Institut für Entwicklungsgenetik, GSF-Forschungszentrum,
Neuherberg

Analysis of promoter and *cis*-regulatory elements that control *Pax9* gene expression during mouse embryonic development

Fabio Santagati

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. A. Gierl

Prüfer der Dissertation: 1. Hon.-Prof. Dr. R. Balling, Technische Universität Carolo-Wilhelmina zu Braunschweig
2. Univ.-Prof. Dr. H. Daniel

Die Dissertation wurde am 05.08.2002 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 14.10.2002 angenommen.

I would like to thank

Prof. Rudi Balling
supervisor of my PhD thesis
former Director of the Institute of Mammalian Genetics of the GSF
and now Scientific Director of the GBF Research Centre, Braunschweig-Germany
who gave me the opportunity to fulfill my PhD work in his laboratory and believed in my capability making me from the very beginning responsible of my own project;

Dr. Kenji Imai
Group leader in the present Institute of Developmental Genetics of the GSF
who directly supervised my experimental work with precious advice and always offered new incentives for interesting discussions;

all the people from my work group and from other groups of the Institutes of Developmental Genetics and of Experimental Genetics, who were always available for constructive interactions and created a friendly atmosphere, in particular Matthias Wahl for his dedication in the critical reading of the manuscript of this thesis.

I would also like to thank all the new and old, near and far friends, in particular those with whom I shared the good and the bad moments of my life in Munich, hoping that this friendship will keep strong for a long time.

Finally, I dedicate this thesis to my family, who gave me support in every decision and in every step of my life with care and affection. I hope they can always be proud of me as I am of them.

Infine dedico questa tesi alla mia famiglia, che mi ha sempre dato sostegno in ogni decisione ed in ogni passo della mia vita con attenzione ed affetto. Spero che possano essere sempre orgogliosi di me come io lo sono di loro.

TABLE OF CONTENTS

1. ZUSAMMENFASSUNG

Pax9 kodiert für einen Transkriptionsfaktor, der eine paired-Domäne enthält und während der Embryogenese essentielle Funktionen hat. *Pax9* wird in verschiedenen embryonalen Geweben exprimiert, einschließlich dem Sklerotom der Somiten, den Extremitätenanlagen, dem Entoderm der Kiementaschen, dem Gesichtsschädelmesenchym, dem Entoderm des Ösophagus und dem hintersten Teil des Entoderms. Die molekularen Mechanismen, die *Pax9* regulieren, sind kaum bekannt. Außerdem sind noch keine *cis*-regulatorischen Elemente von *Pax9* identifiziert worden.

Eine physikalische Karte von einem ~400-kb Bereich, der Maus-*Pax9* umfasst, wurde durch die Isolierung und Charakterisierung von 11 überlappenden BAC-Klonen etabiliert. Diese physikalische Karte legte die Grundlage für die folgende Analyse.

Die Pax9-Exon/Intron-Struktur wurde ermittelt und durch ausführliche Analyse von *Pax9*-Transkripten wurde die Existenz von einem vorher unbekannten Exon (bezeichnet als Exon 0) weiter „upstream“ nachgewiesen. Die Promotoraktivität der zwei entsprechenden putativen Promotoren wurde in *Pax9*-exprimierenden Zelllinien *in vitro* getestet.

Eine breitere Analyse der BAC-Karte enthüllte zusätzlich die anliegenden Gene von *Pax9*: *Nkx2-9* etwa 75-kb „upstream“ von *Pax9* und die letzten Exons vom *Odc*-Gen (mitochondrial oxodicarboxylate carrier) 2-kb „downstream“ vom letzten *Pax9*-Exon.

Die Maus- und Kugelfischsequenz des genomischen Bereiches von *Pax9* wurden mit der entsprechenden humanen genomischen Sequenz verglichen. Die Analyse zeigte 1) die konservierte Syntenie von *Pax9* und seinen anliegenden Genen in diesen drei Tierarten und 2) die Anwesenheit von mehreren nicht-kodierenden genomischen Segmenten (CNS) mit hohem Grad an Sequenzkonservierung, die als starke Kandidaten für *cis*-regulatorische Elemente betrachtet werden können. Die regulatorische Aktivität von zwei CNS-Fragmenten, von denen sich eins (CNS-6) zwischen *Nkx2-9* und *Pax9* und das andere (CNS+2) „downstream“ von *Pax9* befinden, wurde durch konventionelle Transgenese *in vivo* getestet. Diese Segmente wurden vor den minimalen Promotor des *hsp68*-Genes kloniert, um ein Marker-Gen, *lacZ,* zu steuern. Es hat sich herausgestellt, dass CNS–6 ein *cis*-regulatorisches Element ist, das die Expression von *Nkx2-9* im ventralen Teil des Neuralrohres kontrolliert, während CNS+2 die Expression vom

Marker-Gen *lacZ* in einem Bereich des *Pax9*-positiven Gesichtsschädelmesenchym, das der Mundkante der medialen Nasenwülste entspricht, treibt.

Um die Ausdehnung des genomischen Intervalles für den *Pax9*-Lokus zu ermitteln, wurde eine BAC-Transgenese durchgeführt. Ein BAC-Klon wurde durch die 'ET-cloning'-Technik verändert, indem eine IRES-lacZ-neo-Kassette in die 3'-UTR von *Pax9* eingeführt wurde. In transgenen Tieren mit dem veränderten BAC wurde *Pax9-lacZ*-Expression in Extremitätenknospen, in Schwanzentoderm und -muskeln, sowie in einigen Gesichtsschädelbereichen beobachtet. Alle Expressionsdomänen stimmten mit denen von *Pax9* überein. Jedoch spiegelte diese *Pax9-lacZ*-Expression nur teilweise die endogene *Pax9* Expression wider. Zum Beispiel wurde eine Somiten- und Schlundtaschenexpression in den transgenen Mäusen nicht beobachtet.

Das deutet darauf hin, dass der ganze *Pax9*-Lokus sehr groß ist und dass die genomische Organisation der kodierenden und regulatorischen Bereiche komplex ist, und einige wichtige regulatorische Elemente innerhalb anliegender Gene lokalisiert sind.

Diese Folgerung wird als biologische Erklärung der hochkonservierten syntenischen Region von *Pax9* vorgestellt und als Beispiel für einen allgemeinen Mechanismus der Evolution der Genome vorgeschlagen.

2. INTRODUCTION

Developmental biology is the field of biology that studies the development of the organisms; that is the transformations that a fertilized egg cell, or zygote, undergoes, leading to the formation of a new individual.

A zygote divides mitotically to produce all the cells of the body, giving rise to muscle cells, skin cells, neurons, blood cells, and all the other cell types. This generation of cellular diversity is called differentiation and the process that organizes the different cells into tissues and organs is called organogenesis.

The onset of morphological and functional differences that lead the cells of an embryo to differentiate into diverse developmental lines is due to the differential usage of genetic information. Such differential usage occurs through the activation or inactivation of specific sets of genes. This regulation of gene expression can be accomplished at different levels: gene transcription control, RNA processing, translation control and post-translational modifications of proteins. All these control mechanisms ensure that specific proteins are synthesised in the right cell type, at the right time and in the correct amount.

In order to understand the molecular mechanisms that regulate and control embryonic development, it is of prior interest to identify such control genes and their functions.

2.1. *Drosophila* as a model for animal development

The fruitfly *Drosophila melanogaster* has been an extremely useful animal model in the history of developmental biology, which allowed the identification of a great deal of developmental control genes by means of classical genetics in combination with molecular biology methods.

The power of the genetic approach to development has been shown in the analysis of *Drosophila* embryonic axis formation. A polarity along the anterior-posterior axis, responsible for the correct development of anterior and posterior structures in the adult fly, is firmly defined already at the level of oocyte formation before fertilization. This polarity is maintained throughout the embryogenesis, through the larval stage and up to the achievement of the final adult appearance (Nüsslein-Volhard 1991).

The presence of an anterior-posterior axis is not morphologically evident in the first hours of embryonic development, when the high rate of nuclear division is not accompanied by physical cellular separation. Until around the tenth cycle of division the *Drosophila* embryo is an ellipsoid monolayer syncytium of thousands of nuclei surrounding a yolk mass. Most of the nuclei migrate at a certain time point toward the periphery of the egg cell, where they undergo further divisions. At the end of this phase, cell membranes grow down and around the nuclei converting the syncytial monolayer at the periphery of the embryo into a cellular monolayer. The embryo is now at the cellular blastoderm stage. The blastoderm stage is rapidly succeeded by a series of cellular invaginations and movements of cell sheets that constitute gastrulation. During gastrulation three layers of cells segregate - outer ectoderm, inner endoderm and interstitial mesoderm - establishing the multilayer body plan of the organism (a detailed description of *Drosophila* development can be found in Wilkins 1993).

As early as one hour after the onset of gastrulation an important process occurs in the body of the Drosophila embryo; this starts to compartmentalise in segments along the anterior-posterior axis, dividing deeply both the ectodermal and the mesodermal layers. Each of these segments has its own identity and will develop in a corresponding segment of the adult fly. The segmented pattern of the Drosophila embryo establishes, however, even earlier during the cellular blastoderm stage, as cell lineages may already become functionally restricted to segments (Kornberg and Tabata 1993).

