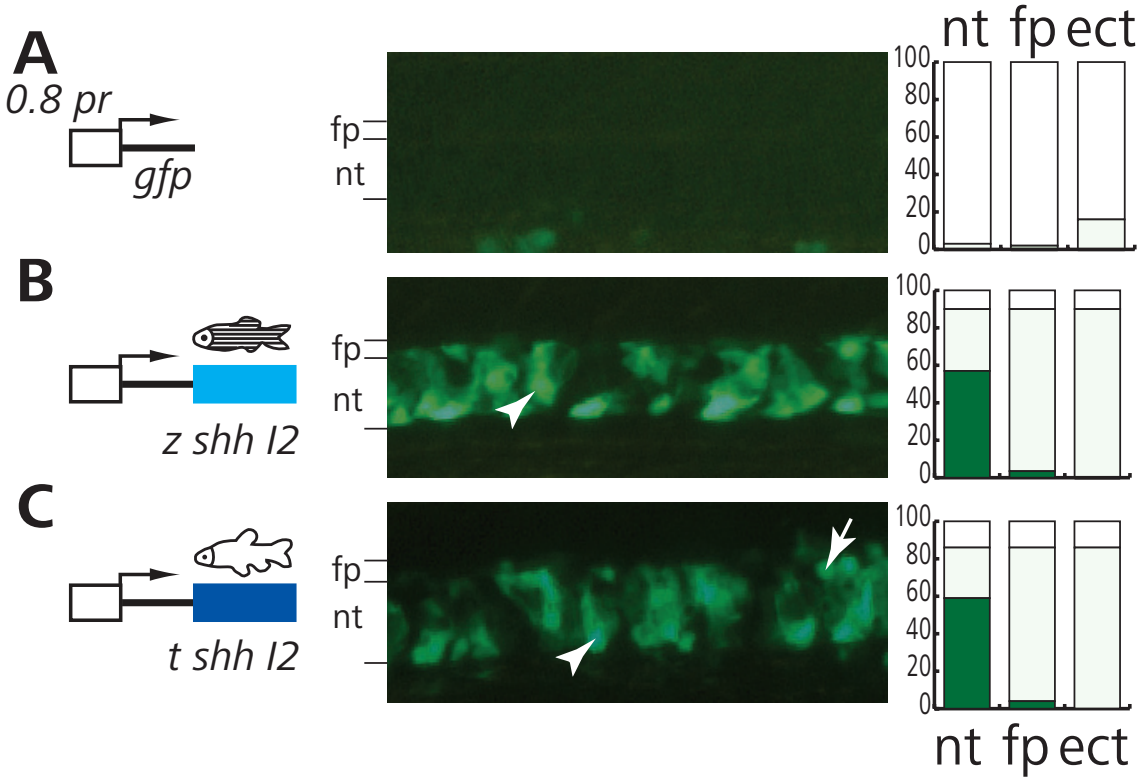

Figure 13. *Shh* intron 2 fragments from cyprinids drive predominantly notochord expression. Microinjected embryos are shown at 24 hpf with lateral view onto the trunk at the level of the midline. A, Zebrafish embryo injected with control *gfp* reporter construct, containing a minimal 0.8 kb zebrafish *shh* promoter. B-C, Embryos injected with *gfp* repoter construct containing *shh* intron 2 sequences from zebrafish (B), tench (C) The stacked-column graphs on the right side represent the quantification of the transient *gfp* expression. Arrows and arrowheads indicate GFP activity in the floor plate and notochord cells respectively. The lines on the left side of each image mark the level of the notochord and the floor plate. The columns show the percentage of the embryos with more than 15 GFP-positive cells per embryo (dark-green), embryos with less than 15 cells (light-green), and non-expressing embryos (white). Numbers of injected embryos are given in Table 3. Abbreviations: z, zebrafish; t, tench; I intron; nt, notochord; fp, floor plate; ect, ectopic; pr, promoter.

tivity of the intron 2 sequences form the sarcopterygian species *Latimeria* and chick (Fig. 14A,B). However, the mouse intron 2 (with the *SFPE2* enhancer) was found to be inactive in zebrafish (Fig. 14C), suggesting that *SFPE2* had functionally diverged during mammalian/mouse evolution either at the cis- or the trans-regulatory level. All together, these data indicate a high degree of functional conservation between *ar-C* sequences among vertebrates.

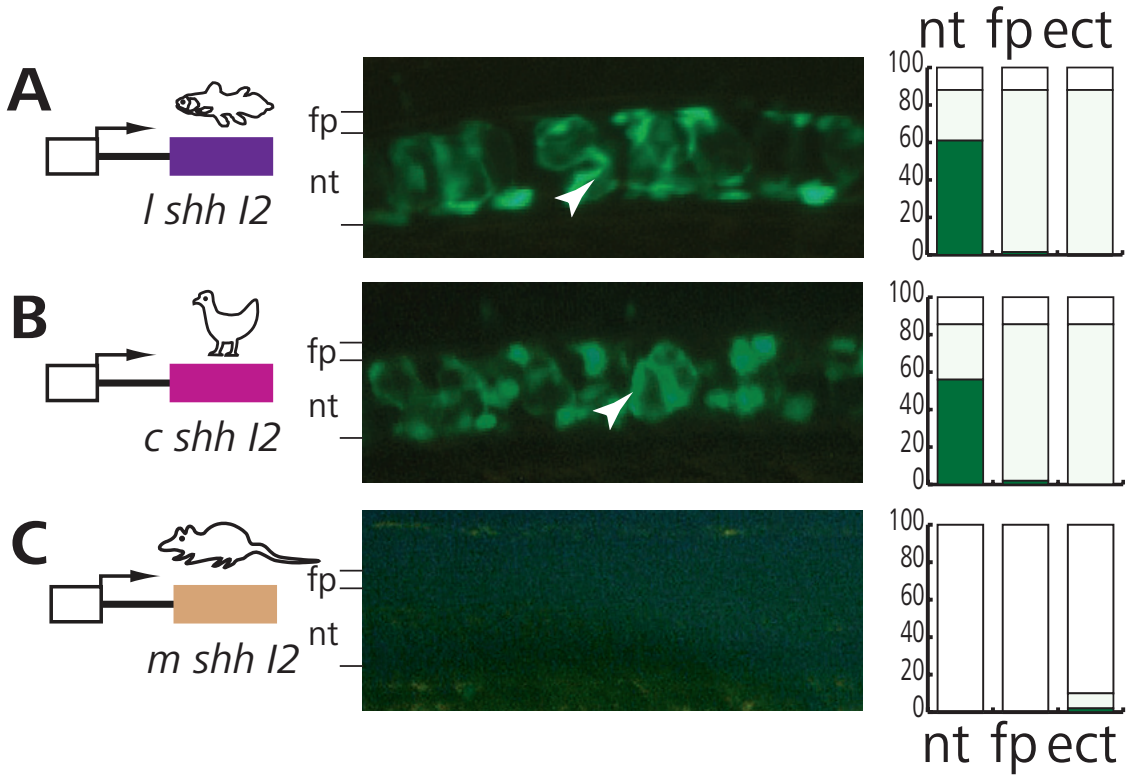

Figure 14. *S*arcopterygian *ar-Cs* **are functional in zebrafish and show notochord activity.** Embryos injected with *gfp* repoter construct containing *shh* intron 2 sequences from *Latimeria* (A), chick (B), and mouse (C). Abbreviations: l, *Latimeria*; c, chick; m, mouse; I intron; nt, notochord; fp, floor plate; ect, ectopic; pr, promoter.

Identification of a putative ar-C enhancer from twhh genes

As next to ask was whether the duplicated *twhh* gene possesses a functional *ar-C* enhancer homolog. The first question was if sequence alignment can reveal conservation of *ar-C* regions between *shh* paralogs. To this end isolation of the *twhh* intron 2 from zebrafish was carried out. Since a genome duplication event has taken place early in actinopterygian evolution it was predicted that the ostariophysian and cyprinid zebrafish as well as all acanthopterygian fish model species whose genomes are known (medaka, stickleback, green spotted pufferfish and fugu) may contain a *twhh* homolog. Analysis of the available genome sequences of these four species of teleost fish indicated that none of them carries a discernible *twhh* homolog suggesting that these lineages (that evolved some 290 myrs after cyprinids (Steinke et al. 2006) may have secondarily lost this *shh* paralog. To collect further evidence for the secondary loss of *twhh* in medaka, a synteny comparison of the respective chromosomes between zebrafish and medaka was performed. As shown on Fig. 15

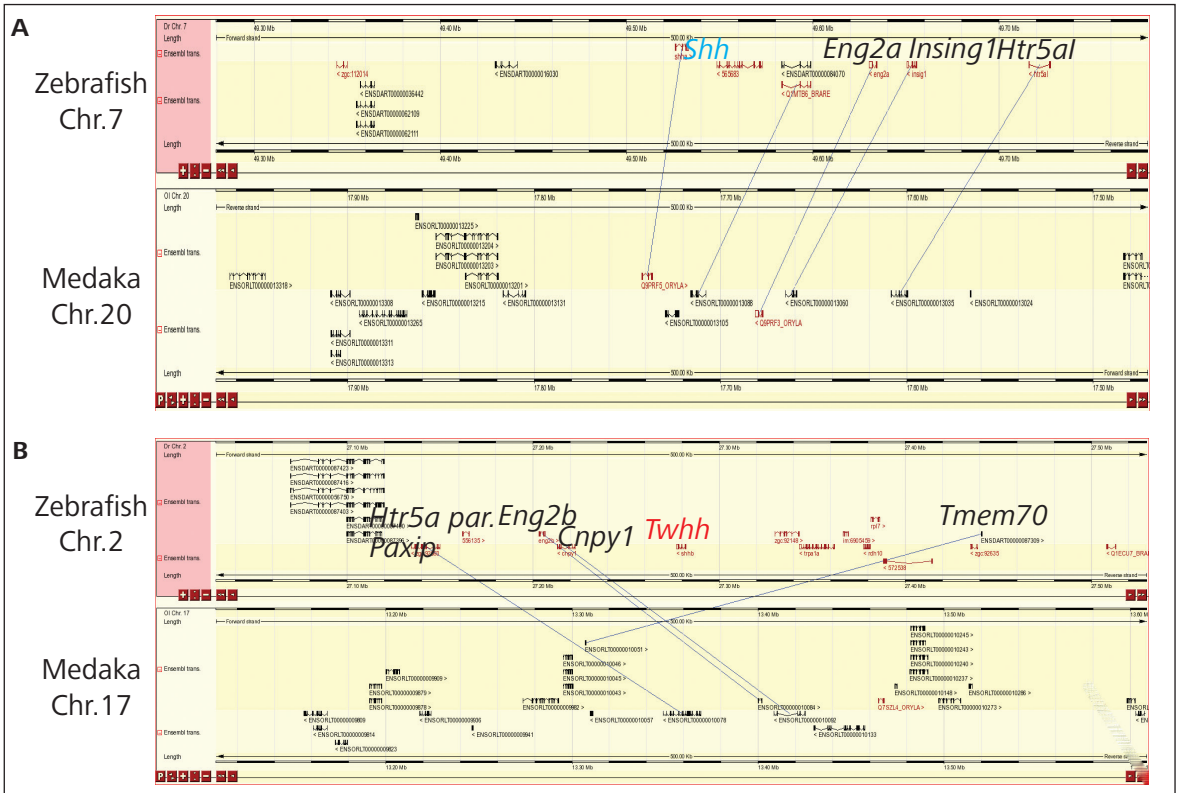

Figure 15. Synteny comparison of *shh* and *twhh* containing chromosomes suggests the loss of a duplicated *shh* paralog gene in medaka. Ensembl view of zebrafish chromosome 7, containing the *shh* locus alongside with medaka chromosome 20 (A) and zebrafish chromosome 2, containing the *twhh* locus alongside with medaka chromosome 17 (B) are shown.

good synteny is observed between the medaka genomic region surrounding *shh* on chromosome (chr) 20 and a region on chr 17, however, chr 17 lacks *twhh*. This result further supports the hypothesis that a *twhh* gene was originally present after duplication but has been lost secondarily during evolution.

To obtain further examples of duplicated shh genes we focused on evolutionary closely related to zebrafish fish species, e. g. species of the *Cyprinidae* family. We were able to detect and isolate *twhh* and its intron 2 from another cyprinid species, tench by PCR using degenerate oligonucleotides that were designed in conserved exon sequences. Importantly, the isolation of more than one *twhh* intron 2 sequences from cyprinids allowed for phylogenetic footprinting of *twhh* genes and search for a putative *ar-C* homolog. we compared the *shh* and *twhh* intron 2 sequences between zebrafish and tench (Fig. 16). The *shh* orthologs between zebrafish and tench show a high degree of sequence similarity which is strongest in the region in which *ar-C* resides. In contrast, comparison of intron 2 from *twhh* and *shh* paralogs of either species revealed no conspicuous conservation. The apparent lack of sequence similarity, however, does not necessarily rule out the possibility that a highly diverged *ar-C* homolog enhancer may still reside in *twhh* intron 2. To test this possibility we compared *twhh* sequences from both species. This indicated striking sequence similarity in the 3` region close to exon 3 where a positionally conserved *ar-C* would be predicted to be located. This suggests, that intron 2 of *twhh* genes of cyprinids may contain a functional enhancer, which has diverged significantly from the *shh ar-C*. Furthermore, the apparent sequence divergence suggests that the putative function of the *twhh* enhancer may also have diverged.

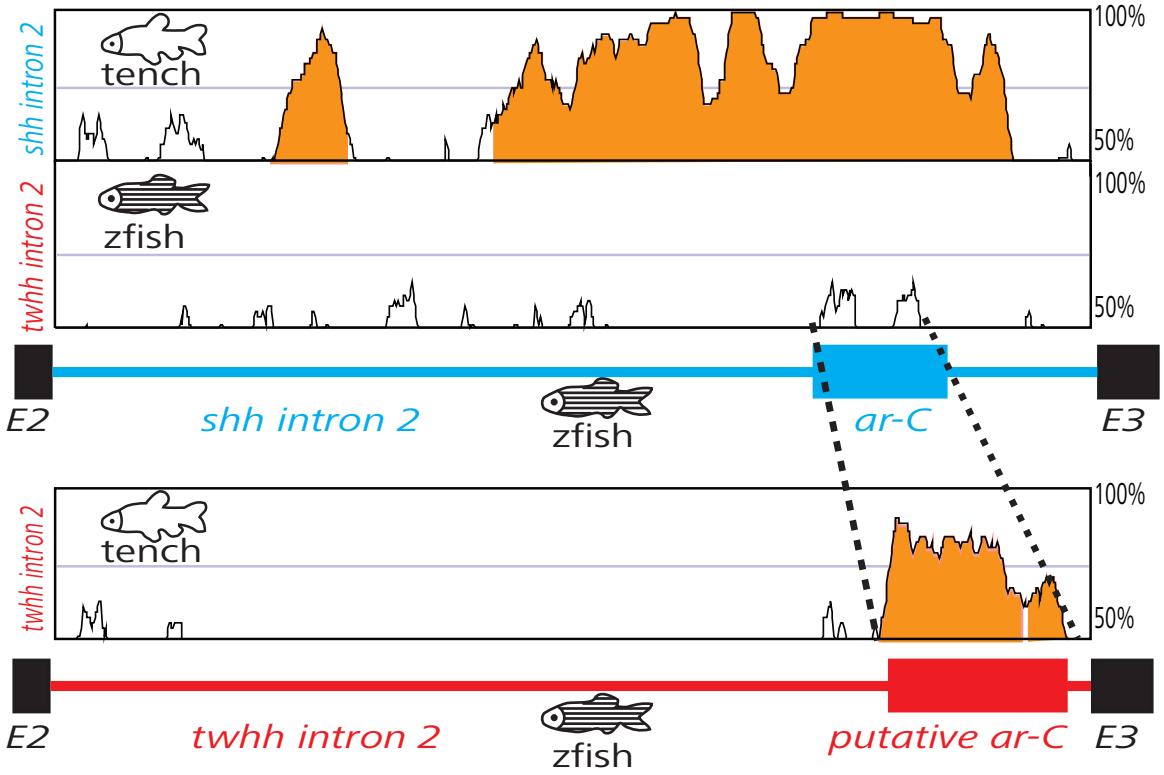

Figure 16. ***Twhh*** **genes carry an** ***ar-C*** **homolog enhancer with diverged sequence.** Top panel: Vista plot comparison (AVID) between zebrafish *shh* intron 2 (base line) zebrafish *twhh* intron 2 and tench *shh* intron 2. Bottom panel: comparison between zebrafish (base line) and tench *twhh* intron 2. The peaks showing more than 70% identity in a 50 base-pairs window are highlighted in orange. The schemes of zebrafish *shh* (top) and *twhh* (bottom) intron 2 mark the position of the *shh ar-C* (blue box), the putative *twhh ar-C* (red box) and exons 2 and 3 (black boxes). Dashed lines demarcate equivalent sequence regions.

The diverged ar-C enhancer of twhh is functionally active

To test whether the conserved sequence in the intron 2 of *twhh* genes functions as enhancer element, we tested several *twhh* fragments representing approximately 10 kb of the *twhh* locus including the 1.4 kb intron 2 in transgenic reporter assays. The *twhh* proximal promoter and 2.7 kb of upstream sequences can activate GFP expression in the notochord (Fig. 17A) but only very weakly in the floor plate similarly to previously published data (Du and Dienhart 2001). Since *twhh* is only expressed in the floor plate but never in the notochord, this GFP expression of the reporter is an ectopic activity and reflects the lack of a notochord repressing functional element probably located elsewhere in the unexplored sequences around the *twhh* locus.

The weak expression in the floor plate suggests that other cis-regulatory modules are required for floor plate activation. In *shh* a floor plate enhancer resides in intron 1 (Muller et al. 1999) (Fig. 7, Fig. 11). To check if a similar enhancer exists in *twhh*, intron 1 of *twhh* was attached to the promoter construct. It was found that it did not enhance the promoter's activity, indicating no obvious enhancer function in this transgenic context (Fig. 17B). Interestingly, the addition of *twhh* intron 2 does, however, result in enhancement of expression in the floor plate (Fig. 17C). This result indicates that intron 2 of *twhh* contains a floor plate enhancer.

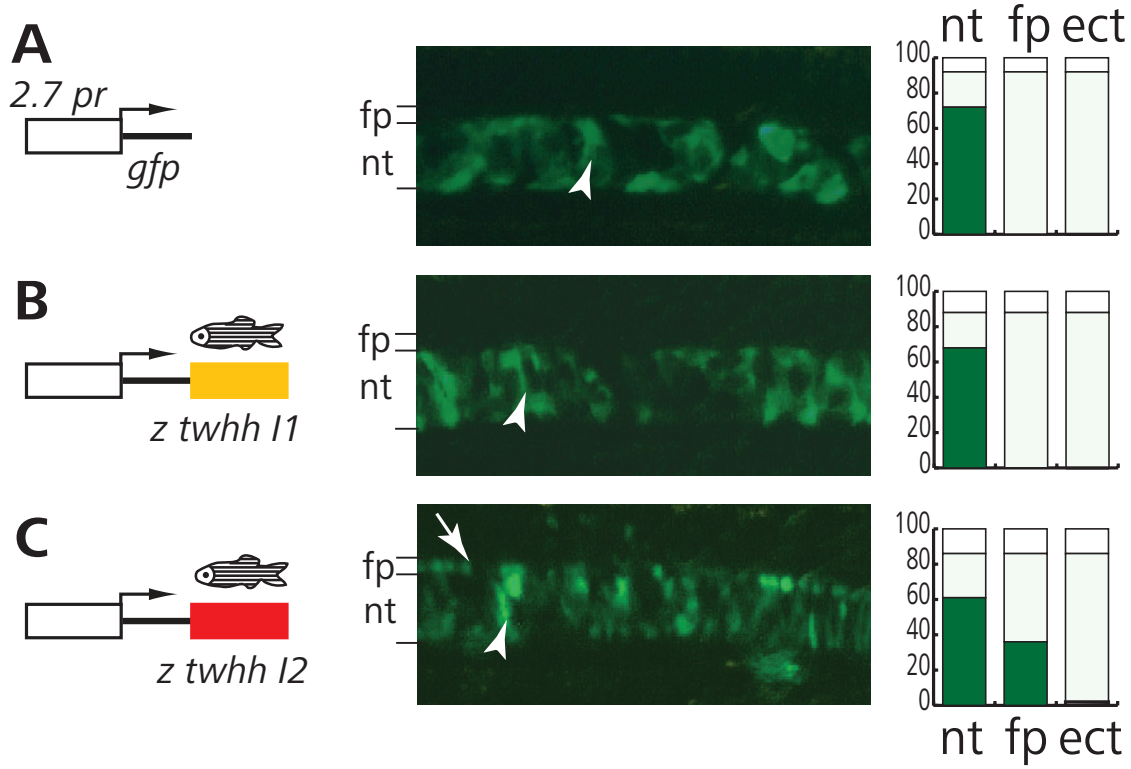

Figure 17. Transgenic analysis of *twhh* genomic fragments for enhancer activity. Embryos injected with the plasmid constructs are shown at 24 hpf, lateral view onto the trunk at the level of midline. Embryos injected with *gfp*-reporter constructs containing zebrafish 2.7 kb *twhh* promoter (A), 2.7 kb *twhh* promoter plus zebrafish *twhh* intron 1 (B) and *twhh* intron 2 (C).

The 2.7 kb upstream and proximal promoter sequence of *twhh* may have influenced the autonomous function of an enhancer in intron 2. To address the activator functions of the identified *shh* and *twhh* enhancers without influence of potential upstream regulatory elements, a series of injection experiments was carried out in which the enhancer activities were analysed with a minimal promoter containing only 0.8 kb of the *shh* promoter (Fig. 18A-D). Moreover, activity of intron 2 sequences from *shh* and *twhh* genes from both zebrafish and tench were systematically compared. Intron 2 fragments of both species consistently resulted in comparable notochord activity (Fig. 18B and Fig. 23B,C), while the *twhh* intron 2 fragment from both species showed the distinct enhancement of expression in the floor plate and reduction of GFP activity in the notochord (Fig. 18C,D).

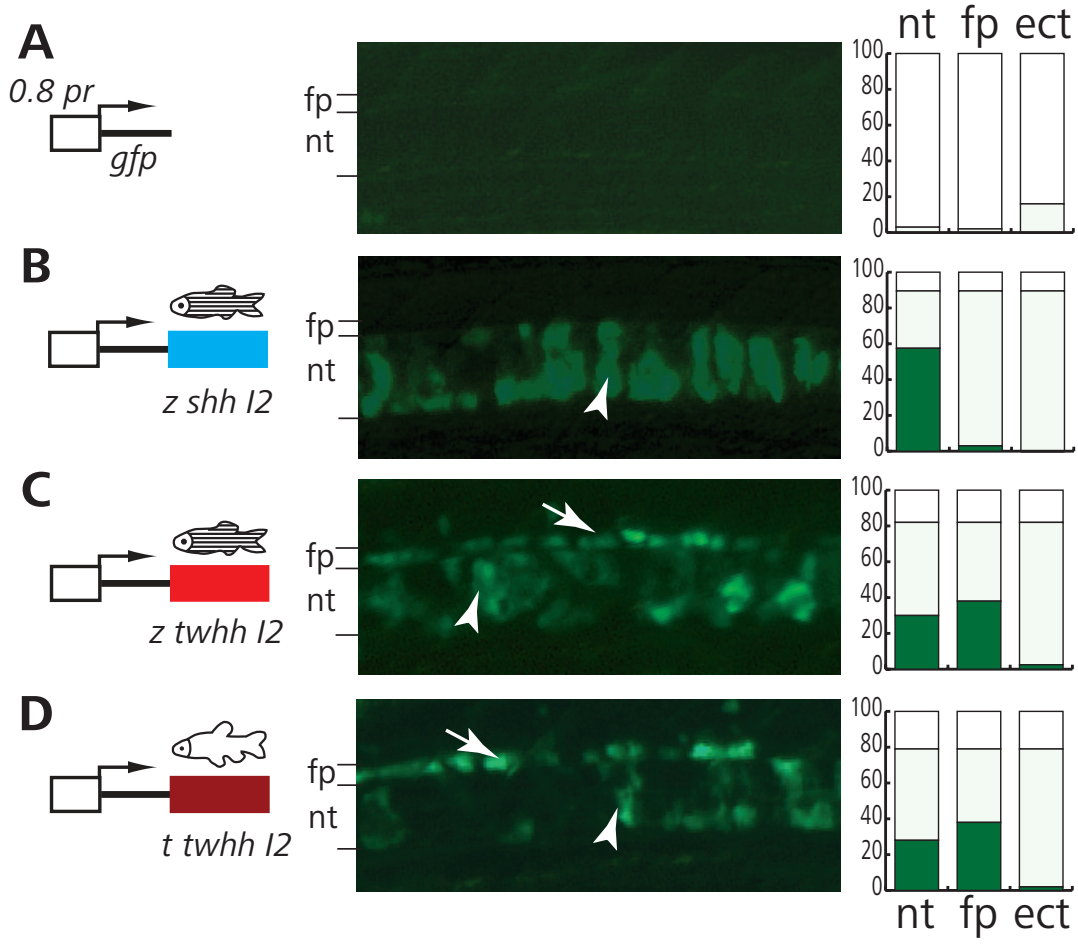

Figure 18. ***Twhh*** **intron 2 from zebrafish and tench contains a functional enhancer diverged in function from that of** ***shh*****.** Embryos injected with promoter-control construct (A), plasmids containing zebrafish *shh* intron 2 (B), zebrafish *twhh* intron 2 (C), tench *twhh* intron 2 (D).

The presence of a highly conserved region within the intron 2 of zebrafish and tench *twhh* genes strongly suggests that the floor plate enhancer activity is the property of this conserved sequence. To test this prediction a set of deletion analysis experiments was carried out. Zebrafish *twhh* intron 2 was cleaved into a 1026 bp fragment of non-conserved and a 380 bp conserved sequence. As shown in Fig. 19A,B, the floor plate specific enhancer effect is retained by the conserved fragment but not the non-conserved sequence, verifying the prediction of the location of the floor plate enhancer. Taken together, a diverged, floor plate active *ar-C* enhancer has been discovered in the *twhh* intron 2, which is consistent with the floor plate specific expression of *twhh* in zebrafish.

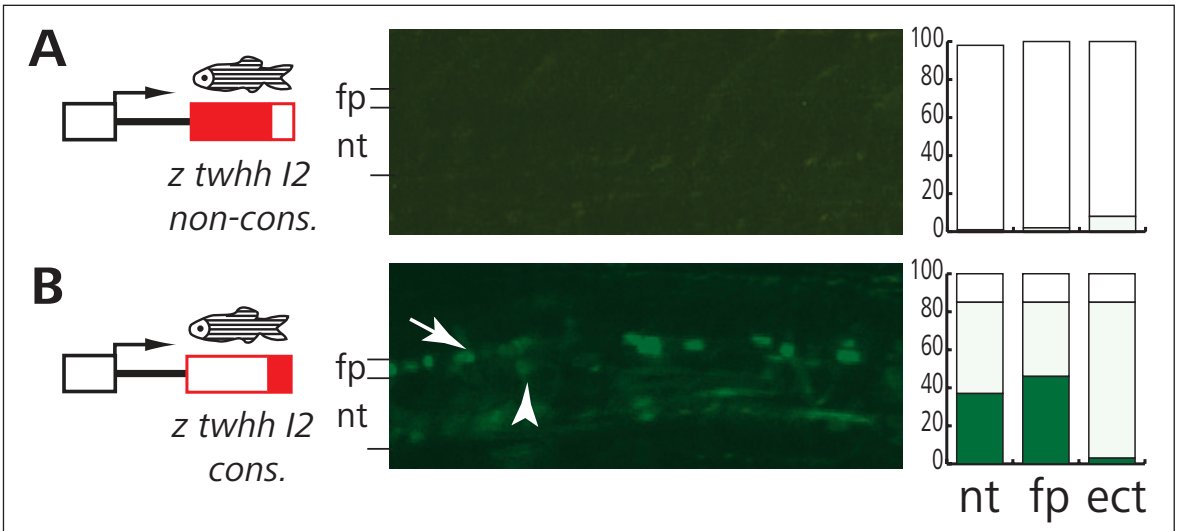

Figure 19. The *twhh ar-C* is located in the conserved part of the second intron. Embryos injected with construct containing the non-conserved part of zebrafish *twhh* intron 2 (A) and the conserved part (putative *ar-C*) (B).

Prediction of functionally relevant motifs by phylogenetic reconstruction

Transcription factor binding sites may be more conserved than the surrounding sequences (Moses et al. 2004). we have hypothesised that sequence similarity between fish and human *ar-C* sequences may reveal conserved motifs, which may reflect conserved transcription factor binding sites (Jeong and Epstein 2003). We postulated that putative transcription factor binding sites and changes in them may be detectable by identification of motifs using local alignment of *ar-C* from large number of pre and post duplicated *shh* orthologs and paralogs. To this end, a CHAOS/DIALIGN (Brudno et al. 2004) multiple alignment was used to compare the functionally active *ar-C* enhancer sequences of zebrafish, as described in (Muller et al. 1999) and equivalent sequences from all major vertebrate classes. The alignments were arranged according to phylogeny (Fig. 20).

A pattern of conserved motifs is detected in the form of homology blocks extending to 20-30 bps. These conserved motifs show distinct distribution characteristics, which reflect phylogenic as well as paralogy and orthology relationships between *shh* genes. For example two homology blocs called C1 and C3 were identified, which are present in all *shh* sequences including *twhh* paralogs in all species analysed. In contrast, two other homology blocks named C2 and C4 were detected only in *shh* genes but absent in *twhh* genes. To exclude possible rearmaments of C2 and C4 blocks, sequence alignments using other algorithms were carried out, but no evidence for reallocation of these blocks was found. Since C2 and C4 are also present in preduplicated enhancers of sarcopterygians the specific and consistent lack of C2 and C4 in *twhh* enhancers is likely due to a secondary loss of these elements after the fish specific gene duplication. This homology blocks may represent putative binding sites. Thus, the two sets of binding sites (C1/C3 and C2/C4 respectively) may be targets for TFs that regulate the differential enhancer activities of *shh* (predominantly notochord expression) and *twhh* (predominant floor plate expression). In conclusion, we have identified a set of putative targets of mutations that may contribute to the divergence of *ar-C* enhancer functions after gene duplication.