The molecular mechanisms that govern the segmentation process started to be analyzed when a systematic search for mutations affecting segmentation was carried out by Nüsslein-Volhard and Wieschaus in 1980. Altogether as many as thirty different loci were identified in this extensive search. The first conclusion was that these loci could be ranged in three general categories according to their mutant phenotypes. In the first class of mutants, broad overlapping non-terminal subregions of the embryo are deleted. The genes associated to these mutations are called gap genes. They are among the first genes transcribed in the embryogenesis. The second group comprises the pair-rule genes, whose mutations lead to repetitive deletions of every other segment throughout the whole body. Finally, the segment polarity genes are responsible for maintaining certain repeated structures within each segment and mutations in this group cause defects that are

reiterated in every segment (Nüsslein-Volhard and Wieschaus 1980). All these genes are known under the name of segmentation genes and they are responsible to mediate an irreversible determination of cell lineages to form segments (Kornberg and Tabata 1993). Another class of genes is of maternal origin. Their mRNAs are accumulated in the cytoplasm of the developing oocyte strictly localized in the anterior or posterior portion of the egg. The products of these mRNAs will be the first anteriorising and posteriorising factors and they will distribute in gradients along the anterior-posterior axis. So for example, the product of the gene *bicoid* will be more abundant at the anterior pole and it will control the formation of anterior structures; while the product of the gene *nanos* will antagonize the function of *bicoid* in the posterior region and promote the formation of posterior structures. The maternal genes co-ordinate the initial expression of the segmentation genes and their corresponding mutants show global effects on the whole segmentation pattern (Nüsslein-Volhard 1991).

Most of the developmental control genes mentioned so far exert their activity at the level of transcription regulation. They are namely so called transcription factors; they recognize specific DNA sequences within other transcriptional units and bind them, resulting in the activation or inactivation of a target gene. Each of these transcription factors can selectively control the expression of several other genes, deciding the functional fate of the cell in which they are expressed. They can also regulate the synthesis of new transcription factors that will in turn generate a new cascade of gene regulation. The cells of a developing organism experience specific and consecutive waves of varying gene activity, that draw them through a series of transformations in each step of the developmental process (Hoch and Jackle 1993).

According to this model, the different classes of segmentation genes follow in order to each other in a series of consequential activations: the maternal genes activate the expression of the gap genes in broad overlapping domains; the different concentrations of the gap gene products cause the pair-rule genes to be transcribed in the primordia of each alternate segment, each giving a striped pattern of seven vertical bands along the anterior-posterior axis; the stripes of the pair-rule gene proteins activate the transcription of the segment polarity genes. Finally, proteins of the gap, pair-rule and segment polarity genes

interact to regulate another class of genes, the homeotic genes, whose transcription determines the developmental fate of each segment (Hoch and Jackle 1993; Kornberg and Tabata 1993).

2.2. Identification of *paired* and paired-box genes in *Drosophila*

After the genetic observations that allowed the identification of developmental control genes in *Drosophila* through the study of related mutants, a molecular biology approach led to the cloning and characterization of these genes. When the DNA and protein sequences of these transcription factors were available, it clearly emerged that these factors appear to share common aminoacidic domains that presumably reflect a functional similarity.

The DNA binding activity of a large number of transcription factors is, for example, a feature that often resides in few specific domains, like homeodomains, zinc-finger domains, helix-loop-helix domains. The presence of one of these domains is normally enough to confer to a protein the ability to bind DNA, while the rest of the protein sequence is required for the specificity of its function via protein-protein interactions with other transcription factors. Genes sharing homologous protein domains are defined as a gene set or a gene family (Dressler and Gruss 1988).

In an evolutionary point of view these functional domains can be regarded as derivatives of a small number of ancestral genes, which combined in various independent assortments originating more complex genes with related functions. The independent assortment of functional domains has the interesting consequence that a particular multidomain gene may belong to more than one gene family (Frigerio et al. 1986).

Paired was one of the *Drosophila* pair-rule genes that were physically isolated after genetic identification. As all the other pair-rule gene mutants, the *paired* mutant shows a deletion of analogous portions at a two-segment periodicity, giving rise to only half of the normal number of segments (Nüsslein-Volhard and Wieschaus 1980). Its expression pattern resembles the one of other pair-rule genes; that is a pattern of seven evenly spaced bands that appear during the late syncytial blastoderm. However at the cellular blastoderm stage its spatial expression undergoes a shift to a fourteen band striped pattern with single segment periodicity (Kilchherr et al. 1986).

Starting from the assumption that any multidomain gene shares common sequence features with several gene sets, a systematic search was initiated in order to find genes that could have sequence similarity to the *paired* gene. This approach indeed brought about the identification of a few genes in *Drosophila*, which showed homology to different regions of *paired*. The sequence comparison between paired and these genes revealed the presence in the *paired* gene of three rather frequently occurring domains. Two domains are in the C-terminus of the paired protein, the prd-repeat also present in the maternal anteriorising factor *bicoid*, and a homeodomain, a DNA binding domain characterized by a helix-turn-helix motif (Frigerio et al. 1986). The homeodomain was originally identified as a common domain in the homeotic gene family but it was subsequently found in several variations in a larger number of transcription factors with a main role in development (Krumlauf 1994).

A different type of domain was for the first time observed on the N-terminal end of the paired protein and it was called paired-domain or paired-box (Bopp et al. 1986). The first identified genes containing this domain were two different transcripts from the *gooseberry* (gsb) locus. Similarly to *paired* the *gsb* genes play as well a role in the segmentation of the *Drosophila* embryo, being part of the group of segment polarity genes. These observations favored the hypothesis that genes bearing homologous domains are involved in related functions. Furthermore, the two *gsb* genes carry a homeodomain that is very similar to the one described in the *paired* gene (Bopp et al. 1986).

A further homology search led, however, to the discovery of other two genes (*pox meso* and *pox neuro*) that contained a paired domain but lacked a homeodomain. This was a strong indication that the two domains evolved separately and that they were brought together in a subset of genes by sequence shuffling. Moreover the paired domains of *pox meso* and *pox neuro* deviate significantly at position characteristically conserved in the *prd*, *gsb*-paired domain and hence represent separate types of paired domains. These two secondarily discovered genes are no segmentation genes, but rather tissue specific transcription factors, presumably acting further downstream in the gene regulatory cascade to which *prd* and the two *gsb* genes belong. The *pox meso* expression was mainly observed in the mesodermal germ layer in the posterior half of each segment, while *pox*

neuro is expressed in a segmental repeated pattern in neural precursors of the peripheral as well as central nervous system (Bopp et al. 1989)

The paired domain was later proven to represent a DNA binding domain. It is composed of two helix-turn-helix subdomains, the N-terminal subdomain (also called PAI) and the C-terminal subdomain (also called RED). Both subdomains can bind to DNA independently, but the main DNA binding activity resides in the N-terminus (Czerny et al. 1993). Thanks to their sequence-specific DNA-binding activity, the paired-domain containing proteins can be involved in transcription regulation processes and therefore play important roles as transcription factors during development (Treisman et al. 1991).

2.3. Paired-box genes in other organisms

A major stimulus in developmental biology has been the discovery that the genes controlling morphogenesis in the fruitfly *Drosophila melanogaster* are conserved in many evolutionarily distant species. The best-studied type of conserved sequence is the homeobox, present in segmentation and homeotic genes of *Drosophila* as well as in many developmental control genes of vertebrates (Krumlauf 1994). Another type of conserved sequence is typical of genes with Zn-finger repeats (Dressler and Gruss 1988). Similarly a set of paired box containing genes was identified in several organisms, including vertebrates and lower metazoans (Dressler et al. 1988; Burri et al. 1989), by homology search with a *Drosophila* paired box probe. In such a way a whole multigene family was isolated in the mouse and when the first mouse paired box containing gene was described (*Pax1*), a remarkable sequence homology with *Drosophila paired* both at the nucleic acid and at the protein level was observed (Deutsch et al. 1988; Walther et al. 1991).

Both in mice and man nine different members of this gene family have been isolated and have been named Pax genes (*Pa*ired-bo*x* genes). These genes are all expressed during embryogenesis and play important roles in patterning and organogenesis (Dahl et al. 1997). Mouse and human Pax genes have been classified into four paralogous groups, which share a specific assembly of two additional structural motifs, other than the paired domain, the octapeptide and the homeodomain. The first group (*Pax1* and *Pax9*) is characterized by the presence of the octapeptide and the absence of the homeodomain, in the second group (*Pax2*, *Pax5* and *Pax8*) only part of the homeodomain is maintained, in

the third (*Pax3* and *Pax7*) both motifs are entirely present and the fourth group (*Pax4* and *Pax6*) lacks the octapeptide and bares only the homeodomain (Fig. 1). Genes within an individual group show very high degree of similarity within the paired domain and a similar expression pattern during embryogenesis (Balczarek et al. 1997; Dahl et al. 1997).