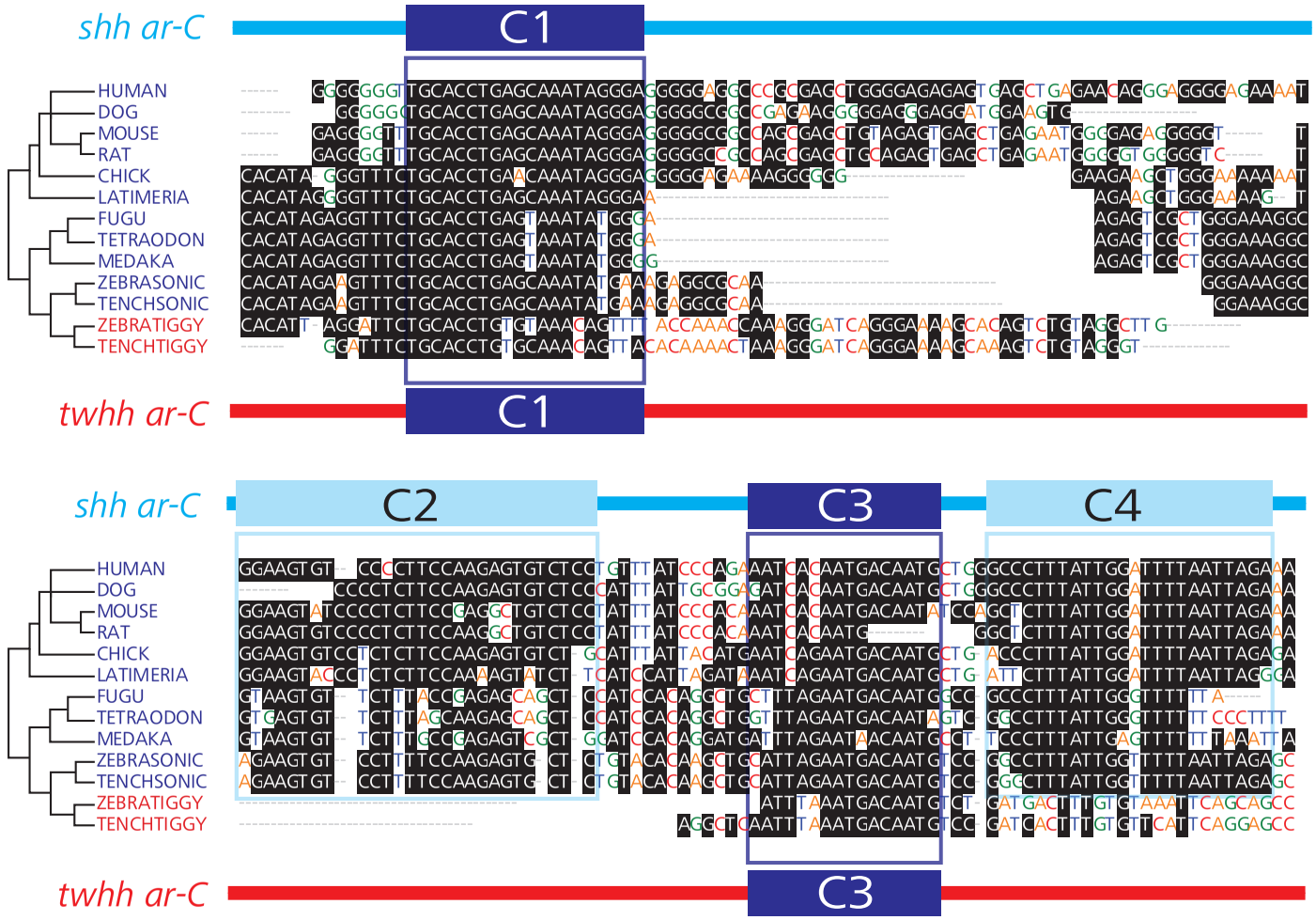

Figure 20. Sequence comparison identifies phylogeny, paralogy and orthology-specific conserved motifs in ar-C sequences. Multiple alignment of *ar-C* homolog sequences of *shh* and *twhh* genes of different vertebrate species was carried out. The phylogenetic tree on the left side represents the evolutionary relationship of the vertebrates. Species in blue correspond to *ar-C* of *shh* genes, in red to *ar-C* of *twhh* genes. Dark-blue boxes depict the conserved motifs, present in both shh and *twhh ar-C* genes. Light-blue boxes mark motifs present only in *twhh* genes.

Functional analysis of conserved motifs reveals the evolutionary changes that likely contributed to enhancer divergence of shh paralogs.

To test the functional significance of the two sets of homology blocks we conducted a systematic mutation analysis of the C1-C4 conserved homology blocks both in *shh* and *twhh* genes. Furthermore, we carried out exchange of homology blocks between *shh* and *twhh ar-C* enhancers to test if the predicted evolutionary changes after gene duplication can be modelled in a transgenic zebrafish system.

As shown in Fig. 21B-E, mutations inserted into homology blocks (C1-4) result in dramatic changes in *shh ar-C* enhancer activity. Replacement of C1 with random sequence results in total loss of *ar-C* enhancer function indicating that this binding site is critical for *shh ar-C* activity (Fig. 21B). By contrast, loss of C3 results in no observable effect suggesting that this conserved block is either not required for enhancer function or only necessary for functions that are not detectable in our transgenic system (Fig. 21D). Importantly, removal of C2 and C4 (the blocks that are only present in *shh* genes) results in strong activation of GFP in the floor plate (Fig 21C, E). In the case of C4 removal, a reduced reporter expression in the notochord has also been observed (Fig 21E). The obtained expression pattern strongly resembles the activity of the wild type *twhh ar-C* enhancer (compare Fig. 21E to Fig. 22A). Thus, removal of *shh* specific motifs from the *shh ar-C* mimics *twhh ar-C* enhancers, both in sequence composition and enhancer function. Moreover, this result is consistent with a model that the C2 and C4 elements are targets for repressors of floor plate expression in the *shh ar-C* enhancer.

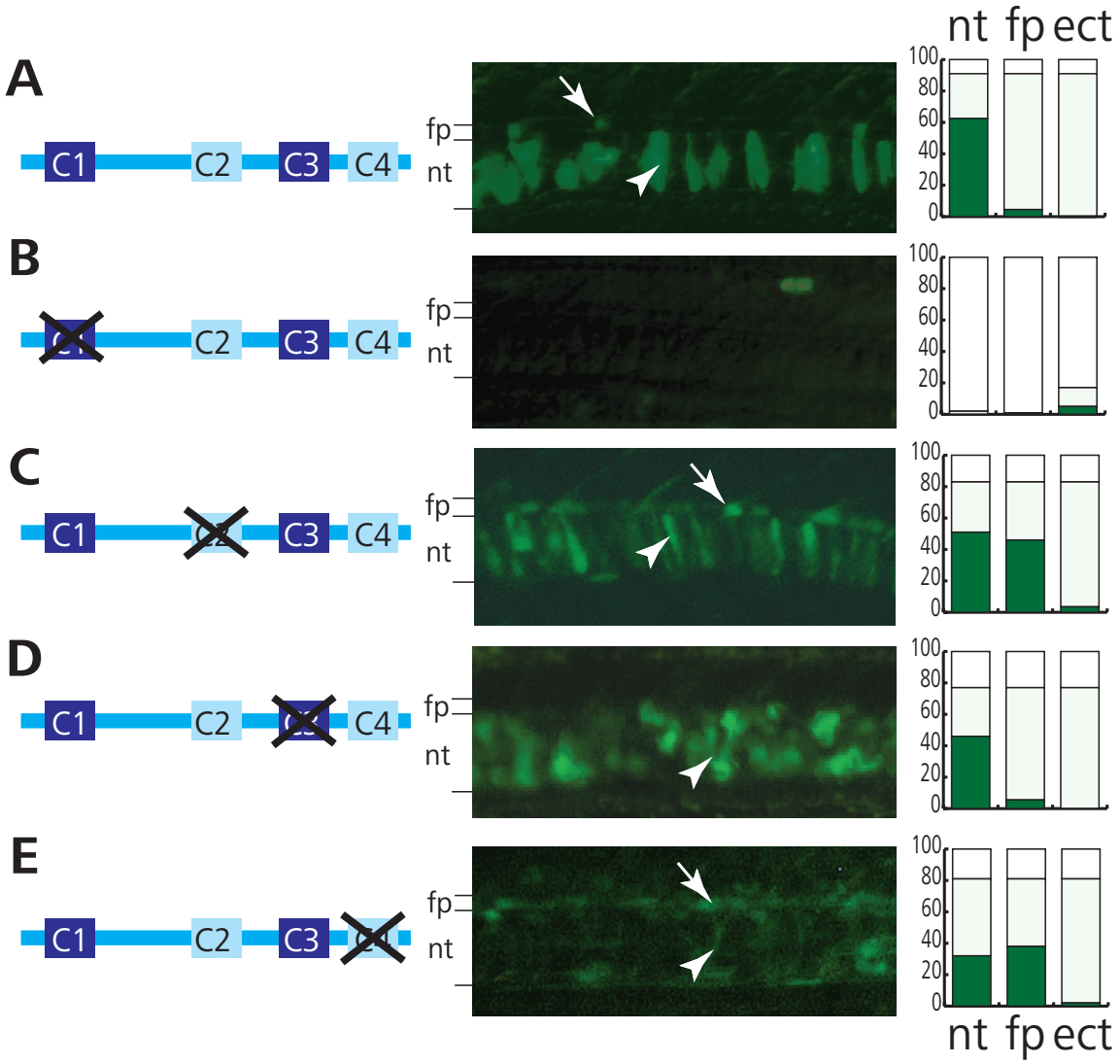

Figure 21. Transgenic analysis of *shh ar-C* motifs by mutations. Embryos injected with the corresponding constructs are shown at 24 hpf lateral view onto the trunk at the level of the midline. Embryos injected with *gfp*-reporter constructs containing wild type zebrafish *shh ar-C* (A), *ar-C* with mutated C1 region (B), mutated C2 (C), mutated C3 (D) and mutated C4 (E).

Next we asked whether *twhh ar-C* is active in the floor plate because it contains the putative midline activator site C1 and lacks the floor plate repressor elements C2 and C4 present in the *shh ar-C* enhancer. To this end, first we tested if the C1 and C3 of *twhh* are required for the function of the *twhh* enhancer. Similarly to the results obtained with *shh*, C1 was found to be critical for the general activity of *twhh ar-C* (Fig. 22B) while loss of C3 had no effect, thus mimicking the findings in *shh* (Fig. 22C). We then introduced the *shh* specific C2 or C4 into the *twhh* enhancer in order to test the functional significance of the lack of C2 and C4 motifs in *twhh*. When a *shh* derived C2 was introduced into *twhh ar-C*, no effect was observed (Fig. 22D), but introduction of the C4 putative floor plate repressor motif from *shh* did result in a dramatic shift in *twhh* enhancer activity (Fig. 22E). The effect was a repression of floor plate expression while notochord activity was retained, thus resembling the wild type or C2 mutant *shh ar-C* enhancer (Fig. 21 A,C). In a control experiment, random DNA sequence was introduced at similar positions into the *twhh ar-C* enhancer. However, this manipulation had no effect on the activity of *twhh ar-C* (Fig. 22F,G), indicating that the changes seen with the C4 insertion are due to the specific sequence of C4. These results together strongly suggest that the function of C4 is to repress floor plate activation by the *shh ar-C* enhancer. Together, these results are consistent with a model that the loss of the C4 motif in the evolution of the *twhh ar-C* has contributed to its floor plate specific activity.

All these transient transgenic analyses were a reliable substitute for the generation of stable transgenic lines as indicated by the identical results obtained with transient analysis and stable transgenic lines made for a subset of the constructs used in this study (Fig. 23)

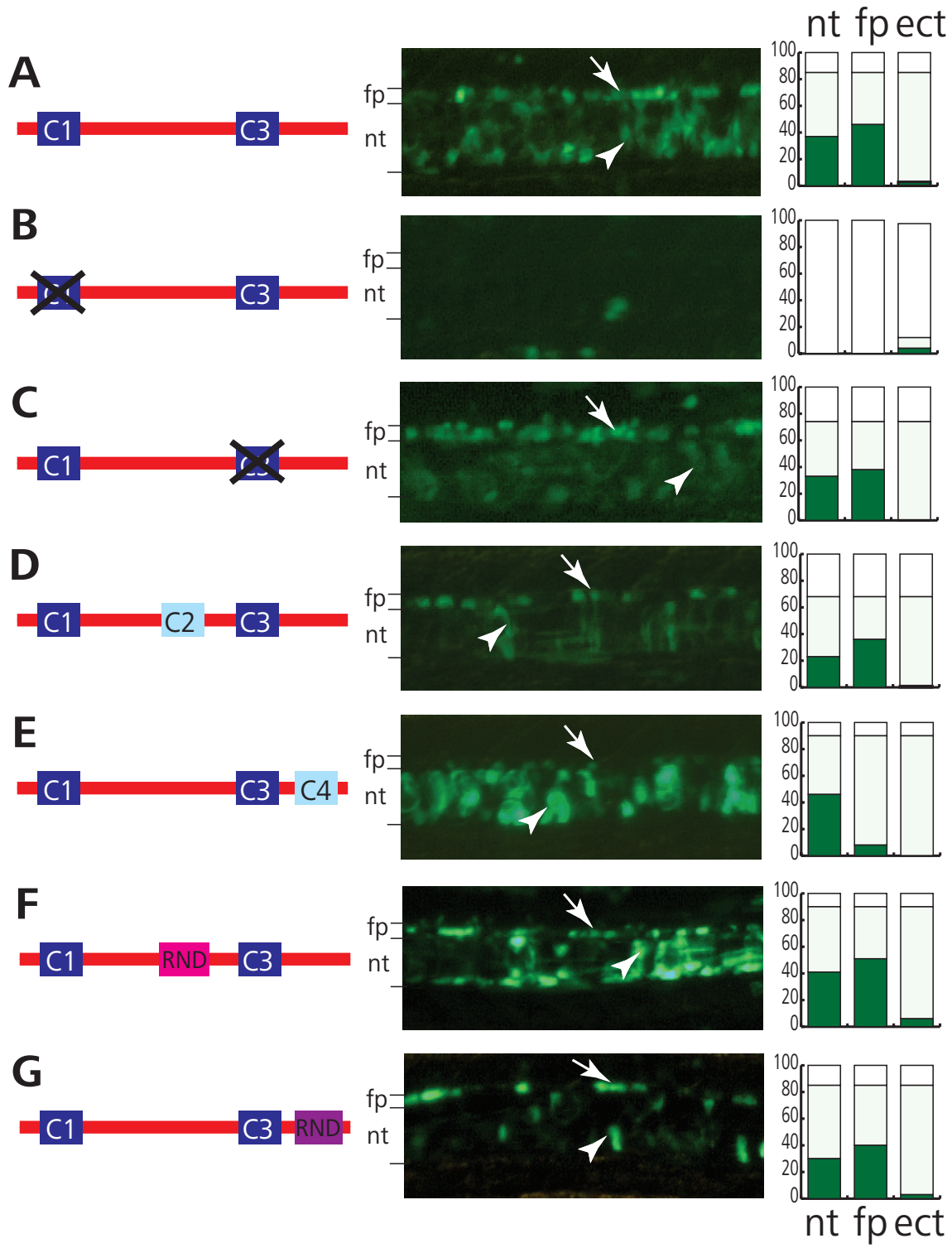

Figure 22. Transgenic analysis of *twhh ar-C* motifs. Embryos injected with *gfp* reporter constructs containing wild type zebrafish *twhh ar-C* (A), *ar-C* with mutated C1 (B), mutated C3 (C), exchange of *twhh* sequence with the zebrafish *shh* C2 (D), with the zebrafish *shh* C4 (E), with random sequence instead of *shh* C2 (F) and with random sequence instead of *shh* C4 (G)

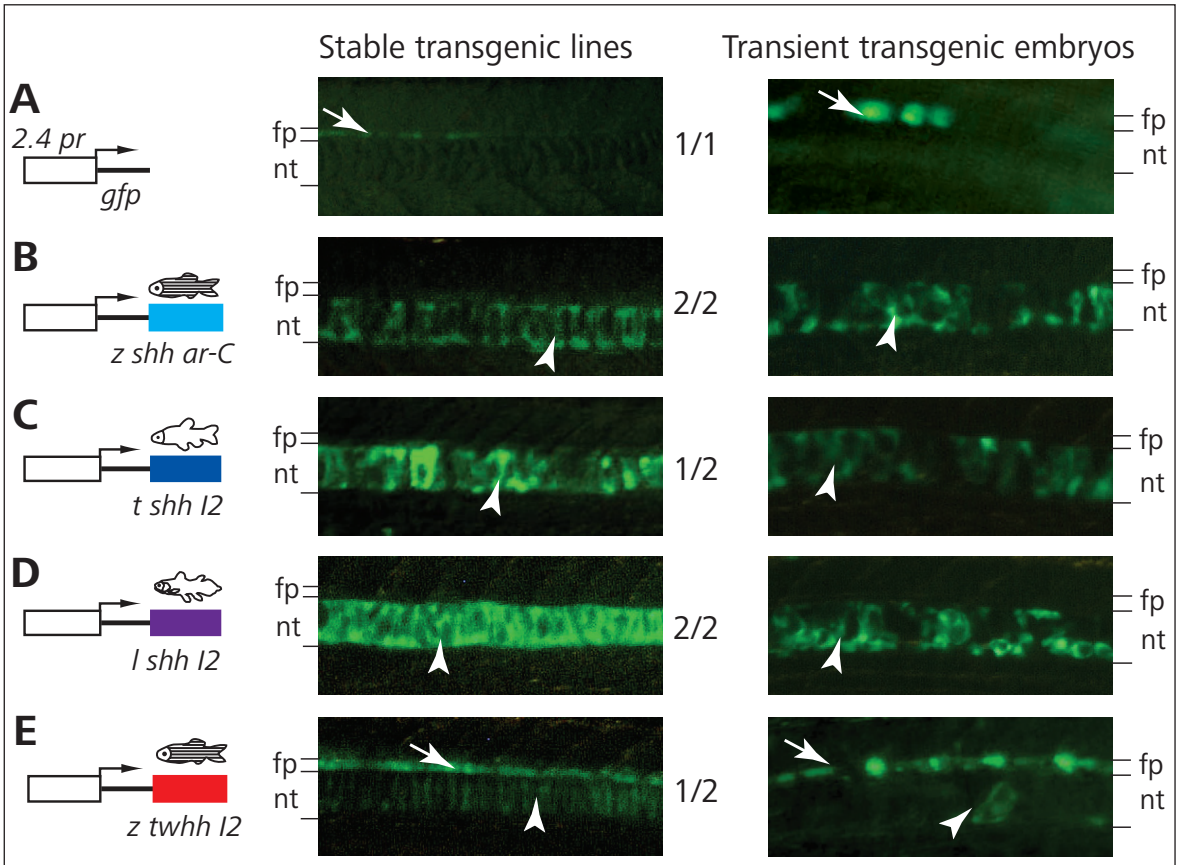

Figure 23. Comparison of the expression pattern between stable transgenic lines and transient transgenic embryos. A. Stable transgenic line (left) and transient transgenic embryos generated with *gfp* reporter construct, containing the 2.4 kb (sequence 2.4 kb upstream from the transcriptional start site) zebrafish *shh* promoter. B-E Transgenic lines and transient transgenic embryos generated with reporter constructs containing the 2.4 kb zebrafish *shh* promoter and zebrafish *ar-C* enhancer (B), tench *shh* intron 2 (C), *Latimeria shh* intron 2 (D) and zebrafish *twhh* intron 2 (E). The numbers on the right side of the images of the stable transgenic lines indicate the number of the transgenic lines showing the expression pattern/ total number of stable lines generated.