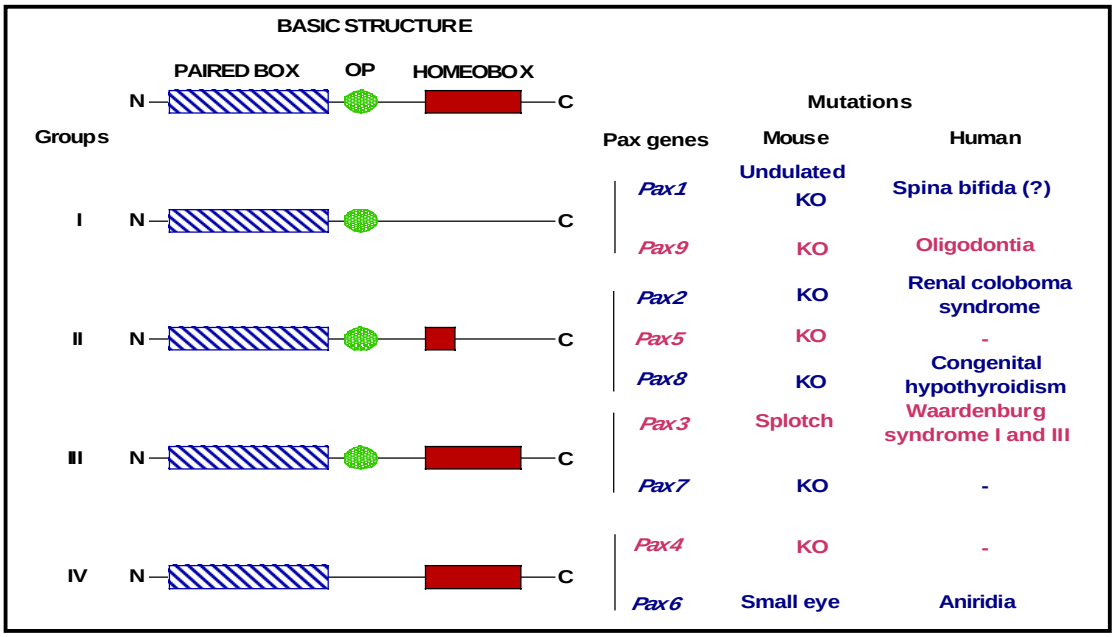


Fig.1 Vertebrate paired-box genes. Group subdivision, targeted (KO) or spontaneous mouse mutations and associated human diseases. OP: octapeptide

If the *Drosophila* paired-box genes are included in this subdivision, sequence analyses suggest that each of the vertebrate Pax groups contains at least one *Drosophila* gene; *pox-meso* can be included in the first group, the segmentation genes *prd*, *gsb-p* and *gsb-d* fall in the third group and the *Drosophila eyeless* gene was found to be the direct orthologue of *Pax6* from the fourth group. Only in the case of the second group the suggested relation with *Drosophila pox-neuro* is not supported by all the authors or by strong sequence analysis data, so that *pox-neuro* has been rather considered as a member of an independent fifth subgroup (Noll 1993; Balczarek et al. 1997; Breitling and Gerber 2000; Galliot and Miller 2000). Considering the degree of homology among the known paired domains of vertebrates and insects, it is clear that at the time of the separation of deuterostomes from protostomes at least two and perhaps as many as four/five different ancestral paired-box genes existed (Noll 1993). The scenario got even more complicated when pax homologues were cloned from lower chordates, like ascidians (Glardon et al.

1997; Wada et al. 1997; Wada et al. 1998; Ogasawara et al. 1999) and amphioxus (Holland et al. 1995; Glardon et al. 1998; Holland et al. 1999; Krelova et al. 2002), and even more distant organisms, namely nematodes, cnidarians, and sponges (Sun et al. 1997; Hoshiyama et al. 1998; Hobert and Ruvkun 1999; Miller et al. 2000). Evolutionary trees were constructed on the basis of the paired domain conservation in the different animal groups and they could explain the acquisition or loss of the other domains during evolution. For example, the prd-type homeodomain, totally or partially present in groups II, III, and IV, was either combined to the paired domain in at least two independent events (that respectively originated the pax subgroup II in one case and the two subgroups III and IV in the other) after the diversification of the paired-domain genes or it was captured in one single event and then subjected to various rounds of modifications during the gene diversification leading to its partial or total loss (respectively in the II and I subgroups) (Noll 1993; Breitling and Gerber 2000).

2.4. Role of Pax genes in Vertebrates

During development, Pax genes are expressed in a highly specific spatial and temporal pattern; they act in early and crucial steps of the generation of a number of organs. The analysis of mouse mutants and human syndromes has uncovered their important role as regulators of normal organ development. Two types of events might be under the control of Pax genes during organogenesis. One is the signal transduction at the interface of epithelium and mesenchyme, where many organs develop (Dahl et al. 1997; Mansouri et al. 1999). The other one is cell proliferation (Dahl et al. 1997). In accordance with a role in proliferation, it has been observed that abnormal expression of pax genes in humans is often associated with tumorigenesis. In particular the pax genes exert their oncogenic potential specifically in the tissues and organs, where they are normally required during development (Dressler and Douglass 1992; Galili et al. 1993; Kozmik et al. 1995; Kroll et al. 2000).

Pax1 belongs together with *Pax9* to the first group of vertebrate Pax genes. It was the first paired-domain containing gene to be identified in a vertebrate genome through homology search with a paired-box probe from the *Drosophila prd* gene. *In situ* hybridizations on developing mouse embryos have shown its main expression domains;

Pax1 is expressed in a segmented pattern in the caudal half of the somites (Deutsch et al. 1988). The somites are the metameric embryonic structures that originate from segmentation of the paraxial mesoderm and consist of epithelial spheres of cells that bud off in anterior-posterior direction flanking on both sides the notochord and the neural tube. Somites later differentiate to give rise dorsally to the dermomyotome, which will yield the skeletal muscles and the dorsal dermis, and ventrally to the sclerotome, which will form the vertebral column (Gossler and Hrabe de Angelis 1998). *Pax1* is expressed in the portion of the somites differentiating into sclerotome and more precisely in that subset of sclerotomal cells which will surround the notochord and give rise to the ventral body of the vertebrae and to the intervertebral discs (Deutsch et al. 1988). Additionally *Pax1* is expressed in the proximal region of the developing limbs at the limb-trunk joint level (Timmons et al. 1994) and in the endoderm of the third and fourth pharyngeal pouches (Wallin et al. 1996). The pharyngeal pouches are metameric structures that form caudally to the head region upon evagination of the endoderm and invagination of the overlying ectoderm.

The role of *Pax1* in the development of the structures where it is expressed became clear, when a point mutation in this gene was associated to a recessive mouse mutant, *undulated* (*un*), which exhibited distortions along the entire vertebral column as well as in the sternum (Balling et al. 1988). Two additional natural mouse mutants of *Pax1*, *undulated-extensive* (un^{ex}) and *Undulated short-tail* (Un^{s}), characterized by a deletion of the last exon of *Pax1* and the whole gene respectively, show similar abnormalities, even though to more extended degrees. In correspondence to each of the three different *Pax1* mutations, the phenotype ranges in its severity from a malformation of the central vertebral structures (vertebral bodies and intervertebral discs) in the mildest case (*un*) to their complete absence in the most severe case (Un^{s}) and it is more pronounced in the lumbar region and in the tail than in the rest of the axial skeleton (Wallin et al. 1994). Moreover other skeletal structures are as well affected, such as the pectoral and pelvic girdles (Timmons et al. 1994) and the thymus, a derivative of the pharyngeal pouches, which is significantly reduced in size and impaired in its function (Wallin et al. 1996). Targeted inactivation of *Pax1* has confirmed the observations made on the natural mutants and the role of the gene in the normal development of these organs (Wilm et al.

1998). A possible connection of human *PAX1* to a form of spina bifida, a malformation characterized by incomplete closure of the neural tube, has been suggested after the finding of an aminoacid substitution in an affected patient (Hol et al. 1996).

A brief description of the other pax genes and of their fundamental roles in development will follow (reviewed in Dahl et al. 1997; Mansouri et al. 1999; Chi and Epstein 2002).

The *Pax2* gene is expressed in the developing mouse kidney and ureter as well as in the optic stalk, the ear, the midbrain-hindbrain junction, and the spinal cord. It participates together with the Wilms tumor 1 gene (*Wt1*) to an important molecular pathway regulating the formation of metanephrons, which differentiate to form the functional kidney in mammals. Loss of *Pax2* in mice results in severe urogenital defects including absence of kidney, ureter and genital organs in addition to ophthalmologic and inner ear defects. In humans, haploinsufficiency of *PAX2* leads to the renal coloboma syndrome, an autosomal dominant disease characterized by renal and ocular defects (Dressler and Woolf 1999). Moreover *Pax2* is absolutely required to maintain the mid-hindbrain region and its misexpression affects the development of deriving structures, like the cerebellum. On the contrary *Pax5*, which is also expressed early in the mid-hindbrain junction, does not seem to be strictly necessary for the maintenance of this structure, since mutant mice show quite a mild brain phenotype. Its main function is rather to be studied in the differentiation of lymphoid precursors to B-cells. Lack of *Pax5* results indeed in a complete block of B-cell maturation (Nutt et al. 2001). *Pax8* is expressed in the developing excretory system and in the thyroid gland. The role of *Pax8* in the formation of the thyroid was clearly observed in the *Pax8* homozygous deficient mice. Also in humans a heterozygote mutation in *PAX8* has been associated to hypothyroidism. The gene is not only essential for the development of the thyroid, but also for its function, by regulating the transcription of thyroid specific genes coding for thyroglobulin and thyroperoxidase (Damante et al. 2001).

Another gene with a pleiotropic function is *Pax3*. The *Splotch* mutant mice, which harbor mutations in this gene, show a wide phenotypic spectrum including neural tube defects, congenital heart disease and coat color defects. In humans, mutations of *PAX3* result in a form of the Waanderburg syndrome characterized by pigmentary disturbance of the iris, hair and skin and hearing problems. In both human and mouse, *Pax3* mutations affect

tissues that receive contributions by neural crest cells, which emerge during embryonic development from the dorsal neural tube where the gene is expressed. *Pax3* is also expressed in the dermomyotome, the dorsal-lateral domain of the elongating somites, and regulates the formation of deriving tissues, such as the muscles of the body wall and of the limbs. The paralogous gene *Pax7* has a fairly overlapping expression pattern with *Pax3* in the dermomyotome, but its function is more focused to the specification and maintenance of satellite cells in the adult muscle, whose function is to differentiate into functional myocytes during healing processes (Mansouri 1998).

The last group includes *Pax4* and *Pax6*. *Pax4* expression is restricted in the developing endocrine pancreas, where it contributes to the differentiation of insulin-producing β-cells and somatostatin-producing δ-cells (Dohrmann et al. 2000). *Pax6* is perhaps the most studied pax gene. Apart from a complementary role in the development of pancreas, mainly aimed to the formation of glucagon-producing α-cells and organization of endocrine cells into proper spherical islets (Dohrmann et al. 2000), this gene has a fundamental and evolutionarily conserved role in eye development. *Pax6* is the only pax gene with real homologous counterparts in invertebrates like the *Drosophila* gene *eyeless*, which acts as a master regulator of eye formation. In vertebrates *Pax6* is expressed in the forming optic cup and in the overlying ectoderm that will from the lens. Heterozygous *PAX6* mutations in humans result in a variety of eye diseases including blindness, aniridia, colobomas and cataracts (Ashery-Padan and Gruss 2001). The spontaneous *Small eye* mouse mutants completely lack mature ocular structures in their homozygous form. Furthermore defects in the *Pax6* function have been proven to cause failure of nasal development in mice, while brain malformations, related to its expression in the developing central nervous system, have been observed both in mice and in some human mutants (Dahl et al. 1997).

2.5. The *Pax9* gene

2.5.1. Isolation and expression pattern

After the identification of eight vertebrate pax genes, *Pax1* still had no direct paralogous counterpart. Thus the assumption that every subgroup of the pax gene family contains at least two members promoted the search for a new gene related to *Pax1*. PCR amplification with degenerated primers from the genome of *Pax1 Un*^{S} mice and hybridization with the PCR product on genomic DNA brought about indeed the discovery of a new gene (*Pax9*) that mapped on mouse chromosome 12 and had a high sequence similarity with *Pax1* (Wallin et al. 1993). At the same time the human *PAX9* gene was isolated in a similar way and assigned to chromosome 14 (Stapleton et al. 1993).

The isolation and sequencing of a *Pax9* clone from a mouse embryo cDNA library allowed the complete sequence alignment with *Pax1* showing an extraordinary high homology even in some regions outside of the paired domain with an overall estimation of about 80% similarity and 66% identity between the two gene products (Neubüser et al. 1995).

A remarkable homology between the two genes was also observed at the expression level. The *Pax9* transcript is detectable during mouse embryogenesis from 8.5 days post coitus (dpc) until around day 15.5. *Pax9* starts to be expressed in the endoderm of the four pharyngeal pouches, partly overlapping with *Pax1* expression, and it remains active in these structures at least until they further develop in their derivatives. Like *Pax1*, the largest *Pax9* expression domain is the sclerotome portion of the caudal half of each single somite; however *Pax9* expression appears slightly later at around day 9 pc, not before de-epithelialization of the sclerotome has occurred. Moreover, while *Pax1* is expressed in the sclerotomal cells directly surrounding the notochord and mainly ventrally located, *Pax9* expression is more dorso-lateral in the area of the sclerotome that will give rise to the vertebral processes as neural arches and proximal ribs. Medially, *Pax9* expression is weaker but also concentrated in the condensed mesenchyme that will give rise to the intervertebral discs. As the mesenchymal anlagen of the vertebrae start differentiating into the chondrocyte lineage, *Pax9* expression progressively reduces, remaining only in the residual mesenchymal tissue surrounding the primordia of the developing vertebrae and ribs. Starting from around day 11.5 pc, *Pax9* is expressed at high levels in the

developing limb buds. Expression peaks in the mesenchyme at the anterior proximal corner of the hand and foot plate and extends from the dorsal to the ventral side of the limb bud. Later on *Pax9* expression extends more posteriorly and restricts more ventrally. It marks the mesenchyme along the forming radius and tibia, in the fore- and hindlimbs respectively, and the dermal thickening of the footpads. Additionally it appears in more distal regions at later stages in part of the mesenchyme of the developing digits. (Neubüser et al. 1995; Peters et al. 1998b).

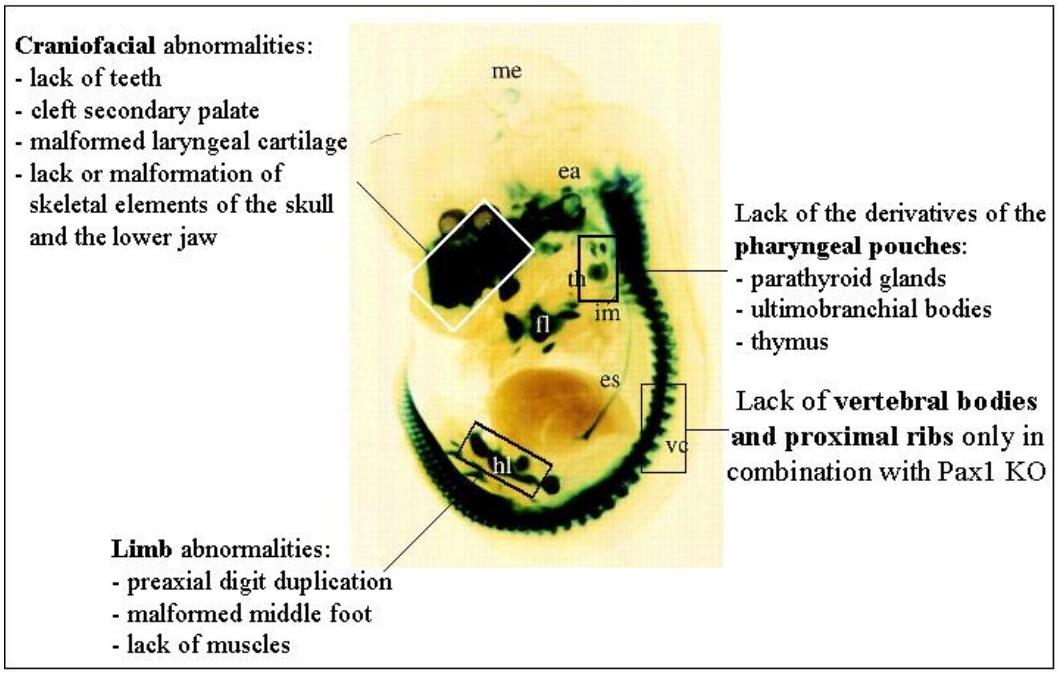


Fig. 2. *Pax9* expression pattern and knock-out phenotype. X-gal staining of a 13.5 dpc *Pax9*^{lacZ} mouse embryo, showing the distribution of the *Pax9* expression domains. The text describes the phenotype of the *Pax9* knock-out relatively to the boxed structures. Modified from Peters et al. 1998b.

Pax9 shows also some specific expression domains, which are not found for *Pax1*. One of these is the craniofacial area, where *Pax9* expression pattern is extremely complex and varies throughout development. The facial mesenchyme of neural crest cell origin that will differentiate in many of the facial bones of the nose, maxillary and mandibular regions starts to express *Pax9* from day 10.5 post coitus on. Strong *Pax9* expression is detectable in the medial and lateral nasal processes and in the mesenchyme between the olfactory epithelium and the external walls of the nasal capsule. *Pax9* mRNA is later

detectable in the mesenchyme of the incisors and molars as well as at the base of the developing skull. The gene is furthermore transcribed in the foregut and oral epithelium and in the hindgut at the tip of the tail, where it later also marks the mesenchyme surrounding the muscle primordia (Neubüser et al. 1995; Peters et al. 1998b).
In the adult mice *Pax9* mRNA was observed in thymus, even though at very low doses (Neubüser et al. 1995); a much higher expression level was detected in other tissues, such as the esophagus, the tongue and the salivary glands, and in the adult human esophagus (Peters et al. 1997; Peters et al. 1998b).