Reporter Construct	Notochord		Floor Plate		Ectopic		Non-expressing	Total number
	>15 cells	<15 cells	>15 cells	<15 cells	>15 cells	<15 cells		
0.8shh:gfp	0%	3% ±1,6	0%	2,3% ±0,9	0%	16% ±3,5	84% ±3,5	224
0.8shh:gfp:z-shh-I2	57% ±2,9	32,9% ± 5,2	3,4% ± 1,2	86,5% ±3	0%	89,9% ±3,8	10,1% ±4,7	301
0.8shh:gfp:t-shh-I2	58,8% ±3,3	27,1% ±6,7	4% ±0,7	82% ±4,6	0%	86% ±4	14% ±4,9	272
0.8shh:gfp:l-shh-I2	61,2% ±8,5	26,4% ±5,2	1,2% ±0,3	86,4% ±3,5	0%	87,6% ±3,4	12,4% ±4,2	325
0.8shh:gfp:c-shh-I2	56,1% ±7,2	28,9% ±11,5	2% ±0,1	83,1% ±4,2	0%	85% ±4,3	15% ±6,1	203
0.8shh:gfp:m-shh-I2	0%	0%	0%	0%	2.2%±1,2	7.5% ±3.1	90.3% ±4,3	250
0.8shh:gfp:z-twhh-I2	30,2% ±5,3	51,6% ±6,9	38,1% ±4,9	43,7% ±8,9	2,5% ±1	79,3% ±4,6	18,2% ±5,4	281
0.8shh:gfp:t-twhh-I2	27,9% ±7,9	50,9% ±7,9	37,8% ±5,7	41% ±6	2,1% ±0,7	76,8% ±2	21,2% ±3,3	248
0.8shh:gfp:z-twhh-I2-non.cons.	0%	1,3% ±1,3	0%	2,1% ±0,8	0%	7,7% ±2,4	92,3% ±3,5	145
0.8shh:gfp:z-twhh-arC	36,7% ±5,7	48,9% ±7,2	46% ±5,4	39,6 ±10,3	3,1% ±0,3	82,4% ±4,8	14,4% ±6,9	409
0.8shh:gfp:z-shh-arC	62,2% ±5,6	28,6% ±2,4	4,4% ±1,1	86,4% ±3,4	0%	90,8% ±3,5	9,2% ±4,3	260
*0.8shh:gfp:z-shh-arCΔ*C1	0%	2,2% ±0,6	0%	1,5% ±0,1	5,2% ±0,3	11,9% ±1	82,9% ±0,9	135
*0.8shh:gfp:z-shh-arCΔ*C2	46,2% ±4,3	31,1% ±8,8	5% ±1,3	72,2% ±3,3	0%	77,2% ±4,5	22,8% ±5,6	347
*0.8shh:gfp:z-shh-arCΔ*C3	51,2% ±3,6	30,5% ±2,6	47,1% ±4,5	34,6% ±5,7	3,7% ±1,3	78% ±1,9	18,3% ±3,7	307
*0.8shh:gfp:z-shh-arCΔ*C4	32,5% ±5,1	48,6% ±6,6	37,6% ±3,1	43,5% ±4,8	2,1% ±1,3	79,1% ±5	18,9% ±4,7	359
*0.8shh:gfp:z-twhh-arCΔ*C1	0%	0%	0%	0%	3,8% ±1,6	10,7% ±7,7	85,5% ±11,3	186
*0.8shh:gfp:z-twhh-arCΔ*C3	33,5% ±3	40,5% ±6	37,8% ±3,9	36,2% ±7,3	0%	74% ±3,5	26% ±4,3	230
0.8shh:gfp:z-twhh-arC+C2	23% ±6,2	44,6% ±8,7	36% ±5,2	31,6% ±7,8	1,3% ±1	66,3% ±3,2	32,4% ±3,2	203
0.8shh:gfp:z-twhh-arC+C4	45,7% ±7,2	43,3% ±4,7	8,2% ±2,4	80,8% ±3,8	0%	89% ±3,2	11% ±3,9	288
0.8shh:gfp:z-twhh-arC+C2rnd	40,5% ±7,7	49,1% ±6,1	50,8% ±7,4	38.8% ±6,5	5.7% ±1,5	83,9% ±5,3	10,4% ±3,8	158
0.8shh:gfp:z-twhh-arC+C4rnd	29,6% ±6,9	55,2% ±5.5	40.1% ±4,3	44,7% ±3.5	3,2% ±1,2	81.6% ±4.8	15,2% ±4.1	182
2.7twhh:gfp	72,4% ±3,1	19,6% ±3,3	0%	92% ±3,4	0%	92% ±3,4	8% ±4,2	308
2.7twhh:gfp:z-twhhI1	68% ±4,9	19,8% ±0,8	0%	87,8% ±4,2	0%	87,8% ±4,2	12,2% ±5,1	339
2.7twhh:gfp:z-twhhI2	61,4% ±4,9	24,7% ±2,9	36,4% ±3,6	49,7% ±3,1	2% ±0,8	84,1% ±5,6	13,9% ±7,7	296

Table 3. Quantification of the *gfp* expression for each reporter construct

Discussion

It has long been suggested (Carroll 2000; King and Wilson 1975) that a major driving force of evolution of animal shape is the divergence of cis-regulatory elements of genes. Recent years provided evidence in support of this hypothesis (Gompel et al. 2005; Hughes et al. 2006; Jeong et al. 2006; Prud'homme et al. 2006; Wittkopp et al. 2004; Wittkopp et al. 2002). However, the mechanisms of cis-regulatory evolution are still poorly understood (Ludwig et al. 2000; Ludwig et al. 2005; Ludwig et al. 1998; Wittkopp 2006). In this thesis, we have systematically analysed the evolutionary history of a single enhancer of orthologous and paralogous *shh* genes during vertebrate phylogeny. By construction of multiple alignments, we were able to predict which motifs within the *ar-C* enhancer represent regulatory input. Through specific mutations and exchanges of motifs, we mimicked likely evolutionary events in transgenic analysis and identified the lineage specific modifications leading to discernible changes in tissue specific enhancer activity in embryo development.

Identification and functional verification of a diverged ar-C enhancer

Using phylogenetic footprinting of intron 2 of *twhh* genes we have identified a conserved *ar-C* homolog enhancer in two species of cyprinids. Our results by transgenic analysis indicate that the *ar-C* sequences in intron 2 together with the promoter activity of *twhh* (Du and Dienhart 2001) contribute to this gene's activity in the floor plate. While *shh* enhancers retained significant sequence similarity with their orthologs the whole of the *twhh* gene and its *ar-C* enhancer is grossly changed from that of *shh* paralogs. This paralog specific change happened despite the fact that *twhh* had equal time and chance to diverge as *shh* after duplication from an ancestral *sonic hedgehog* gene. This result is in line with observations from several reports. For example Zerucha et al. (Zerucha et al. 2000), studying the cis-regulatory elements of the *dlx* genes in zebrafish, have observed that the non-coding sequences form orthologous *dlx* genes are much more conserved than those of the paralogous genes, result very similar to the case of *shh* and *twhh*. A similar conclusion has been

achieved by McEwen and co-workers (McEwen et al. 2006) in a global comparative study of cis-regulatory elements of duplicated genes from vertebrates. Our results, together with the above reports provide experimental support to the notion that differential divergence of non-coding conserved elements of paralogs may be a general phenomenon in vertebrates. Such selective accumulation of mutations in one of the duplicated paralogs are more in line with the proposed classical model for evolution of duplicated genes, e. g. nonfunctionalisation, when deleterious mutations accumulated only in one of the copies resulting in its lost during evolution. In fact, this is most likely to have happened with the *twhh* in Acantopterygian lineage (medaka, fugu and tetraodon) (Fig. 10). Genomic analysis has failed so far to detect *twhh* gene in those species (Fig. 15) and although our results indicate subfunctionalisation between *shh* paralogs in zebrafish, we can not exclude that this is just an intermediate evolutionary state *twhh* in on way to be lost in zebrafish as well. However the high degree of sequence divergence between paralogs is unlikely to cause nonfunctionalisation in general, this is inconsistent with the existence of high number functional paralogs over million of years, and for many of them subfunctionalisation on spatial and temporal expression domains has been shown. Thus, the evolutinary mechanisms behind the observed selective diverges of paralogs, remain to be understood.

Identification of putative transcription factor binding sites by local alignment of multiple species and phylogenetic reconstruction of enhancer divergence.

The use of a local sequence alignment approach of representative species of major vertebrate lineages allowed us to predict functionally relevant motifs within the *ar-C* enhancers. Our findings are most consistent with a model that these motifs are individual or multimeric transcription factor binding sites. Since mutation and transgenic analysis verified the functional relevance of these motifs in driving expression in the midline, thus the most parsimonious explanation for the conservation of these sequence elements is that they represent functional binding sites for

developmental regulator transcription factors.

The *ar-C* enhancer is composed of motifs with different regulatory capacity (Fig. 24). Motifs exist that are crucial for the overall activity of the enhancer (C1) functioning as midline activator which can activate expression in both floor plate and notochord. Other motifs (C2 and C4) refine the enhancer activity by repressing the floor plate expression. This indicates that the overall activity output of an enhancer in midline tissues is subject to both activator and repressor functions acting in concert. These results agree with the previously proposed grammar of developmentally regulated gene expression (Falb and Maniatis 1992; Gompel et al. 2005; Gray et al. 1994; Howard and Davidson 2004; Lemon and Tjian 2000; Levine and Davidson 2005; Minokawa et al. 2005).

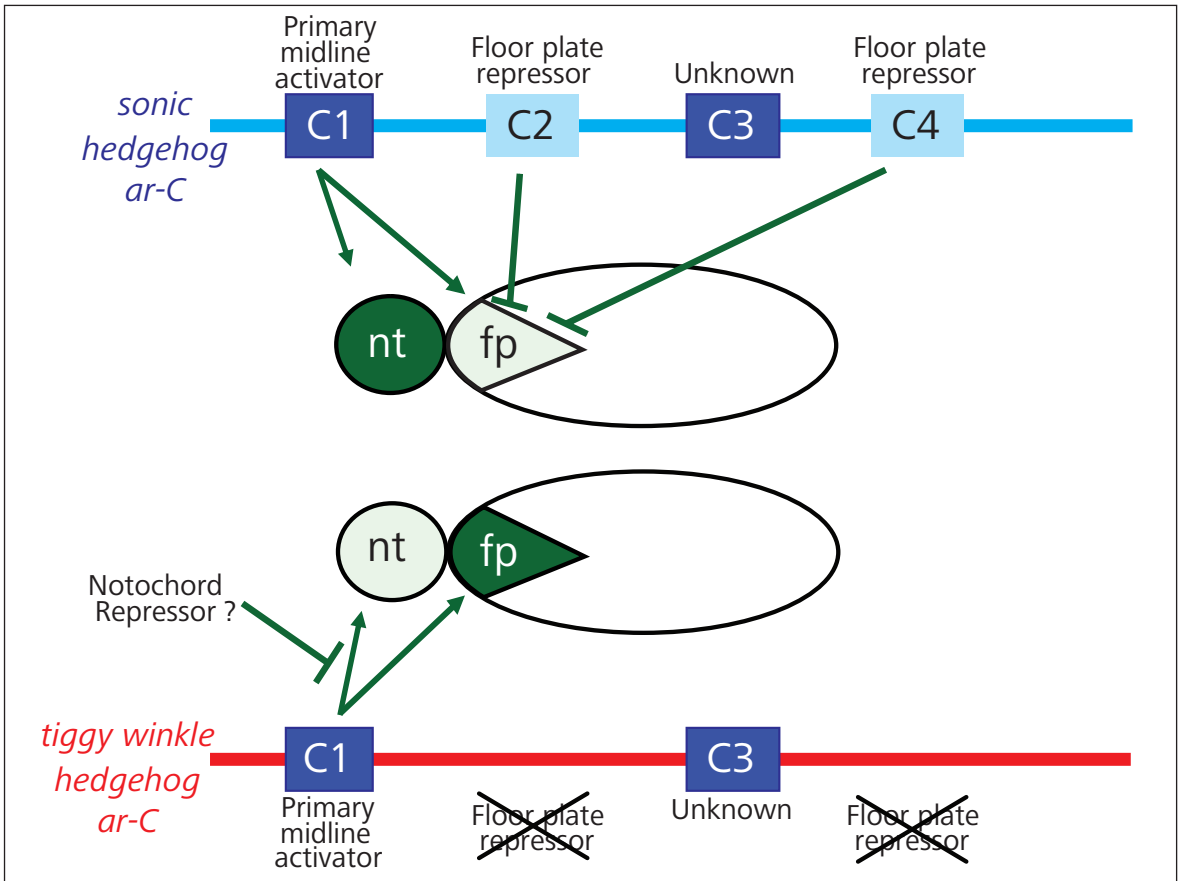

Figure 24. Model for motif structure and interaction in *ar-C* enhancers involved in the regulation of midline expression and functional divergence of *ar-C* in *sonic* and *tiggy winkle hedgehogs* in zebrafish. Schemes on the top and bottom represent the structure of the *ar-C* enhancer of shh (blue) and *twhh* (red) with the position of the conserved motifs indicated in coloured boxes as in Fig. 4 and 5. In the middle schematic cross-sections of the neural tube with the floor plate (fp) and the notochord (nt) are shown (ventral to the left). Dark green indicates strong enhancer activity. Arrows indicate activator, blunt arrows indicate repression function by individual motifs.