2.5.2. The Pax9 knock-out mouse

Since no natural *Pax9* mutant was known, the first evidence about the role of the gene came with the creation of a knockout mouse model (Peters et al. 1998b).
The *Pax9* deficient mice show no apparent phenotype at the heterozygous state; conversely the homozygous mice die soon after birth due to impairment of respiration. An inspection of the skull revealed that all mutants have a cleft secondary palate. The aberrant morphology of the palatal shelves is likely to account for the defect in canalizing the air into the respiratory ways and therefore causing suffocation. This provided a first clue of the role of *Pax9* in the development of the facial skeleton. A closer examination of *Pax9* mutants showed a larger number of affected bones, including several elements of both jaws and of the base of the skull. These malformations extend also to the cartilage elements of the larynx and the thyroid, even though no *Pax9* expression is detected in these structures during development. Further skeletal defects can be observed at the level of the limbs. Homozygous *Pax9* mutants develop preaxial duplications in both fore- and hindlimbs, which result only in the hindlimbs in a small supernumerary toe (Peters et al. 1998b).
Surprisingly, despite the evident expression of *Pax9* in the developing vertebral column and the asserted role of the homologous gene *Pax1* at this level, no abnormal phenotype was observed in the vertebral column of the $Pax9^{-/-}$ mice. That *Pax9* has indeed a role in the development of the axial skeleton could only be proven through generation of *Pax1*/*Pax9* double knockout mice. In these mice the vertebral malformations already described in the $Pax1^{-/-}$ condition (Wilm et al. 1998) showed a much greater severity. No

vertebral bodies or intervertebral discs whatsoever are formed, thereby dramatically reducing the overall length of the body axis. Furthermore, the proximal parts of most ribs and all skeletal elements of the tail are missing. These results are a strong indication for the synergistic role of *Pax1* and *Pax9* in vertebral column development and for their functional redundancy, that allows compensation from one gene in the absence of the other one, partially in the *Pax1*$^{-/-}$ state or totally in the *Pax9*$^{-/-}$ state. Besides, intermediate genotypic conditions, going from wild-type to *Pax1*$^{-/-}$/*Pax9*$^{-/-}$ and passing through all the series of allelic combinations, correspond to intermediate phenotypes in the vertebral column defects, suggesting a dosage dependent co-operation of the two genes (Peters et al. 1999).

Apart from the skeletal defects *Pax9*$^{-/-}$ mice display lack of the derivatives of the third and fourth pharyngeal pouches, such as parathyroid glands and ultimobranchial bodies; in contrast derivatives of the first and second pharyngeal pouches appear unaffected (Peters et al. 1998b). The thymus, whose epithelial portion develops from the third pharyngeal pouch, appears severely affected in size in *Pax9*$^{-/-}$ mice. The thymic rudiment reaches a certain point of maturation, but it arrests abruptly showing impairment in the thymopoiesis (Hetzer-Egger et al. 2002).

Another interesting aspect of the *Pax9* deficient mice is the total absence of teeth in both jaws in accordance to previous observations of *Pax9* expression in the mesenchymal compartment of the developing teeth (Neubüser et al. 1995). A closer investigation in mutant embryos showed that *Pax9* is essential for tooth development to proceed beyond the epithelial bud stage. *Pax9* is required to maintain the BMP signaling from the epithelium that will promote the expression of downstream genes (like *Msx1* and *Lef1*) responsible for earlier events of tooth formation (Peters et al. 1998a; Peters et al. 1998b).

The *Pax9* function in tooth development has become a topic of great impact since the finding that some human patients affected by a form of oligodontia carried a mutation in the *PAX9* gene. Oligodontia is the agenesis of six or more teeth without other associated disorders. A heterozygous point mutation in *PAX9*, causing a frameshift and therefore abnormal protein synthesis, proved to be enough to determine a mutant dental phenotype, first evidence of the haploinsufficiency of this gene (Stockton et al. 2000). Since then other *PAX9* mutations associated to oligodontia have been described, always in a

heterozygous form and resulting in abnormal or truncated protein synthesis (Nieminen et al. 2001; Frazier-Bowers et al. 2002) or consisting of a large genomic deletion including the whole gene (Das et al. 2002). This finding places *PAX9* among the pax genes with a known role in human diseases.

2.5.3. Comparative analysis of Pax9 in other species

The function of the *Pax9* gene is most probably conserved also outside of the mammalian class. This was demonstrated upon *Pax9* isolation from other non-mammalian vertebrates. *Pax9* shows in the chick an expression pattern very closely resembling the situation in the mouse. It starts in the developing foregut pocket that will give rise to the pharyngeal pouches, where the gene still remains expressed (Peters et al. 1995; Müller et al. 1996). At a later stage *Pax9* becomes also visible in the sclerotome of the somites, even though to a much lower temporal and spatial extent than *Pax1*, being absent in early sclerotomal cells at the caudal end of the embryo (Peters et al. 1995; Müller et al. 1996). *Pax9* transcripts were also detected in an anterior-proximal and anterior-distal domain of the developing limb buds (Müller et al. 1996) and later on in the metatarsal mesenchyme (Peters et al. 1995). Like in the mouse, chick *Pax9* is expressed in distinct areas of the developing olfactory organ (Peters et al. 1995) and to a lower extent in the mesenchyme of the mandible (Chen et al. 2000). Expression of *Pax9* in the esophagus and the thymus of the adult chick resembles again the situation in the mouse (Peters et al. 1995). This comparative analysis of the gene in mouse and chick suggests that its developmental role has maintained fairly conserved in higher vertebrates.

Two transcript isoforms of *Pax9* were found in zebrafish both with the same expression pattern, the ventral part of the somites, corresponding to the sclerotome, and two anterior stripes underlying laterally the hindbrain on both sides, described as facial mesenchyme (Nornes et al. 1996), but probably coinciding with the originating pharyngeal endoderm (see Results section).

Studies in the lower chordates, like amphioxus (*Branchiostoma lanceolatum*) and the two ascidian species *Halocynthia roretzi* and *Ciona intestinalis*, and in one hemichordate species (the acorn worm *Ptychodera flava*), brought about the identification of only one *Pax1/9* related gene, supporting the hypothesis that the two genes derived from

duplication of a common ancestor in the vertebrate lineage (see Discussion). Remarkably the *Pax1/9* gene of these organisms is mainly or exclusively expressed in the endoderm of the pharyngeal gills. In all cases the expression appears rather late in development and persists in the pharyngeal gills of the adult animal (Holland et al. 1995; Ogasawara et al. 1999). This suggests that originally *Pax1* and *Pax9* might have been connected to the development of these structures and only later they acquired the known expression in other domains. Significantly, the expression of *Pax9* in lampreys, that represent an intermediate animal species between upper vertebrates and lower chordates, is also restricted to the pharyngeal endoderm that will form the gills. A weak expression is already visible in some anterior derivatives of neural crest cell origin, prelude of the facial mesenchyme expression in higher vertebrates, but no *Pax9* transcript is detected in the somites (Ogasawara et al. 2000).

2.5.4. Pax9 regulation

In order to better understand the molecular pathways in which *Pax9* takes part and the role of the gene in the development of the structures where it is expressed, it is important to study the upstream events that lead to the tissue specific transcription of *Pax9*.

Some considerations about *Pax9* regulation in the sclerotome have been already suggested upon observations made both in the chick and in the mouse.

The mouse mutant *Danforth's short tail* (*Sd*) is a skeleton mutant in which the notochord is affected, such that in the cervical and thoracic region a notochord primarily forms but subsequently degenerates. As a result the vertebrae present various morphological defects (Dunn et al. 1940). It had been previously shown that *Pax1* expression in the sclerotome appears extremely reduced in the *Sd* homozygous mice stopping abruptly in the thoracic region and consequently completely missing in the lumbar, sacral and caudal regions (Koseki et al. 1993). Similar results were observed for *Pax9*. In *Sd* homozygotes *Pax9* expression is detectable up to the thoracic level but not in the lumbar, sacral, or caudal regions, while the other expression domains appear unaffected (Neubüser et al. 1995). These observations suggest that expression of *Pax1* and *Pax9* in sclerotome depend on signals from the notochord. A similar notochord dependent expression was observed in the chick. Surgical removal of the notochord from chick embryos, that were subsequently

allowed to further develop, resulted in disruption of the somite morphology and loss of *Pax9* transcript in the cells lying directly beneath the neural tube where the sclerotome is normally located. Those cells had switched to a dermomyotomal fate and showed ectopic expression of related markers, like *Pax3* and *Pax7*. Vice versa when notochord grafts were applied in ectopic positions, an increase in *Pax9* expression was observed adjacent to the ectopic notochord, where myotome would normally be present, and the cells would rather differentiate into the sclerotomal lineage (Goulding et al. 1994).

The extracellular factor Sonic hedgehog (Shh) produced by the notochord is considered the main mediator and activator of the signaling cascade that exerts this transcriptional control. Graft experiments with SHH-expressing cells were able to mimic the effect of notochord grafts dorsally and laterally to the neural tube, causing an enlargement of the *Pax1* positive area and inducing as a consequence overdevelopment of cartilage (Watanabe et al. 1998). Conversely, signals coming from the lateral plate mesoderm strongly down-regulated both *Pax1* and *Pax9* expression in chick explant cultures (Müller et al. 1996). It has been suggested that the transcription factor Uncx4.1 might be involved in the regulatory pathway, since mouse defective of the corresponding gene display *Pax9* downregulation in the caudal half of the sclerotome (Leitges et al. 2000; Mansouri et al. 2000).

In the jaw mesenchyme *Pax9* expression marks the sites of tooth formation (Neubüser et al. 1997). Members of the Fibroblast Growth Factor family (FGFs) and of the Bone Morphogenetic Protein family (BMPs) secreted from the overlying ectoderm determine the location of *Pax9* activation in an antagonistic manner. Experiments on cultured explants of the mandibular arch of mouse embryos at around E10.5 showed induction of *Pax9* expression upon application of FGF8-soaked beads, as FGF8 is known to be expressed in the prospective dental ectoderm. Conversely, when BMP2- or BMP4-soaked beads were applied *Pax9* expression was inhibited. These results, together with the observation of the *Fgf8*, *Bmp2* and *Bmp4* expression patterns in the mandibular ectoderm, elucidate the mechanism of *Pax9* regulation in the tooth mesenchyme (Neubüser et al. 1997). These signal factors are however only required for the initial induction of *Pax9* and not for its maintenance in later stages (from E11.5 on) when *Pax9* expression is already established and probably self-sustaining (Neubüser et al. 1997; Mandler and

Neubüser 2001). Interestingly, *Pax9* transcription in these cells appears independent of Shh signaling, which is also arising from the mandibular ectoderm (Dassule et al. 2000), suggesting a tissue-specific competence of the gene to respond to particular signaling factors. However it cannot be ruled out that Shh may activate at this level a different molecular pathway than the one acting during the sclerotome induction.

The signaling function of the adjacent tissues on the induction of *Pax9* expression seems to be a common regulatory mechanism. However this is not a general situation. By separating axial and lateral parts of the prospective chick foregut region or by grafting prospective pharyngeal endoderm into different parts of the developing chick embryo, no change in *Pax1* and *Pax9* expression was observed, suggesting that activation of the two genes in the endoderm is rather intrinsically regulated and very early determined (Müller et al. 1996)

2.6. Aim of the work

The present work collocates within the attempt of determining which molecular factors directly regulate *Pax9* tissue-specific during mouse embryogenesis.

I concentrated my work at the DNA level, first establishing the *Pax9* genomic structure and then searching for sequence elements that drive the embryonic expression of the gene, believing that the identification of such *cis*-regulatory elements can be of extremely high impact for the identification of direct binding factors. In order to do that, I used two different approaches. One consists of a comparative sequencing of the *Pax9* genomic region from different species, searching for short conserved non-coding sequences that can be taken as best candidates for regulatory elements. The other one is based on the creation of a transgenic mouse model using large genomic regions encompassing the *Pax9* locus, in order to reproduce the entire gene expression pattern and subsequently narrow down the single functional elements.

3. MATERIALS AND METHODS

3.1. Materials

In this work the following materials were used

E. coli bacterial strains

DH5α (Gibco BRL):	F^- Φ80dlacZΔM15 Δ(lacZYA-argF) U169 deoR recA1 endA1 hsdR17 (r_k^-, m_k^+) phoA supE44 λ^- thi-1 gyrA96 relA1
DH10b (Gibco BRL):	F^- mcrA Δ(mrr-hsdRMS-mcrBC) Φ80dlacZΔM15 ΔlacX74 deoR recA1 endA1 araΔ139 Δ(ara, leu)7697 galU galK λ^- rpsL nupG λ^- tonA
DM1 (Gibco BRL)	F^- dam^-13::Tn9(Cm^R) dcm^-mcrB $hsdR^-M^+$ gal1 gal2 ara^- lac^- thr^- leu^- ton^R tsx^R su^0
TOP10 (Invitrogen):	F^- mcrA Δ(mrr-hsdRMS-mcrBC) Φ80dlacZΔM15 ΔlacX74 deoR recA1 araΔ139 Δ(ara, leu)7697 galU galK λ^- rpsL (Str^R) endA1 nupG

Commercial vectors

pBluescript® II KS+ (pBSKS) (Stratagene) was used for most of the cloning procedures.

pCR2.1®-TOPO® (Invitrogen) and pCR II®-TOPO® (Invitrogen) were used for directly cloning of most of the PCR products if no cloning ends were added to the primers.

pEGFP-C1 (Clontech) was used for subcloning of the EGFP coding sequence.

pIRES-EGFP (Clontech) was used for subcloning of the IRES-EGFP cassette.

pGL3-Basic Vector, pGL3-Promoter Vector, pGL3-Control Vector and pRL-SV40 (Promega) were used for the luciferase assay in cell culture.

Plasmids and vectors obtained from other people/ groups

pcPax9-k5 containing the 2400 bp cDNA sequence of mouse *Pax9* (Neubüser et al. 1995), available in our own lab.

pcPax9-WM containing a 1370 bp EcoRI-SspI fragment of mouse Pax9 cDNA (Neubüser et al. 1995), available in our own lab.

pP9paired containing the mouse *Pax9* paired domain cloned in pCR2.1TOPO, obtained from J. Gerber, IEG, GSF.

pzPax9a containing the 2 kb zebrafish Pax9a cDNA sequence, obtained from Dr. Terje Johansen, University of Tromsø, Norway (Nornes et al. 1996).

pGT1.8IresβgeoA containing the IRESβgeo cassette, obtained from Dr. K. Araki, Kumamoto University, Japan.

pSV1.RecA for cloning of targeting cassette and expression of the *recA* gene in the RecA mediated BAC modification method, obtained from Dr. Yang, The Rockefeller University, New York (Yang et al. 1997).

pASShsp68lacZpA containing the lacZ gene under the *hsp68* basal promoter control for generation of constructs for conventional transgenesis, obtained from Dr. H. Sasaki, Osaka University (Sasaki and Hogan 1996).

pGETrec for expression of the arabinose inducible recET recombination machinery for the ET cloning method, obtained from Dr. PA Ioannou, The Murdoch Institute for Research into Birth Defects, Royal Children's Hospital, Melbourne (Narayanan et al. 1999).

pGK-FRT containing the kanamycin resistance gene flanked by FRT sites for preparation of targeting cassettes and 706-pMJ-tet for the bacterial expression of FLIP-recombinase both used for the ET cloning method, obtained from Dr. F. Stewart, EMBL Heidelberg

pzhsp70-nβgal for the synthesis of a *lacZ in situ* probe, obtained from Dr. Laure Bally-Cuif, GSF-ISG

DNA libraries

RPCI – 23 Female (C57BL/6J) Mouse BAC Library, constructed in Peter deJong's lab at the Roswell Park Cancer Institute, consisting of partially EcoRI-digested DNA cloned into the EcoRI site of pBACe3.6 vector. 11.2 fold mouse genome coverage. Host E. coli strain DH10b. Supplied by Research Genetics.

Fugu cosmid library no. 66, constructed by Carola Burgtorf, containing MboI partial digests of *Fugu rubripes* genomic DNA cloned into Lawrist4. Host *E. coli* strain DH10b. Supplied by Resource Center / Primary Database of the German Human Genome Project.
Mouse full-length cDNA library, constructed by the Genome Exploration Research Group, Genomic Science Center, Genome Science Laboratory, Tsukuba Life Science Center, The Institute of Physical and Chemical Research, Riken, Japan (Bono et al. 2002).

Cell lines

NIH3T3:	mouse transformed embryo fibroblasts
MLB13 myc clone 14:	skeletal progenitor cell line derived from 13-dpc mouse embryo limb buds (Rosen et al. 1994)
AT478:	mouse squamous cell carcinoma (Guttenberger et al. 1990)

3.2. Molecular biology methods

3.2.1. Plasmid DNA preparation

E. coli cells containing plasmid DNA were usually grown, if not differently specified, in autoclave sterilized LB-medium (10 g bacto-tryptone, 5 g yeast extract, 10 g NaCl in 1 l H_20) with a selective specific antibiotic, ampicillin (100 µg/ml) or kanamycin (30 µg/ml), over-night at 37°C. Medium (25 ml culture) and large (100 ml culture) scale preparations of plasmid DNA were carried out by means of the respectively Plasmid Midi- and Plasmid Maxi-Kit from QIAGEN, according to the provided enclosed protocol. Elution from the column was performed with water.
In case of low copy number plasmids, like pSV1.RecA and 706-pMJ-tet, the following modifications were applied:
- growth in 100 ml (midi) or in 500 ml (maxi) LB + 10 µg/ml tetracycline over-night at 30°C;
- elution from the QIAGEN column with QE buffer, previously warmed up to 60°C.
Small scale preparations (minipreps) were realized with the QIAGEN QiaPrep Mini-Kit from a 5 ml culture, in case DNA was needed for downstream applications, like sequencing or further cloning steps.

For mere colony screening minipreps, 2 ml cultures were processed according to Birnboim and Doly (Birnboim and Doly 1979) and DNA was dissolved in 50 µl TER (10 mM TRIS·HCl pH 7.5, 1 mM EDTA, 10 µg/ml RNase A). This method was as well applied for small scale preparation of cosmid DNA.

3.2.2. BAC DNA preparation

Large scale preparations were executed starting from 500 ml LB cultures with 12.5 µg/ml chloramphenicol in over-night growth at 37°C and bacterial cells were then processed with the Large Construct Kit from QIAGEN or the Nucleobond BAC 100 Kit from Machelerey-Nagel.