An open question remains however, why the *ar-C* enhancer of *shh* should be repressed in the floor plate while the *shh* gene is well known to be active in this tissue? The level of the Hedgehog morphogen signal emanating from the floor plate is critical for correct patterning of the ventral neural tube (Roelink et al. 1995). Animals with only one gene encoding the Sonic hedgehog protein (sarcopterygians and fishes without *twhh*) achieve this by controlled activation of *shh* in the floor plate as a result of a combination of several synergistic enhancers (Epstein et al. 1999; Ertzer et al. 2007). In species with a second copy of *shh* paralog contributing to Shh production only one of the paralog floor plate enhancers may be subjected to selection pressure. For example, to counter the overproduction of Hedgehog levels the reduction of transcription can occur by blocking the activity of one of synergistically active enhancers (in this case *ar-C*). It is important to note however, that the *shh ar-C* enhancers are not exclusively expressed in the notochord and retained a weaker but still noticeable, capacity to activate expression in the floor plate. This function seem to be critical for the synergistic initiation of *shh* floor plate expression together with *ar-B* in early developmental stages (Ertzer et al. 2007). Thus, the output of Shh/Twhh levels appears to be a subject of quantitative regulation of paralog enhancer activities. Alternatively, it is feasible that there are time points when the two genes are not overlapping in expression and the complementing specificities of *twhh* and *shh ar-C* enhancers reflect the non-overlapping production of Hedgehog proteins in the two midline tissues (Etheridge et al. 2001).

Importantly, the order and combination of motifs of *ar-C* are conserved. This is a very different result from that proposed for the stripe 2 enhancers of drosophilids where the functional conservation of cis-regulatory modules was a result of stabilising selection of reshuffled transcription factor binding site composition (Ludwig et al. 2000; Ludwig et al. 1998). The evolutionary pressure to keep the order and composition of binding sites within enhancers may be limited to transcription factor and developmental regulatory genes (Bejerano et al. 2004; Plessy et al. 2005). The high conservation level, however, may be a consequence of selective pressure

acting on a secondary function of enhancer sequences such as conserved non-coding RNAs which can be involved in transcriptional regulation, as shown for the *dlx* genes (Feng et al. 2006).

Previously, individual binding sites were identified through comparative approaches in vertebrates e.g. (Bejder and Hall 2002; Jeong and Epstein 2003; Shashikant et al. 1998). These examples, together with our systematic analysis of conserved motifs in the *ar-C* enhancers demonstrate that functionally relevant motifs detected by sequence alignment may aid in identifying yet unknown and uncharacterised functional transcription factor binding sites

The use of large numbers of species spanning long evolutionary distance allowed us to generate a phylogenetic reconstruction of enhancer divergence before and after gene duplication (Fig. 25). By generating artificial enhancers with mutations that mimic the predicted lineage specific changes in motif composition of *twhh* and *shh* enhancers we were able to reconstruct the likely evolutionary events leading to divergence of the *ar-C* enhancer function. For example, insertion of the floor plate repressor C4 element into *twhh* resulted in enhancer activity reminiscent of *shh ar-C* in which the C4 site had been identified. These results argue that the very changes resulting in the divergence of the enhancer function have been identified.

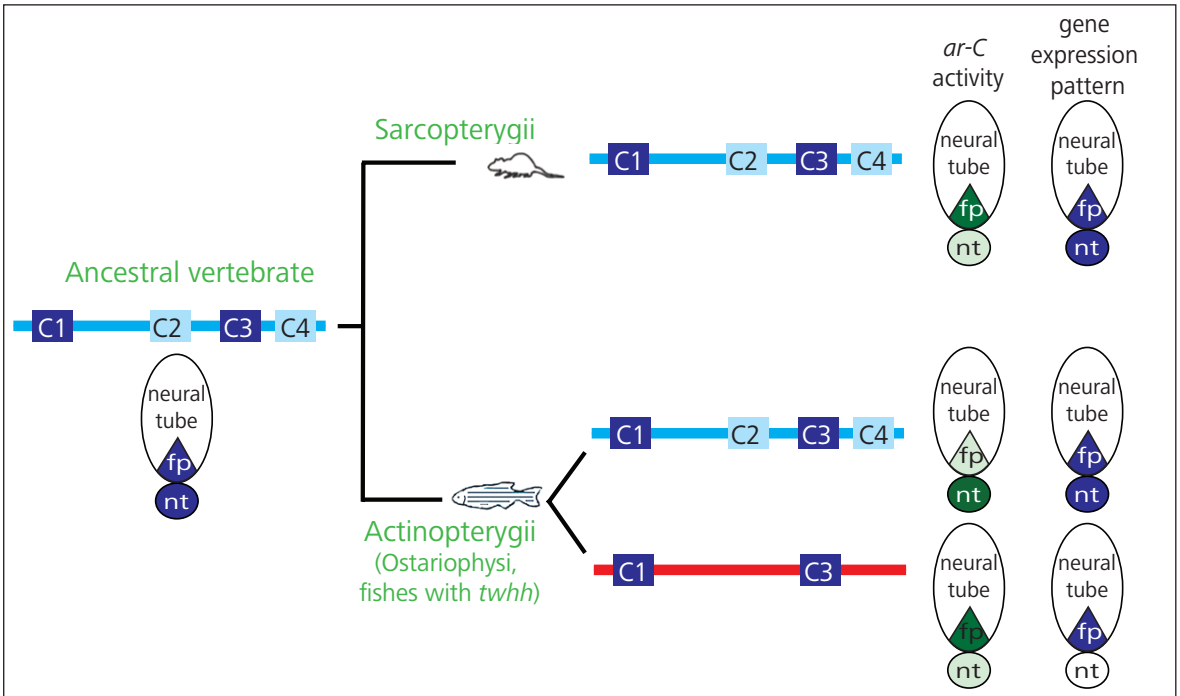

Figure 25. Phylogenetic relationship of the genes and the motif composition of the respective *ar-C* enhancers are shown. In blue shh gene enhancers in red twhh gene enhancers are shown. On the left, a predicted pre-duplicated ancestral *shh ar-C* enhancer is shown. Below, the predicted activity of the ancestral *shh* gene is depicted in blue in a schematic cross section of an embryonic midline. On the right, schematic cross sections of midlines in green indicate *ar-C/SFPE2* enhancer activities; shades of green indicate strength of enhancer activity in the respective midline tissues. In blue the expression activity of the respective *shh/twhh* genes are shown.

Subfunctionalisation by fission or binary switch in midline-specificity of enhancers during evolution

One of the main goals of our study was to better understand the diversity of subfunctionalisation mechanisms that may act on paralog enhancers during evolution. One of the main questions we asked was which are the evolvable units, involved in the process of subfunctionalisation? Because of the random nature of DNA mutation, this process can occur on the level of whole CRMs (enhancers) or on level of transcription factor binding sites within one enhancer (Fig 26). In this study we were able to show that the subfunctionalisation between *shh* and *twhh* in zebrafish happen predominantly on the level of putative binding sites within a single enhancer (*ar-C*). We propose that the presence or absence of the C4 site functions as a binary switch to modulate *ar-C* enhancer activity specific to one of two midline tissues after gene duplication. By selective removal of repressor and activator bind-

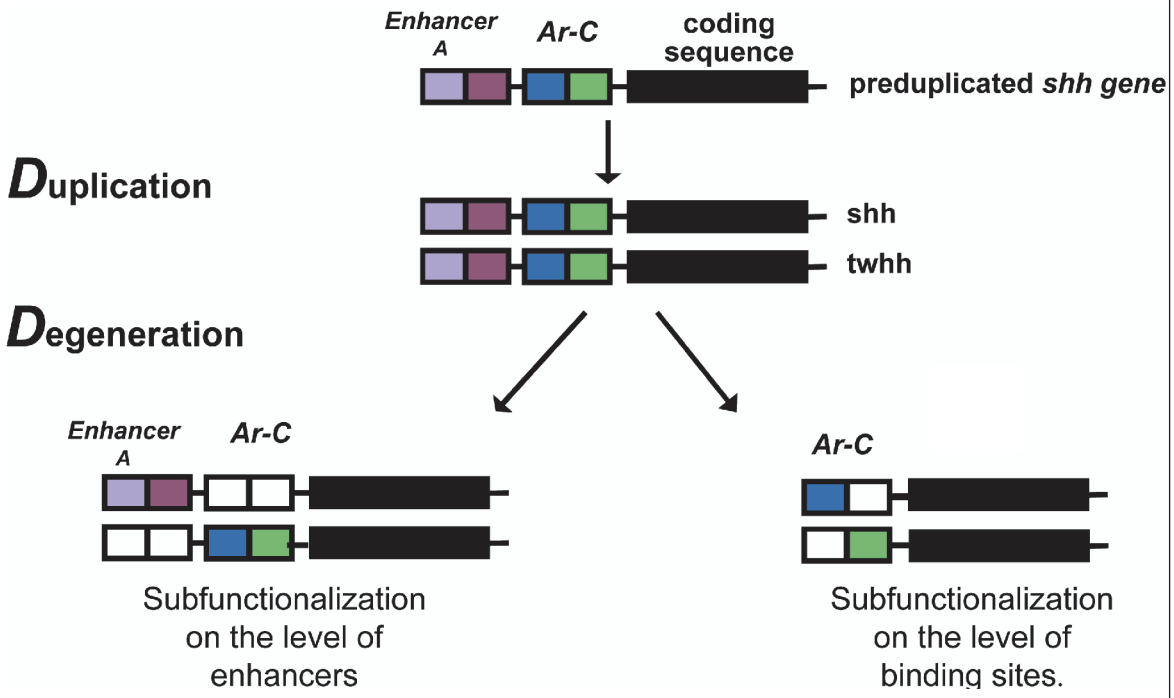

Figure 26. Schematic representations of the possible subfunctionalisation mechanisms of cis-regulatory modules in *shh/twhh* genes in zebrafish, according the DDC model. The left panel shows subfunctionalisation on the level of enhancers. The right subfunctionalisation on level of binding sites within one enhancer (*ar-C*), which have been confirmed by our results. The green blue purpule nad dark pink boxes indicate individual binding sites. The black boxes coding sequence. Modified from Force et al. 1999.

ing sites the subfunctionalisation of the *ar-C* enhancer to floor plate or notochord can thus occur (Fig 24). However the persistent notochord activity of the *twhh ar-C* and other elements of analysed *twhh* genomic region (Fig. 17, 18, 23) leads us to propose the involvement of another element in the subfunctionalisation, outside of these regions, needed to repress the expression of *twhh* in the notochord. Thus, our results do not provide exclusive support to one or the other mechanism underling the subfunctionalisation of cis-regulatory modules in global. This is in-line with two recent reports that demonstrate experimental verification of subfunctionalisation of Hox gene enhancers (Tumpel et al. 2006; Tvrdik and Capecchi 2006). The first report by Tvrdik et al. describes a subfunctionalisation process on level of whole cis-regulatory modules (enhancers). On the other hand Trumpel and co-workers show that such subfunctionalisation can happen on the level of binding sites within one enhancer, which is very similar to our finding (Fig. 26). However, we were able

to demonstrate that such changes can be predicted by comparative genomics (phylogenetic footprinting) without previous, experimentally confirmed knowledge of the exact transcription factor binding site composition of the enhancer in question. This is in contrast to their findings, based mainly on comparison of already known transcription factor binding sites.

The subfunctionalisation model would argue for the existence of a pre-duplication (sarcopterygian) *ar-C* enhancer that is equally active in both the floor plate and notochord. However, in fish all *shh ar-C* enhancers from sarcopterygian lineages show notochord specific enhancer activity and even more strikingly, the mouse *SFPE2* (the homolog of zebrafish *ar-C*) has been inactive in fish. These differences may be explained both by subfunctionalisation mechanisms as well as by trans-acting factor changes, like changes of the expression domains or binding site specificity of the transcription factors involved in the regulation of *shh,* between zebrafish and mouse. In support of trans-changes the mouse *SFPE2* enhancer is mainly active in the floor plate of the mouse (Epstein et al. 1999) and can activate notochord expression in a multimerised form (Jeong and Epstein 2003) (Fig. 7, 8). In addition, the mainly notochord specific zebrafish *ar-C* showed both floor pate and notochord activity in the mouse (Muller et al. 1999). However, the lack of enhancer activity of *SFPE2* in zebrafish can be a result not only of trans-regulatory change but a mouse-specific sequence divergence as well. As can be seen on Fig. 12 the mouse *SFPE2* is the most diverged on sequence level form the zebrafish *ar-C*, in comparison to the other sarcopterygians (*Latimeria*, chicken and human). Thus the subfunctionalisation of duplicated *ar-C shh* enhancers is a composite result of selective loss of several motifs including negative regulatory elements in one enhancer (*twhh*) paralleled by modifications either on the cis or on the trans level to restrict the activity of the less diverged sister paralog enhancer (*shh*). The prediction from this model is that fish species without *twhh* gene (acantopterygii) may have floor plate active *ar-C* enhancer. Interestingly, the floor plate repressor elements (C2/C4) of *shh ar-C* of acanthopterygians (e.g. medaka, fugu) are present but diverged from all other *shh* homologs

(Fig. 27), and may thus represent the evolutionary changes leading to retention of *shh ar-C* floor plate activity in these fish lineages, a hypothesis, which confirmation will be a subject of future studies.

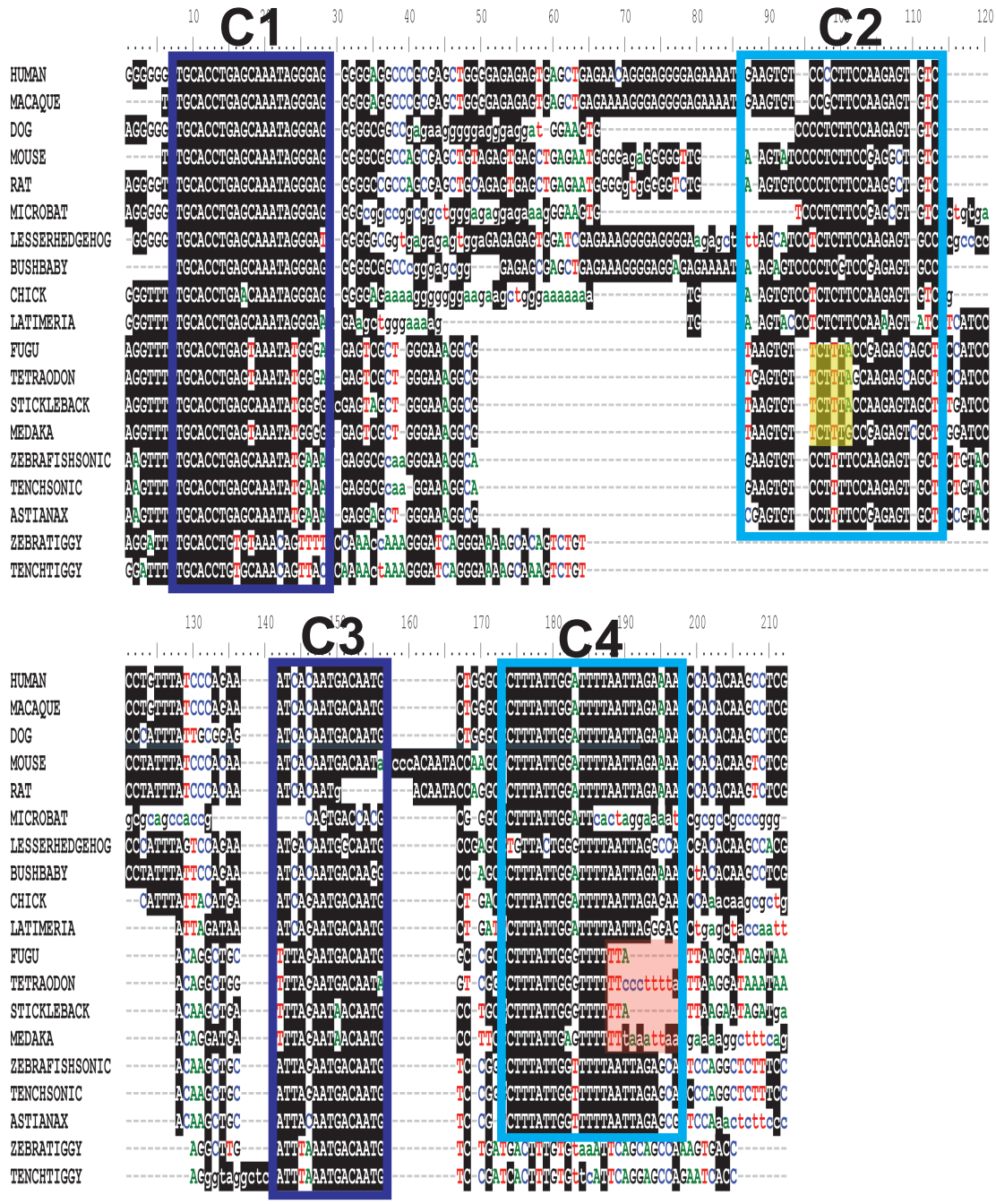

Figure 27 Specific divergence of homology block C2 and C4 (the blocks involved in the repression of the floor plate activity of the zebrafish *shh*) between acanthopterygians (fishes without *twhh* gene) and ostariophysians (fishes with *twhh* gene). The sequence differences are marked by coloured boxes, yellow for C2 and red for C4.

The combination of both negative and positive regulatory sites within a single enhancer indicates the integration of activating and repressing signals to modulate the resulting transcriptional activity. This could be achieved through multiple trans-acting factors that interact with a series of binding sites within the *ar-C* enhancer. Which transcription factors bind to the C1-C4 blocks will remain a challenge for future research. Predictions can be made based on known transcription factor recognition sequences. For instance, C1 contains a *foxa2* binding sequence which is consistent with the previously suggested role of this factor in regulating *shh* gene expression in the midline of mouse (Ang and Rossant 1994; Jeong and Epstein 2003) frog (Ruiz i Altaba 1994) and fish (Chang et al. 1997). Interestingly, C4 carries a sequence identical to the homeobox binding site that has been described to be present in the mouse *SFPE2* enhancer (Jeong and Epstein 2003). This binding site is required for floor plate activity in the mouse. The identity of the mouse binding factor and whether the same transcription factor acts (by probably repressing floor plate activity) in the *ar-C* enhancer in zebrafish, is yet unknown. The relevance of specific transcription factors from large protein families in binding to the *ar-C* binding sites remains a challenging question.

In conclusion, we were able to predict the motifs that are required for the tissue-specific activity of the paralog enhancers by phylogenetic reconstruction and we identified the putative transcription factor binding sites that were the likely targets of evolutionary changes underlying the functional divergence of the two *ar-C* enhancers of the *shh* paralogs. By engineering and exchanging mutations in both of the two enhancers of *shh* and *twhh* followed by transgenic analysis of the mutated enhancers we were able to recapitulate the predicted evolutionary events and thus provide evidence for the likely mechanism of enhancer evolution after gene duplication.

References

Alberts, B. 2002. *Molecular biology of the cell*. Garland Science, New York.

Alcedo, J., M. Ayzenzon, T. Von Ohlen, M. Noll, and J.E. Hooper. 1996. The Drosophila smoothened gene encodes a seven-pass membrane protein, a putative receptor for the hedgehog signal. *Cell* **86:** 221-232.

Alexandre, C., A. Jacinto, and P.W. Ingham. 1996. Transcriptional activation of hedgehog target genes in Drosophila is mediated directly by the cubitus interruptus protein, a member of the GLI family of zinc finger DNA-binding proteins. *Genes Dev* **10:** 2003-2013.

Alkema, W.B., O. Johansson, J. Lagergren, and W.W. Wasserman. 2004. MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res* **32:** W195-198.

Amores, A., A. Force, Y.L. Yan, L. Joly, C. Amemiya, A. Fritz, R.K. Ho, J. Langeland, V. Prince, Y.L. Wang et al. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* **282:** 1711-1714.

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

Amsterdam, A. and T.S. Becker. 2005. Transgenes as screening tools to probe and manipulate the zebrafish genome. *Dev Dyn* **234:** 255-268.

Amsterdam, A. and N. Hopkins. 2004. Retroviral-mediated insertional mutagenesis in zebrafish. *Methods Cell Biol* **77:** 3-20.

Ang, S.L. and J. Rossant. 1994. HNF-3 beta is essential for node and notochord formation in mouse development. *Cell* **78:** 561-574.

Arbeitman, M.N., E.E. Furlong, F. Imam, E. Johnson, B.H. Null, B.S. Baker, M.A. Krasnow, M.P. Scott, R.W. Davis, and K.P. White. 2002. Gene expression during the life cycle of Drosophila melanogaster. *Science* **297:** 2270-2275.

Arnone, M.I. and E.H. Davidson. 1997. The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124:** 1851-1864.

Avaron, F., L. Hoffman, D. Guay, and M.A. Akimenko. 2006. Characterization of two new zebrafish members of the hedgehog family: atypical expression of a zebrafish indian hedgehog gene in skeletal elements of both endochondral and dermal origins. *Dev Dyn* **235:** 478-489.

Bai, C.B., W. Auerbach, J.S. Lee, D. Stephen, and A.L. Joyner. 2002. Gli2, but not Gli1, is required for initial Shh signaling and ectopic activation of the Shh pathway. *Development* **129:** 4753-4761.

Bai, C.B., D. Stephen, and A.L. Joyner. 2004. All mouse ventral spinal cord patterning by hedgehog is Gli dependent and involves an activator function of Gli3. *Dev Cell* **6:** 103-115.

Bale, A.E. and K.P. Yu. 2001. The hedgehog pathway and basal cell carcinomas. *Hum Mol Genet* **10:** 757-762.

Barth, K.A. and S.W. Wilson. 1995. Expression of zebrafish nk2.2 is influenced by sonic hedgehog/vertebrate hedgehog-1 and demarcates a zone of neuronal differentiation in the embryonic forebrain. *Development* **121:** 1755-1768.

Beddington, R.S. 1994. Induction of a second neural axis by the mouse node. *Development* **120:** 613-620.

Bejder, L. and B.K. Hall. 2002. Limbs in whales and limblessness in other vertebrates: mechanisms of evolutionary and developmental transformation and loss. *Evol Dev* **4:** 445-458.

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

Belting, H.G., C.S. Shashikant, and F.H. Ruddle. 1998. Modification of expression and cis-regulation of Hoxc8 in the evolution of diverged axial morphology. *Proc Natl Acad Sci U S A* **95:** 2355-2360.

Berezikov, E., V. Guryev, R.H. Plasterk, and E. Cuppen. 2004. CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res* **14:** 170-178.

Berman, B.P., B.D. Pfeiffer, T.R. Laverty, S.L. Salzberg, G.M. Rubin, M.B. Eisen, and S.E. Celniker. 2004. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biol* **5:** R61.

Berthelsen, J., V. Zappavigna, E. Ferretti, F. Mavilio, and F. Blasi. 1998. The novel homeoprotein Prep1 modulates Pbx-Hox protein cooperativity. *Embo J* **17:** 1434-1445.

Bingham, S., A. Nasevicius, S.C. Ekker, and A. Chandrasekhar. 2001. Sonic hedgehog and tiggy-winkle hedgehog cooperatively induce zebrafish branchiomotor neurons. *Genesis* **30:** 170-174.

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

Briscoe, J., Y. Chen, T.M. Jessell, and G. Struhl. 2001. A hedgehog-insensitive form of patched provides evidence for direct long-range morphogen activity of sonic hedgehog in the neural tube. *Mol Cell* **7:** 1279-1291.

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

Brudno, M., R. Steinkamp, and B. Morgenstern. 2004. The CHAOS/DIALIGN WWW server for multiple alignment of genomic sequences. *Nucleic Acids Res* **32:** W41-44.

Burgess-Beusse, B., C. Farrell, M. Gaszner, M. Litt, V. Mutskov, F. Recillas-Targa, M. Simpson, A. West, and G. Felsenfeld. 2002. The insulation of genes from external enhancers and silencing chromatin. *Proc Natl Acad Sci U S A* **99 Suppl 4:** 16433-16437.

Burke, R., D. Nellen, M. Bellotto, E. Hafen, K.A. Senti, B.J. Dickson, and K. Basler. 1999. Dispatched, a novel sterol-sensing domain protein dedicated to the release of cholesterol-modified hedgehog from signaling cells. *Cell* **99:** 803-815.

Butler, J.E. and J.T. Kadonaga. 2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev* **16:** 2583-2592.

Carroll, S.B. 2000. Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101:** 577-580.

Carroll, S.B. 2005. Evolution at two levels: on genes and form. *PLoS Biol* **3:** e245.

Carroll, S.B., J.K. Grenier, and S.D. Weatherbee. 2005. *From DNA to diversity : molecular genetics and the evolution of animal design*. Blackwell Pub., Malden, Mass. ; Oxford.

Castillo-Davis, C.I., D.L. Hartl, and G. Achaz. 2004. cis-Regulatory and protein evolution in orthologous and duplicate genes. *Genome Res* **14:** 1530-1536.

Catala, M., M.A. Teillet, E.M. De Robertis, and M.L. Le Douarin. 1996. A spinal cord fate map in the avian embryo: while regressing, Hensen's node lays down the notochord and floor plate thus joining the spinal cord lateral walls. *Development* **122:** 2599-2610.

Cawley, S., S. Bekiranov, H.H. Ng, P. Kapranov, E.A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A.J. Williams et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116:** 499-509.

Chandrasekhar, A., J.T. Warren, Jr., K. Takahashi, H.E. Schauerte, F.J. van Eeden, P. Haffter, and J.Y. Kuwada. 1998. Role of sonic hedgehog in branchiomotor neuron induction in zebrafish. *Mech Dev* **76:** 101-115.

Chang, B.E., P. Blader, N. Fischer, P.W. Ingham, and U. Strahle. 1997. Axial (HNF3beta) and retinoic acid receptors are regulators of the zebrafish sonic hedgehog promoter. *Embo J* **16:** 3955-3964.

Charrier, J.B., F. Lapointe, N.M. Le Douarin, and M.A. Teillet. 2002. Dual origin of the floor plate in the avian embryo. *Development* **129:** 4785-4796.

Cheng, A.M., B. Thisse, C. Thisse, and C.V. Wright. 2000. The lefty-related factor Xatv acts as a feedback inhibitor of nodal signaling in mesoderm induction and L-R axis development in xenopus. *Development* **127:** 1049-1061.

Chiang, C., Y. Litingtung, E. Lee, K.E. Young, J.L. Corden, H. Westphal, and P.A. Beachy. 1996. Cyclopia and defective axial patterning in mice lacking Sonic hedgehog gene function. *Nature* **383:** 407-413.

Chiu, C.H., C. Amemiya, K. Dewar, C.B. Kim, F.H. Ruddle, and G.P. Wagner. 2002. Molecular evolution of the HoxA cluster in the three major gnathostome lineages. *Proc Natl Acad Sci U S A* **99:** 5492-5497.

Choo, Y. and A. Klug. 1997. Physical basis of a protein-DNA recognition code. *Curr Opin Struct Biol* **7:** 117-125.

Christ, B., R. Huang, and M. Scaal. 2004. Formation and differentiation of the avian sclerotome. *Anat Embryol (Berl)* **208:** 333-350.

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

Chuang, P.T. and A.P. McMahon. 1999. Vertebrate Hedgehog signalling modulated by induction of a Hedgehog-binding protein. *Nature* **397:** 617-621.

Cleaver, O. and P.A. Krieg. 2001. Notochord patterning of the endoderm. *Dev Biol* **234:** 1-12.

Colamarino, S.A. and M. Tessier-Lavigne. 1995. The role of the floor plate in axon guidance. *Annu Rev Neurosci* **18:** 497-529.

Cooke, J., M.A. Nowak, M. Boerlijst, and J. Maynard-Smith. 1997. Evolutionary origins and maintenance of redundant gene expression during metazoan development. *Trends Genet* **13:** 360-364.

Dale, K., N. Sattar, J. Heemskerk, J.D. Clarke, M. Placzek, and J. Dodd. 1999. Differential patterning of ventral midline cells by axial mesoderm is regulated by BMP7 and chordin. *Development* **126:** 397-408.

Damen, W.G. and D. Tautz. 1998. A Hox class 3 orthologue from the spider Cupiennius salei is expressed in a Hox-gene-like fashion. *Dev Genes Evol* **208:** 586-590.

Danke, J., T. Miyake, T. Powers, J. Schein, H. Shin, I. Bosdet, M. Erdmann, R. Caldwell, and C.T. Amemiya. 2004. Genome resource for the Indonesian coelacanth, Latimeria menadoensis. *J Exp Zoolog A Comp Exp Biol* **301:** 228-234.

Danos, M.C. and H.J. Yost. 1995. Linkage of cardiac left-right asymmetry and dorsal-anterior development in Xenopus. *Development* **121:** 1467-1474.

Davidson, E.H. 2001. *Genomic regulatory systems : development and evolution*. Academic Press, San Diego ; London.

Dickmeis, T. and F. Muller. 2005. The identification and functional characterisation of conserved regulatory elements in developmental genes. *Brief Funct Genomic Proteomic* **3:** 332-350.

Dickmeis, T., C. Plessy, S. Rastegar, P. Aanstad, R. Herwig, F. Chalmel, N. Fischer, and U. Strahle. 2004. Expression profiling and comparative genomics identify a conserved regulatory region controlling midline expression in the zebrafish embryo. *Genome Res* **14:** 228-238.

Dodd, J., T.M. Jessell, and M. Placzek. 1998. The when and where of floor plate induction. *Science* **282:** 1654-1657.

Driever, W., L. Solnica-Krezel, A.F. Schier, S.C. Neuhauss, J. Malicki, D.L. Stemple, D.Y. Stainier, F. Zwartkruis, S. Abdelilah, Z. Rangini et al. 1996. A genetic screen for mutations affecting embryogenesis in zebrafish. *Development* **123:** 37-46.

Du, S.J. and M. Dienhart. 2001. Zebrafish tiggy-winkle hedgehog promoter directs notochord and floor plate green fluorescence protein expression in transgenic zebrafish embryos. *Dev Dyn* **222:** 655-666.

Dudley, A.T. and E.J. Robertson. 1997. Overlapping expression domains of bone morphogenetic protein family members potentially account for limited tissue defects in BMP7 deficient embryos. *Dev Dyn* **208:** 349-362.

Ekker, S.C., A.R. Ungar, P. Greenstein, D.P. von Kessler, J.A. Porter, R.T. Moon, and P.A. Beachy. 1995. Patterning activities of vertebrate hedgehog proteins in the developing eye and brain. *Curr Biol* **5:** 944-955.

Entchev, E.V. and M.A. Gonzalez-Gaitan. 2002. Morphogen gradient formation and vesicular trafficking. *Traffic* **3:** 98-109.

Epstein, D.J., A.P. McMahon, and A.L. Joyner. 1999. Regionalization of Sonic hedgehog transcription along the anteroposterior axis of the mouse central nervous system is regulated by Hnf3-dependent and -independent mechanisms. *Development* **126:** 281-292.

Ertzer, R., F. Muller, Y. Hadzhiev, S. Rathnam, N. Fischer, S. Rastegar, and U. Strahle. 2007. Cooperation of sonic hedgehog enhancers in midline expression. *Dev Biol* **301:** 578-589.

Etheridge, L.A., T. Wu, J.O. Liang, S.C. Ekker, and M.E. Halpern. 2001. Floor plate develops upon depletion of tiggy-winkle and sonic hedgehog. *Genesis* **30:** 164-169.

Falb, D. and T. Maniatis. 1992. Drosophila transcriptional repressor protein that binds specifically to negative control elements in fat body enhancers. *Mol Cell Biol* **12:** 4093-4103.

Feldman, B., M.A. Gates, E.S. Egan, S.T. Dougan, G. Rennebeck, H.I. Sirotkin, A.F. Schier, and W.S. Talbot. 1998. Zebrafish organizer development and germ-layer formation require nodal-related signals. *Nature* **395:** 181-185.

Feng, J., C. Bi, B.S. Clark, R. Mady, P. Shah, and J.D. Kohtz. 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:** 1470-1484.

Fisher, R.A. 1935. The sheltering of lethals. *Am. Nat.* **69:** 446-455.

Fondon, J.W., 3rd and H.R. Garner. 2004. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A* **101:** 18058-18063.