In order to increase the yield of the preparation, a twice or three times as big bacterial culture was inoculated and the final purified DNA samples were pooled together.

Minipreps of BAC DNA were performed with the normal alkaline lysis method from Birnboim and Doly. The only relevant modification to the protocol was the addition of 450 µl of 5M LiCl after solution III in order to facilitate protein precipitation.

During preparation shearing of BAC DNA was mineralized by avoiding vortexing and vigorous mixing and pipetting.

BAC DNA was stored in TE pH8 at 4°C.

3.2.3. Genomic DNA preparation

Mouse genomic DNA was extracted from tail tips or from yolk sacs of respectively adult mice (at least 3 weeks old) and mouse embryos (from 10 dpc up to 14 dpc). Tissue samples were incubated in an appropriate volume of lysis buffer (50 mM KCl, 10 mM Tris·HCl pH8.3, 0.1 mg/ml gelatin, 0.45% Nonidet NP-40, 0.45% Tween20) rendered 0.5 mg/ml proteinase K shaking overnight at 65°C. For PCR applications, the treatment was followed by 10 minutes at 95°C and quick spinning down of the debris; 1 µl of crude sample was used for every single PCR reaction.

For Southern-blot analysis, phenol/chloroform/isoamylic alcohol extraction and ethanol (EtOH) precipitation followed the overnight incubation. DNA pellets were washed once with 70% EtOH and redissolved in 50 µl TE pH8.

Alternatively, tissue samples were incubated in „Tail buffer“ (50 mM TRIS·HCl pH8, 50 mM EDTA, 100 mM NaCl, 0.5% SDS) freshly supplemented with 500 µg/ml proteinase K, shaking overnight at 55°C and DNA was isolated by two phenol extraction steps and EtOH precipitation as above. This second method was only used to obtain cleaner DNA for Southern-blot analysis.

3.2.4. Restriction digest of DNA samples

Restriction digestion of DNA was performed for screening of plasmid clones and related orientation analysis, for Southern blot analysis of BAC, cosmid or genomic DNA, and for isolation and preparation of DNA fragment in cloning procedures. Restriction enzymes from the following suppliers were used, Gibco BRL, Roche, New England Biolab. Enzyme units to use were empirically determined for each reaction and working buffers were chosen in accordance to the information provided by the suppliers. Incubations took place at 37°C, if not differently specified, for a minimum time of 30 minutes up to overnight.

3.2.5. DNA Gel electrophoresis

Conventional gel electrophoresis for separation of DNA molecules in the range of 100 bp – 20 kb was usually performed on 1% agarose gels (Ultra Pure Agarose, Gibco BRL) in 1x TAE buffer (40 mM Tris-acetate: 1 mM EDTA). The agarose percentage was otherwise adjusted between 0.7% and 2% according to the desired separation range for specific purposes as described in the single cases. For separation of DNA molecules between 80 bp and 200 bp, 3% agarose gels were performed using MetaPhore® agarose (BMA). Gel run was performed with variable time and volt conditions according to the separation range and agarose percentage using an electrophoresis power supply (Consort). For size comparison, a DNA molecular weight marker was loaded on gel next to the samples (SmartLadder, Eurogentec, or 100bp-ladder, Gibco BRL). DNA was stained with the intercalating fluorescent reagent ethidium bromide (EtBr), which was added either in the gel before solidification or in TAE buffer for after-run staining at the concentration of 0.5 µg/ml. Stained DNA was visualized on a UV-transilluminator at a wavelength of 254 nm and photographed with a gel documentation apparatus (Herolab).

Pulse-field gel electrophoresis (PFGE)

PFGE was performed for separation of high molecular weight DNA molecules (20-200 kb) derived from restriction digestion of BAC DNA. SeaKem-LE Agarose (FMC Bioproducts) or peqGOLD Pulsed Field Agarose (peQLab) were used at the concentration of 1% in 0.5% TBE (45 mM Tris-borate, 1 mM EDTA). Run was carried out in 0.5% TBE by means of the Chef-Mapper apparatus (BioRad). Lambda Ladder PFG Marker and MidRange I & II PFG Markers (New England Biolab) were used for size estimation. The temperature was maintained at 14°C by means of a cooling pump. The run time and the switch time were automatically calculated by the apparatus according to the input for separation range. Gel was stained with EtBr after run and visualized as described above.

3.2.6. Southern blot

This method was used for transfer of DNA from agarose gel onto nylon membranes for subsequent hybridization with specific probes.

Electrophoresis was executed at low voltage overnight to ensure a better separation of the bands. Gels were usually let run without EtBr and stained after the run. After the staining and photographic documentation, the DNA was nicked in a UV crosslinker at 60 mJ/cm^2 to facilitate the transfer of larger DNA fragments (over 10 kb). DNA was denatured by bathing the gel twice in denaturation solution (0.5 M NaOH, 1.5 M NaCl) for at least 15 min each time.

3.2.6.1. Alkaline capillary blotting

Two large sheets of gel blotting paper (Schleicher and Schuell) were prewetted with denaturation solution and laid on a glass plate; the agarose gel was placed face down on it and covered with prewetted nylon membrane (Hybond N$^+$, Amersham) avoiding formation of air bubbles between the gel and the membrane. Dry gel blotting paper and a stack of paper towels were piled up above and kept pressed down overnight with a weight to allow capillary flow of liquid. For genomic DNA Southern blot, the lower blotting paper was dipped in solution to increase the flow rate and the efficiency of the transfer. After the transfer, the nylon membrane was marked at the position of the wells and

neutralized in 1 M Tris-HCl pH7.5, 1.5 M NaCl. DNA was fixed on the membrane by UV crosslinking at 120 mJ/cm^2.

3.2.6.2. Radioactive and non-radioactive hybridization

Labeling. 25-50 ng of DNA probe were labeled with 50 µCi α[^{32}P]dCTP (Amersham) by means of the Megaprime DNA Labeling Kit (Amersham) as described in enclosed protocol. Labeled probe was purified from free nucleotides through MicroSpin S-300 HR Columns (Pharmacia). Efficiency of labeling was checked by measuring 2 µl of the flow-through in an isotope counter Ersicount 400 (Scottlab).

Pre-hybridization. Membranes were saturated in 20-25 ml pre-hybridization buffer at 65°C for at least 3 hrs shaking in hybridization oven (Hybaid, Shake' N' Stack). Pre-hybridization solutions: standard buffer (5x SSC*, 0.02% SDS, 0.1% N-lauroylsarcosine, 1% Blocking Reagent (Roche)) or Church buffer (1 mM EDTA, 500 mM $NaPO_4$ pH7.5, 7% SDS + 0.1 mg/ml salmon sperm DNA or 1% BSA to add fresh before use).

Hybridization. After the pre-hybridization, solution was replaced with 5-10 ml fresh pre-hybridization buffer containing the probe, previously denatured 5 min at 95°C, and the membrane was incubated overnight at 65°C.

Washing. In order to remove the aspecifically bound probe, the following washing steps were carried out in a shaking water bath at 65°C:

2x 30 min in 2x SSC, 0.1% SDS

2x 30 min in 0.1x SSC, 0.1% SDS

Membranes were rinsed in 2x SSC and tightly sealed in plastic bags or plastic wrap.

Exposition. Biomax MS autoradiographic films (Kodak) were exposed on the membranes inside light-proof autoradiographic cassettes with enhancer screens at –80°C for few hours in case of hybridization on cloned DNA (plasmid, cosmid or BAC DNA) up to overnight in case of genomic DNA. Development of films was accomplished by means of a Curix 60 film developer (Agfa).

Stripping. If the membrane was to be reused for hybridization with a different probe, the old probe was removed by bringing the membrane to boil in 1% SDS and let cool down to room temperature. After rinsing with 2x SSC the membrane was ready to be used.

* 20x SSC: 3 M NaCl, 0.3 M sodium citrate pH7

For low stringency hybridizations (for instance interspecies hybridization with mouse probes on *F. rubripes* genomic clones) the following modifications were applied.

Hybridization standard buffer with 7x SSC instead of 5x SSC

Washing: 2x briefly in 2x SSC, 0.1% SDS at room temperature

2x 30 min in 2x SSC, 0.1% SDS at 65°C

Hybridization experiments for BAC end cloning were carried out with non-radioactive probes. In this case the DIG-High Prime kit (Roche) was used to label the probes with digoxigenin-conjugated dUTP (DIG-11-dUTP), according to enclosed protocol. Hybridization was carried out in standard buffer (see above) with 20-30 ng/ml probe. The hybridization signal was detected with an alkalyne phosphatase (AP)-conjugated anti-digoxigenin antibody (Roche) and subsequent chemiluminescent revelation with CSPD substrate (Roche). Guidelines for hybridization and detection procedures are described in the DIG System User's Guide for Filter Hybridization supplied by Roche.

3.2.7. Colony Hybridization

Hybond N^+ (Amersham) membranes were laid for 2 min onto agar plates with *E. coli* colonies for transfer and in the meanwhile they were marked with reference ink dots. Subsequently the membranes were laid twice face up on 750 µl 0.5 N NaOH for 2 min and similarly twice on 1 M Tris-HCl pH7.5 for 2 min (the excessive solution was each time drained with Whatman paper to avoid washing off of the colonies). Then they were let dry up and used for hybridization.