Force, A., M. Lynch, F.B. Pickett, A. Amores, Y.L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531-1545.

Fouquet, B., B.M. Weinstein, F.C. Serluca, and M.C. Fishman. 1997. Vessel patterning in the embryo of the zebrafish: guidance by notochord. *Dev Biol* **183:** 37-48.

Frazer, K.A., L. Pachter, A. Poliakov, E.M. Rubin, and I. Dubchak. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32:** W273-279.

Gaiano, N., M. Allende, A. Amsterdam, K. Kawakami, and N. Hopkins. 1996a. Highly efficient germ-line transmission of proviral insertions in zebrafish. *Proc Natl Acad Sci U S A* **93:** 7777-7782.

Gaiano, N., A. Amsterdam, K. Kawakami, M. Allende, T. Becker, and N. Hopkins. 1996b. Insertional mutagenesis and rapid cloning of essential genes in zebrafish. *Nature* **383:** 829-832.

Gallet, A., L. Ruel, L. Staccini-Lavenant, and P.P. Therond. 2006. Cholesterol modification is necessary for controlled planar long-range activity of Hedgehog in Drosophila epithelia. *Development* **133:** 407-418.

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

Glazko, G.V., E.V. Koonin, I.B. Rogozin, and S.A. Shabalina. 2003. A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. *Trends Genet* **19:** 119-124.

Goldstein, A.M. and M.C. Fishman. 1998. Notochord regulates cardiac lineage in zebrafish embryos. *Dev Biol* **201:** 247-252.

Gompel, N., B. Prud'homme, P.J. Wittkopp, V.A. Kassner, and S.B. Carroll. 2005. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature* **433:** 481-487.

Gray, S., P. Szymanski, and M. Levine. 1994. Short-range repression permits multiple enhancers to function autonomously within a complex promoter. *Genes Dev* **8:** 1829-1838.

Gritsman, K., J. Zhang, S. Cheng, E. Heckscher, W.S. Talbot, and A.F. Schier. 1999. The EGF-CFC protein one-eyed pinhead is essential for nodal signaling. *Cell* **97:** 121-132.

Hadrys, T., V. Prince, M. Hunter, R. Baker, and S. Rinkwitz. 2004. Comparative ge-

nomic analysis of vertebrate Hox3 and Hox4 genes. *J Exp Zoolog B Mol Dev Evol* **302:** 147-164.

Hadrys, T., B. Punnamoottil, M. Pieper, H. Kikuta, G. Pezeron, T.S. Becker, V. Prince, R. Baker, and S. Rinkwitz. 2006. Conserved co-regulation and promoter sharing of hoxb3a and hoxb4a in zebrafish. *Dev Biol* **297:** 26-43.

Haffter, P., M. Granato, M. Brand, M.C. Mullins, M. Hammerschmidt, D.A. Kane, J. Odenthal, F.J. van Eeden, Y.J. Jiang, C.P. Heisenberg et al. 1996. The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. *Development* **123:** 1-36.

Haldane, J.B.S. 1933. The part played by recurrent mutation in evolution. *Am. Nat.* **67:** 5-9.

Han, C., T.Y. Belenkaya, B. Wang, and X. Lin. 2004. Drosophila glypicans control the cell-to-cell movement of Hedgehog by a dynamin-independent process. *Development* **131:** 601-611.

Harland, B.F., S.A. Smith, R. Ellis, R. O'Brien, and E.R. Morris. 1992. Comparison of the nutrient intakes of blacks, Siouan Indians, and whites in Columbus County, North Carolina. *J Am Diet Assoc* **92:** 348-350.

Harland, R. and J. Gerhart. 1997. Formation and function of Spemann's organizer. *Annu Rev Cell Dev Biol* **13:** 611-667.

Hemmati-Brivanlou, A., O.G. Kelly, and D.A. Melton. 1994. Follistatin, an antagonist of activin, is expressed in the Spemann organizer and displays direct neuralizing activity. *Cell* **77:** 283-295.

Higuchi, R., B. Krummel, and R.K. Saiki. 1988. A general method of in vitro preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions. *Nucleic Acids Res* **16:** 7351-7367.

Hoegg, S., H. Brinkmann, J.S. Taylor, and A. Meyer. 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J Mol Evol* **59:** 190-203.

Holland, P.W. 1997. Vertebrate evolution: something fishy about Hox genes. *Curr Biol* **7:** R570-572.

Holland, P.W., J. Garcia-Fernandez, N.A. Williams, and A. Sidow. 1994. Gene duplications and the origins of vertebrate development. *Dev Suppl***:** 125-133.

Hooper, J.E. and M.P. Scott. 2005. Communicating with Hedgehogs. *Nat Rev Mol Cell Biol* **6:** 306-317.

Howard, M.L. and E.H. Davidson. 2004. cis-Regulatory control circuits in development. *Dev Biol* **271:** 109-118.

Hughes, K.A., J.F. Ayroles, M.M. Reedy, J.M. Drnevich, K.C. Rowe, E.A. Ruedi, C.E. Caceres, and K.N. Paige. 2006. Segregating variation in the transcriptome: cis regulation and additivity of effects. *Genetics* **173:** 1347-1355.

Ingham, P.W. 1998. Transducing Hedgehog: the story so far. *Embo J* **17:** 3505-3511.

Ingham, P.W. and A.P. McMahon. 2001. Hedgehog signaling in animal development: paradigms and principles. *Genes Dev* **15:** 3059-3087.

Ingham, P.W. and M. Placzek. 2006. Orchestrating ontogenesis: variations on a theme by sonic hedgehog. *Nat Rev Genet* **7:** 841-850.

Isaacs, H.V., M.E. Pownall, and J.M. Slack. 1995. eFGF is expressed in the dorsal midline of Xenopus laevis. *Int J Dev Biol* **39:** 575-579.

Iyer, V.R., C.E. Horak, C.S. Scafe, D. Botstein, M. Snyder, and P.O. Brown. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and

MBF. *Nature* **409:** 533-538.
J. Wittbrodt, A.M.M.S. 1998. More genes in fish? *BioEssays* **20:** 511-515.
Jaillon, O., J.M. Aury, F. Brunet, J.L. Petit, N. Stange-Thomann, E. Mauceli, L. Bouneau, C. Fischer, C. Ozouf-Costaz, A. Bernot et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431:** 946-957.
Jegga, A.G., S.P. Sherwood, J.W. Carman, A.T. Pinski, J.L. Phillips, J.P. Pestian, and B.J. Aronow. 2002. Detection and visualization of compositionally similar cis-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res* **12:** 1408-1417.
Jeong, S., A. Rokas, and S.B. Carroll. 2006a. Regulation of body pigmentation by the Abdominal-B Hox protein and its gain and loss in Drosophila evolution. *Cell* **125:** 1387-1399.
Jeong, Y., K. El-Jaick, E. Roessler, M. Muenke, and D.J. Epstein. 2006b. A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. *Development* **133:** 761-772.
Jeong, Y. and D.J. Epstein. 2003. Distinct regulators of Shh transcription in the floor plate and notochord indicate separate origins for these tissues in the mouse node. *Development* **130:** 3891-3902.
Jessell, T.M. 2000. Neuronal specification in the spinal cord: inductive signals and transcriptional codes. *Nat Rev Genet* **1:** 20-29.
Johansson, O., W. Alkema, W.W. Wasserman, and J. Lagergren. 2003. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics* **19 Suppl 1:** i169-176.
Jones, S., P. van Heyningen, H.M. Berman, and J.M. Thornton. 1999. Protein-DNA interactions: A structural analysis. *J Mol Biol* **287:** 877-896.
Joseph, E.M. and D.A. Melton. 1997. Xnr4: a Xenopus nodal-related gene expressed in the Spemann organizer. *Dev Biol* **184:** 367-372.
Kane, D.A. and C.B. Kimmel. 1993. The zebrafish midblastula transition. *Development* **119:** 447-456.
Kawakami, K., A. Koga, H. Hori, and A. Shima. 1998. Excision of the tol2 transposable element of the medaka fish, Oryzias latipes, in zebrafish, Danio rerio. *Gene* **225:** 17-22.
Kawakami, K., H. Takeda, N. Kawakami, M. Kobayashi, N. Matsuda, and M. Mishina. 2004. A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. *Dev Cell* **7:** 133-144.
Kimmel, C.B., W.W. Ballard, S.R. Kimmel, B. Ullmann, and T.F. Schilling. 1995. Stages of embryonic development of the zebrafish. *Dev Dyn* **203:** 253-310.
King, M.C. and A.C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188:** 107-116.
Knoepfler, P.S. and M.P. Kamps. 1995. The pentapeptide motif of Hox proteins is required for cooperative DNA binding with Pbx1, physically contacts Pbx1, and enhances DNA binding by Pbx1. *Mol Cell Biol* **15:** 5811-5819.
Kondaiah, P., M. Taira, U.D. Vempati, and I.B. Dawid. 2000. Transforming growth factor-beta5 expression during early development of Xenopus laevis. *Mech Dev* **95:** 207-209.
Korzh, V.P. 2001. [T-Box genes and developmental decisions that cells make]. *Ontogenez* **32:** 196-203.

Krauss, S., J.P. Concordet, and P.W. Ingham. 1993. A functionally conserved homolog of the Drosophila segment polarity gene hh is expressed in tissues with polarizing activity in zebrafish embryos. *Cell* **75:** 1431-1444.

Kuras, L. and K. Struhl. 1999. Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme. *Nature* **399:** 609-613.

Laemmli, U.K., E. Kas, L. Poljak, and Y. Adachi. 1992. Scaffold-associated regions: cis-acting determinants of chromatin structural loops and functional domains. *Curr Opin Genet Dev* **2:** 275-285.

Latchman, D.S. 1997. Transcription factors: an overview. *Int J Biochem Cell Biol* **29:** 1305-1312.

Lauderdale, J.D., S.K. Pasquali, R. Fazel, F.J. van Eeden, H.E. Schauerte, P. Haffter, and J.Y. Kuwada. 1998. Regulation of netrin-1a expression by hedgehog proteins. *Mol Cell Neurosci* **11:** 194-205.

Le Douarin, N.M. and M.E. Halpern. 2000. Discussion point. Origin and specification of the neural tube floor plate: insights from the chick and zebrafish. *Curr Opin Neurobiol* **10:** 23-30.

Lee, T.I. and R.A. Young. 2000. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* **34:** 77-137.

Lemon, B. and R. Tjian. 2000. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev* **14:** 2551-2569.

Lenhard, B., A. Sandelin, L. Mendoza, P. Engstrom, N. Jareborg, and W.W. Wasserman. 2003. Identification of conserved regulatory elements by comparative genome analysis. *J Biol* **2:** 13.

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

Levine, M. and E.H. Davidson. 2005. Gene regulatory networks for development. *Proc Natl Acad Sci U S A* **102:** 4936-4942.

Lewin, B. 2000. *Genes VII*. Oxford University Press, Oxford.

Lewis, P.M., M.P. Dunn, J.A. McMahon, M. Logan, J.F. Martin, B. St-Jacques, and A.P. McMahon. 2001. Cholesterol modification of sonic hedgehog is required for long-range signaling activity and effective modulation of signaling by Ptc1. *Cell* **105:** 599-612.

Li, Y., H. Zhang, Y. Litingtung, and C. Chiang. 2006. Cholesterol modification restricts the spread of Shh gradient in the limb bud. *Proc Natl Acad Sci U S A* **103:** 6548-6553.

Lim, L.P., M.E. Glasner, S. Yekta, C.B. Burge, and D.P. Bartel. 2003. Vertebrate microRNA genes. *Science* **299:** 1540.

Linsenmayer, T.F., E. Gibney, and T.M. Schmid. 1986. Segmental appearance of type X collagen in the developing avian notochord. *Dev Biol* **113:** 467-473.

Locker, J. 2001. *Transcription factors*. BIOS ;
Academic Press, Oxford
San Diego, CA.

Lohr, J.L., M.C. Danos, and H.J. Yost. 1997. Left-right asymmetry of a nodal-related gene is regulated by dorsoanterior midline structures during Xenopus development. *Development* **124:** 1465-1472.

Loots, G.G. and I. Ovcharenko. 2004. rVISTA 2.0: evolutionary analysis of tran-

scription factor binding sites. *Nucleic Acids Res* **32:** W217-221.
Loots, G.G., I. Ovcharenko, L. Pachter, I. Dubchak, and E.M. Rubin. 2002. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res* **12:** 832-839.
Ludwig, M.Z. 2002. Functional evolution of noncoding DNA. *Curr Opin Genet Dev* **12:** 634-639.
Ludwig, M.Z., C. Bergman, N.H. Patel, and M. Kreitman. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403:** 564-567.
Ludwig, M.Z., A. Palsson, E. Alekseeva, C.M. Bergman, J. Nathan, and M. Kreitman. 2005. Functional evolution of a cis-regulatory module. *PLoS Biol* **3:** e93.
Ludwig, M.Z., N.H. Patel, and M. Kreitman. 1998. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development* **125:** 949-958.
Lum, L. and P.A. Beachy. 2004. The Hedgehog response network: sensors, switches, and routers. *Science* **304:** 1755-1759.
Ma, Y., A. Erkner, R. Gong, S. Yao, J. Taipale, K. Basler, and P.A. Beachy. 2002. Hedgehog-mediated patterning of the mammalian embryo requires transporter-like function of dispatched. *Cell* **111:** 63-75.
Marcellini, S. and P. Simpson. 2006. Two or Four Bristles: Functional Evolution of an Enhancer of scute in Drosophilidae. *PLoS Biol* **4:** e386.
Marino, S. 2005. Medulloblastoma: developmental mechanisms out of control. *Trends Mol Med* **11:** 17-22.
Marti, E. 2000. Expression of chick BMP-1/Tolloid during patterning of the neural tube and somites. *Mech Dev* **91:** 415-419.
Marti, E., R. Takada, D.A. Bumcrot, H. Sasaki, and A.P. McMahon. 1995. Distribution of Sonic hedgehog peptides in the developing chick and mouse embryo. *Development* **121:** 2537-2547.
Mayor, C., M. Brudno, J.R. Schwartz, A. Poliakov, E.M. Rubin, K.A. Frazer, L.S. Pachter, and I. Dubchak. 2000. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:** 1046-1047.
Mazet, F. and S.M. Shimeld. 2002. Gene duplication and divergence in the early evolution of vertebrates. *Curr Opin Genet Dev* **12:** 393-396.
McEwen, G.K., A. Woolfe, D. Goode, T. Vavouri, H. Callaway, and G. Elgar. 2006. Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. *Genome Res* **16:** 451-465.
Meyer, A. 2003. Molecular evolution: Duplication, duplication. *Nature* **421:** 31-32.
Meyer, A. and E. Malaga-Trillo. 1999. Vertebrate genomics: More fishy tales about Hox genes. *Curr Biol* **9:** R210-213.
Meyer, A. and M. Schartl. 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr Opin Cell Biol* **11:** 699-704.
Minokawa, T., A.H. Wikramanayake, and E.H. Davidson. 2005. cis-Regulatory inputs of the wnt8 gene in the sea urchin endomesoderm network. *Dev Biol* **288:** 545-558.
Moses, A.M., D.Y. Chiang, D.A. Pollard, V.N. Iyer, and M.B. Eisen. 2004. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol* **5:** R98.
Muller, F., S. Albert, P. Blader, N. Fischer, M. Hallonet, and U. Strahle. 2000. Direct

action of the nodal-related signal cyclops in induction of sonic hedgehog in the ventral midline of the CNS. *Development* **127:** 3889-3897.

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

Mullis, K.B. and F.A. Faloona. 1987. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol* **155:** 335-350.

Munsterberg, A.E. and A.B. Lassar. 1995. Combinatorial signals from the neural tube, floor plate and notochord induce myogenic bHLH gene expression in the somite. *Development* **121:** 651-660.

Nakano, Y., H.R. Kim, A. Kawakami, S. Roy, A.F. Schier, and P.W. Ingham. 2004. Inactivation of dispatched 1 by the chameleon mutation disrupts Hedgehog signalling in the zebrafish embryo. *Dev Biol* **269:** 381-392.

Nasevicius, A. and S.C. Ekker. 2000. Effective targeted gene 'knockdown' in zebrafish. *Nat Genet* **26:** 216-220.

Neumann, C.J., H. Grandel, W. Gaffield, S. Schulte-Merker, and C. Nusslein-Volhard. 1999. Transient establishment of anteroposterior polarity in the zebrafish pectoral fin bud in the absence of sonic hedgehog activity. *Development* **126:** 4817-4826.

Nusslein-Volhard, C. and E. Wieschaus. 1980. Mutations affecting segment number and polarity in Drosophila. *Nature* **287:** 795-801.

O'Brien, S.J., J.F. Eisenberg, M. Miyamoto, S.B. Hedges, S. Kumar, D.E. Wilson, M. Menotti-Raymond, W.J. Murphy, W.G. Nash, L.A. Lyons et al. 1999. Genome maps 10. Comparative genomics. Mammalian radiations. Wall chart. *Science* **286:** 463-478.

Odenthal, J., F.J. van Eeden, P. Haffter, P.W. Ingham, and C. Nusslein-Volhard. 2000. Two distinct cell populations in the floor plate of the zebrafish are induced by different pathways. *Dev Biol* **219:** 350-363.

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

Oppenheimer, J.M. 1936. Transplantation experiments on developing teleosts (Fundulus and Perca). *Journal of Experimental Zoology* **72:** 409-437.

Pan, Y., C.B. Bai, A.L. Joyner, and B. Wang. 2006. Sonic hedgehog signaling regulates Gli2 transcriptional activity by suppressing its processing and degradation. *Mol Cell Biol* **26:** 3365-3377.

Parinov, S., I. Kondrichin, V. Korzh, and A. Emelyanov. 2004. Tol2 transposon-mediated enhancer trap to identify developmentally regulated zebrafish genes in vivo. *Dev Dyn* **231:** 449-459.

Placzek, M. and J. Briscoe. 2005. The floor plate: multiple cells, multiple signals. *Nat Rev Neurosci* **6:** 230-240.

Placzek, M., J. Dodd, and T.M. Jessell. 2000. Discussion point. The case for floor plate induction by the notochord. *Curr Opin Neurobiol* **10:** 15-22.

Placzek, M., T.M. Jessell, and J. Dodd. 1993. Induction of floor plate differentiation by contact-dependent, homeogenetic signals. *Development* **117:** 205-218.

Plessy, C., T. Dickmeis, F. Chalmel, and U. Strähle. 2005. Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. *Trends in Genetics* **in press**.

Pogoda, H.M., L. Solnica-Krezel, W. Driever, and D. Meyer. 2000. The zebrafish

forkhead transcription factor FoxH1/Fast1 is a modulator of nodal signaling required for organizer formation. *Curr Biol* **10:** 1041-1049.

Postlethwait, J.H., Y.L. Yan, M.A. Gates, S. Horne, A. Amores, A. Brownlie, A. Donovan, E.S. Egan, A. Force, Z. Gong et al. 1998. Vertebrate genome evolution and the zebrafish gene map. *Nat Genet* **18:** 345-349.

Pourquie, O., M. Coltey, M.A. Teillet, C. Ordahl, and N.M. Le Douarin. 1993. Control of dorsoventral patterning of somitic derivatives by notochord and floor plate. *Proc Natl Acad Sci U S A* **90:** 5242-5246.

Prince, V.E. and F.B. Pickett. 2002. Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet* **3:** 827-837.

Prud'homme, B., N. Gompel, A. Rokas, V.A. Kassner, T.M. Williams, S.D. Yeh, J.R. True, and S.B. Carroll. 2006. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature* **440:** 1050-1053.

Rabinow, P. 1996. *Making PCR : a story of biotechnology*. University of Chicago Press, Chicago ; London.

Rebagliati, M.R., R. Toyama, C. Fricke, P. Haffter, and I.B. Dawid. 1998. Zebrafish nodal-related genes are implicated in axial patterning and establishing left-right asymmetry. *Dev Biol* **199:** 261-272.

Roelink, H., J.A. Porter, C. Chiang, Y. Tanabe, D.T. Chang, P.A. Beachy, and T.M. Jessell. 1995. Floor plate and motor neuron induction by different concentrations of the amino-terminal cleavage product of sonic hedgehog autoproteolysis. *Cell* **81:** 445-455.

Ruiz i Altaba, A. 1994. Pattern formation in the vertebrate neural plate. *Trends Neurosci* **17:** 233-243.

Sambrook, J. and D.W. Russell. 2001. *Molecular cloning : a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Sandelin, A., W. Alkema, P. Engstrom, W.W. Wasserman, and B. Lenhard. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32 Database issue:** D91-94.

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

Sasai, Y., B. Lu, H. Steinbeisser, D. Geissert, L.K. Gont, and E.M. De Robertis. 1994. Xenopus chordin: a novel dorsalizing factor activated by organizer-specific homeobox genes. *Cell* **79:** 779-790.

Scemama, J.L., M. Hunter, J. McCallum, V. Prince, and E. Stellwag. 2002. Evolutionary divergence of vertebrate Hoxb2 expression patterns and transcriptional regulatory loci. *J Exp Zool* **294:** 285-299.

Schafer, M., D. Kinzel, C. Neuner, M. Schartl, J.N. Volff, and C. Winkler. 2005. Hedgehog and retinoid signalling confines nkx2.2b expression to the lateral floor plate of the zebrafish trunk. *Mech Dev* **122:** 43-56.

Schauerte, H.E., F.J. van Eeden, C. Fricke, J. Odenthal, U. Strahle, and P. Haffter. 1998. Sonic hedgehog is not required for the induction of medial floor plate cells in the zebrafish. *Development* **125:** 2983-2993.

Scholpp, S., O. Wolf, M. Brand, and A. Lumsden. 2006. Hedgehog signalling from the zona limitans intrathalamica orchestrates patterning of the zebrafish diencephalon. *Development* **133:** 855-864.

Shamim, H., R. Mahmood, C. Logan, P. Doherty, A. Lumsden, and I. Mason. 1999.

Sequential roles for Fgf4, En1 and Fgf8 in specification and regionalisation of the midbrain. *Development* **126:** 945-959.

Shapiro, M.D., M.E. Marks, C.L. Peichel, B.K. Blackman, K.S. Nereng, B. Jonsson, D. Schluter, and D.M. Kingsley. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428:** 717-723.

Shashikant, C.S., C.B. Kim, M.A. Borbely, W.C. Wang, and F.H. Ruddle. 1998. Comparative studies on mammalian Hoxc8 early enhancer sequence reveal a baleen whale-specific deletion of a cis-acting element. *Proc Natl Acad Sci U S A* **95:** 15446-15451.

Sirotkin, H.I., M.A. Gates, P.D. Kelly, A.F. Schier, and W.S. Talbot. 2000. Fast1 is required for the development of dorsal axial structures in zebrafish. *Curr Biol* **10:** 1051-1054.

Smith, W.C. and R.M. Harland. 1992. Expression cloning of noggin, a new dorsalizing factor localized to the Spemann organizer in Xenopus embryos. *Cell* **70:** 829-840.

Smits, P. and V. Lefebvre. 2003. Sox5 and Sox6 are required for notochord extracellular matrix sheath formation, notochord cell survival and development of the nucleus pulposus of intervertebral discs. *Development* **130:** 1135-1148.

Sorek, R. and G. Ast. 2003. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res* **13:** 1631-1637.

Spemann, H. and H. Mangold. 1924. Über die Induktion von Embryonalanlagen durch Implantation artfremder Organisatoren. *Roux's Arch. Entw.Mech.Org.* **100:** 599–638.

Stauber, M., A. Prell, and U. Schmidt-Ott. 2002. A single Hox3 gene with composite bicoid and zerknullt expression characteristics in non-Cyclorrhaphan flies. *Proc Natl Acad Sci U S A* **99:** 274-279.

Steinke, D., W. Salzburger, and A. Meyer. 2006. Novel relationships among ten fish model species revealed based on a phylogenomic analysis using ESTs. *J Mol Evol* **62:** 772-784.

Stemple, D.L. 2004. The notochord. *Curr Biol* **14:** R873-874.

Stemple, D.L. 2005. Structure and function of the notochord: an essential organ for chordate development. *Development* **132:** 2503-2512.

Stern, D.L. 1998. A role of Ultrabithorax in morphological differences between Drosophila species. *Nature* **396:** 463-466.

Stone, D.M., M. Hynes, M. Armanini, T.A. Swanson, Q. Gu, R.L. Johnson, M.P. Scott, D. Pennica, A. Goddard, H. Phillips et al. 1996. The tumour-suppressor gene patched encodes a candidate receptor for Sonic hedgehog. *Nature* **384:** 129-134.

Stormo, G.D. 2000. DNA binding sites: representation and discovery. *Bioinformatics* **16:** 16-23.

Strahle, U., P. Blader, and P.W. Ingham. 1996. Expression of axial and sonic hedgehog in wildtype and midline defective zebrafish embryos. *Int J Dev Biol* **40:** 929-940.

Strahle, U., C.S. Lam, R. Ertzer, and S. Rastegar. 2004. Vertebrate floor-plate specification: variations on common themes. *Trends Genet* **20:** 155-162.

Sucena, E., I. Delon, I. Jones, F. Payre, and D.L. Stern. 2003. Regulatory evolution of shavenbaby/ovo underlies multiple cases of morphological parallelism.

Nature **424:** 935-938.
Sucena, E. and D.L. Stern. 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:** 4530-4534.
Swiderski, R.E. and M. Solursh. 1992. Localization of type II collagen, long form alpha 1(IX) collagen, and short form alpha 1(IX) collagen transcripts in the developing chick notochord and axial skeleton. *Dev Dyn* **194:** 118-127.
Tagle, D.A., B.F. Koop, M. Goodman, J.L. Slightom, D.L. Hess, and R.T. Jones. 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* **203:** 439-455.
Tautz, D. 2000. Evolution of transcriptional regulation. *Curr Opin Genet Dev* **10:** 575-579.
Taylor, J.S., I. Braasch, T. Frickey, A. Meyer, and Y. Van de Peer. 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res* **13:** 382-390.
Taylor, J.S., Y. Van de Peer, and A. Meyer. 2001a. Genome duplication, divergent resolution and speciation. *Trends Genet* **17:** 299-301.
Taylor, J.S., Y. Van de Peer, and A. Meyer. 2001b. Revisiting recent challenges to the ancient fish-specific genome duplication hypothesis. *Curr Biol* **11:** R1005-1008.
Tenzen, T., B.L. Allen, F. Cole, J.S. Kang, R.S. Krauss, and A.P. McMahon. 2006. The cell surface membrane proteins Cdo and Boc are components and targets of the Hedgehog signaling pathway and feedback network in mice. *Dev Cell* **10:** 647-656.
Tompa, M., N. Li, T.L. Bailey, G.M. Church, B. De Moor, E. Eskin, A.V. Favorov, M.C. Frith, Y. Fu, W.J. Kent et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23:** 137-144.
Tumpel, S., F. Cambronero, L.M. Wiedemann, and R. Krumlauf. 2006. Evolution of cis elements in the differential expression of two Hoxa2 coparalogous genes in pufferfish (Takifugu rubripes). *Proc Natl Acad Sci U S A* **103:** 5419-5424.
Tvrdik, P. and M.R. Capecchi. 2006. Reversal of Hox1 gene subfunctionalization in the mouse. *Dev Cell* **11:** 239-250.
van den Heuvel, M. and P.W. Ingham. 1996. smoothened encodes a receptor-like serpentine protein required for hedgehog signalling. *Nature* **382:** 547-551.
Varga, Z.M., A. Amores, K.E. Lewis, Y.L. Yan, J.H. Postlethwait, J.S. Eisen, and M. Westerfield. 2001. Zebrafish smoothened functions in ventral neural tube specification and axon tract formation. *Development* **128:** 3497-3509.
Volff, J.N. 2005. Genome evolution and biodiversity in teleost fish. *Heredity* **94:** 280-294.
Waddington, C.H. 1930. Developmental mechanics of chick and duck embryos. *Nature* **125:** 924 -925.
Wang, B., J.F. Fallon, and P.A. Beachy. 2000. Hedgehog-regulated processing of Gli3 produces an anterior/posterior repressor gradient in the developing vertebrate limb. *Cell* **100:** 423-434.
Wang, R.L., A. Stec, J. Hey, L. Lukens, and J. Doebley. 1999. The limits of selection during maize domestication. *Nature* **398:** 236-239.
White, K.P., S.A. Rifkin, P. Hurban, and D.S. Hogness. 1999. Microarray analysis of

Drosophila development during metamorphosis. *Science* **286:** 2179-2184.
White, R.J. 2001. *Gene transcription : mechanisms and control*. Blackwell Science, Oxford.
Wienholds, E., F. van Eeden, M. Kosters, J. Mudde, R.H. Plasterk, and E. Cuppen. 2003. Efficient target-selected mutagenesis in zebrafish. *Genome Res* **13:** 2700-2707.
Wingender, E., P. Dietze, H. Karas, and R. Knuppel. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24:** 238-241.
Wittkopp, P.J. 2006. Evolution of cis-regulatory sequence and function in Diptera. *Heredity* **97:** 139-147.
Wittkopp, P.J., S.B. Carroll, and A. Kopp. 2003. Evolution in black and white: genetic control of pigment patterns in Drosophila. *Trends Genet* **19:** 495-504.
Wittkopp, P.J., B.K. Haerum, and A.G. Clark. 2004. Evolutionary changes in cis and trans gene regulation. *Nature* **430:** 85-88.
Wittkopp, P.J., K. Vaccaro, and S.B. Carroll. 2002. Evolution of yellow gene regulation and pigmentation in Drosophila. *Curr Biol* **12:** 1547-1556.
Wray, G.A., M.W. Hahn, E. Abouheif, J.P. Balhoff, M. Pizer, M.V. Rockman, and L.A. Romano. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* **20:** 1377-1419.
Yamada, T., S.L. Pfaff, T. Edlund, and T.M. Jessell. 1993. Control of cell pattern in the neural tube: motor neuron induction by diffusible factors from notochord and floor plate. *Cell* **73:** 673-686.
Yamada, T., M. Placzek, H. Tanaka, J. Dodd, and T.M. Jessell. 1991. Control of cell pattern in the developing nervous system: polarizing activity of the floor plate and notochord. *Cell* **64:** 635-647.
Yamamoto, Y., D.W. Stock, and W.R. Jeffery. 2004. Hedgehog signalling controls eye degeneration in blind cavefish. *Nature* **431:** 844-847.
Yao, S., L. Lum, and P. Beachy. 2006. The ihog cell-surface proteins bind Hedgehog and mediate pathway activation. *Cell* **125:** 343-357.
Zardoya, R., E. Abouheif, and A. Meyer. 1996a. Evolution and orthology of hedgehog genes. *Trends Genet* **12:** 496-497.
Zardoya, R., E. Abouheif, and A. Meyer. 1996b. Evolutionary analyses of hedgehog and Hoxd-10 genes in fish species closely related to the zebrafish. *Proc Natl Acad Sci U S A* **93:** 13036-13041.
Zerucha, T., T. Stuhmer, G. Hatch, B.K. Park, Q. Long, G. Yu, A. Gambarotta, J.R. Schultz, J.L. Rubenstein, and M. Ekker. 2000. A highly conserved enhancer in the Dlx5/Dlx6 intergenic region is the site of cross-regulatory interactions between Dlx genes in the embryonic forebrain. *J Neurosci* **20:** 709-721.
Zhang, J., W.S. Talbot, and A.F. Schier. 1998. Positional cloning identifies zebrafish one-eyed pinhead as a permissive EGF-related ligand required during gastrulation. *Cell* **92:** 241-251.
Zhang, W., J.S. Kang, F. Cole, M.J. Yi, and R.S. Krauss. 2006. Cdo functions at multiple points in the Sonic Hedgehog pathway, and Cdo-deficient mice accurately model human holoprosencephaly. *Dev Cell* **10:** 657-665.