These modifications were applied to the hybridization protocol described above.

Labeling. Oligonucleotides were used as probes. Oligo probes (100 ng) were end-labeled at the 5' end with 5 units T4 Polynucleotide Kinase (PNK, MBI Fermentas) for 30 min at 37°C using 30 µCi (10 mCi/ml) $\gamma[^{32}P]$ATP (Amersham) and added directly to the hybridization solution without purification.

(Pre)-Hybridization solution: 6x SSC, 1% SDS, 10% dextran sulfate

Pre- and Hybridization temperature: 50-55°C

Washing: 1x 15' RT in 6x SSC, 0.1% SDS

1x 15' 50°C in 6x SSC, 0.1% SDS

1-2x 30' 65°C in 2x SSC, 0.1% SDS (if too much background)

3.2.8. Extraction of DNA fragments from agarose gel

After the gel electrophoresis, the desired DNA bands were cut out with a scalpel under irradiation with low-energetic long wavelength UV light (320 nm) to minimize damaging of DNA itself. The fragments were eluted from the gel pieces by means of the QIAquick Gel Extraction Kit (QIAGEN) according to the provided protocol.

In case of preparation of fragments to inject into fertilized oocytes for generation of transgenic mice, contact with EtBr and UV irradiation was absolutely avoided. Small aliquots of the restricted DNA were loaded on both sides of the sample as markers and run was executed in an EtBr-free electrophoresis chamber. The markers were subsequently separated from the rest of the gel and stained with EtBr to determine the position of the band. The cutting of the band was done at the level of the marked position. Elution was performed with the kit described above. DNA fragments were eluted from the columns with injection buffer (10 mM Tris-HCl pH7.5, 0.1 mM EDTA).

In case of high molecular weight fragments for generation of BAC transgenic mice, the DNA was extracted from the gel pieces through electroelution inside dialysis bags. Electroelution was carried out in an electrophoresis chamber in 0.5% TBE for 3 hrs at 3 volts/cm and then for 40 sec at inverted polarity to detach DNA from the dialysis membrane. Afterwards dialysis was performed in TE pH8 for at least 2 hrs, DNA was recovered from the bag, EtOH precipitated and redissolved in BAC injection buffer: 10 mM Tris-HCl pH7.5, 0.1 mM EDTA pH8, 100 mM NaCl, 1x Polyamines (1000x polyamine stock: 30 mM Spermine, 70 mM Spermidine)

3.2.9. Cloning and Transformation

Cloning of DNA fragments obtained from restriction digestion was accomplished into linearized plasmid vectors with compatible ends. If vector ends were compatible to each other, appropriate 5’-end dephosphorilation was executed by incubation with 1 unit of shrimp alkaline phosphatase (USB) for 1 hr at 37°C and subsequent purification with the QIAquick Nucleotide Removal Kit (QIAGEN).

Blunting of 5'-protruding ends was required for ligation of incompatible ends. This was accomplished with 2 units of Klenow enzyme (Roche) in restriction buffer H (Roche) with 200 µM dNTP mix for 1 hr at RT.

Cloning of short linkers, supplied as single stranded oligonucleotides, required previous annealing carried out by 5 min incubation at 90°C and slow cooling down to room temperature.

Blunt PCR products, obtained from amplification with a proof reading DNA polymerase (see later), were purified with the QUIAquick PCR Purification Kit (QIAGEN) and phosphorilated with 20 units of T4 Polynucleotide Kinase (MBI, Fermentas) in presence of 200 picomoles of ATP at 37°C and subsequently purified with the QIAquick kit. Alternatively, PCR products with A-overhangs, obtained from amplification with Taq DNA polymerase, were cloned into the pCR2.1 or pCRII TOPO vectors (Invitrogen) conforming to the provided protocol. PCR products with blunt ends, obtained from amplification with proof-reading DNA polymerases, could be cloned into the pCR vectors only upon addition of A-overhangs through incubation with Taq DNA pol and dNTPs for 10 min at 72°C.

Ligations were performed in a molecular ratio of insert and vector 4:1 with total 100 ng DNA and 1 unit of T4 DNA ligase (Gibco, BRL) overnight at 14°C.

3.2.9.1. Competent cells preparation and transformation

Chemically competent *E. coli* cells of DH5α and DH10b strains were prepared as follows. One single colony from an LB-Agar plate was inoculated in 2 ml LB and let grow overnight at 37°C. On the next day 500 µl of starter culture were transferred into 100 ml LB and shaken at 37°C until cell density reached an optical density at 600 nm wavelength (OD_{600}) between 0.4 and 0.6. Growth was stopped by placing the culture on ice for 15 min. The cells were subsequently centrifuged at 3000 rpm for 10 min at 4°C and the pellet was resuspended in 15 ml TFB1[1]. The same centrifugation was repeated and the cells were this time resuspended in 4 ml TFB2[2] and left on ice for 15 min. Finally, cells were split in 100 µl aliquots, frozen down in liquid nitrogen and stored at -80°C.

[1] TFB1: 30 mM potassium acetate, 50 mM $MnCl_2$, 100 mM RbCl, 10 mM $CaCl_2$, 15% glycerol → pH5.8 with HCl

[2] TFB2: 10 mM MOPS, 75 mM $CaCl_2$, 10 mM RbCl, 15% glycerol → pH7 with NaOH

For transformation 100 µl cell aliquots were incubated for 30 min on ice for uptake with up to 10 µl of a 20 µl ligation mix or 1 to 10 ng of plasmid DNA for retransformation. After uptake, the cells were heat-shocked at 42°C for 90 sec and subsequently put on ice for 5 min. 200 µl SOC medium (Gibco, BRL) were added and cells were incubated shaking at 37°C for 30 min - 1 hr. 100 µl of cells were plated out on plates of LB supplemented with 15% Agar and the appropriate antibiotic for selection. Plates were then incubated overnight at 37°C, if not differently required. Single colonies were picked and inoculated in LB medium + antibiotic and let grow overnight at 37°C for miniprep analysis. Alternatively, they were directly screened by colony PCR; in this case each single colony was dispersed in 20 µl H_2O and 1 µl of it was used as PCR template. The positive clones were inoculated and expanded.

In case of a first cloning step into pBluescript or a TOPO vector, blue-white selection of the colonies was possible. For this purpose LB-Agar plates were previously added with 20 µl of 100 mM IPTG water solution and 40 µl of 40 mg/ml X-Gal in dimethylformamide solution. White colonies after 37°C overnight incubation were picked for screening.

Bacterial clones were stored as glycerol stocks at –80°C (1 volume bacterial culture + 1 volume 50% glycerol).

3.2.10. Polymerase chain reaction (PCR)

This technique for DNA amplification had several different applications, colony screening, production of DNA fragments to subclone, checking of correctness of clones and constructs, probe synthesis, gene expression analysis, genotyping of mice. According to the specific purpose, PCR amplifications were realized from various types of template, genomic DNA, first strand cDNA and cloned DNA, like plasmids, cosmids and BACs. Moreover, different types of DNA polymerases were chosen. In general, when no subcloning of the PCR product was required, a normal Taq polymerase was employed from MPI-Fermentas or self-produced by Dr. J Adamski, Institute of Experimental Genetics, GSF. In both cases 10x PCR buffer and 25 mM $MgCl_2$ from MBI-Fermentas were used. All the primers were synthesized by U. Linzner, Institute of Pathology, GSF.

In general, if not differently specified, PCR reaction mixes were set up with 1x PCR buffer, 1.5 mM $MgCl_2$, 10 picomoles of each primer, 2.5-5 units of Taq polymerase in 25 or 50 µl volume. The amount of DNA template was specifically determined in each single case.

PCR reactions were performed in a GeneAmp PCR System 9600 (Perkin-Elmer) with the following general program

4 min at 94°C (denaturation)
15 sec at 94°C
30 sec at Ta (annealing)
1 min per kb of expected product at 72°C (extension)
} 25-35 cycles as specified for each reaction
5-10 min at 72°C (elongation)

The annealing temperature (Ta) of each primer was generally calculated 5°C higher than as specified in the synthesis report. Primers were designed so that their Ta ranged between 55°C and 60°C, if not differently required. The lower Ta of each primer pair was chosen for the reaction.

If the PCR product was to be cloned, the proofreading *Pfx* DNA polymerase was used (Gibco, BRL). Before downstream applications, the PCR product was purified by means of the QIAquick PCR Purification Kit (QIAGEN). Sometimes a better performance was obtained by using the Platinum *Taq* DNA Polymerase or Platinum *Pfx* DNA Polymerase (Gibco, BRL), for hardly amplified DNA sequences, with or without addition of provided Enhancer Solution in the PCR mix. Likewise, a High-Fidelity Platinum *Taq* DNA Polymerase (Gibco, BRL) was used when mutation-free difficult PCR amplifications had to be performed. For each of these commercial enzymes, supplied reagents and protocols were used.

A complete list of all the PCR reactions performed in this work is provided in table formats at the end of this section.

3.2.11. DNA sequencing

Sequencing reactions of DNA were accomplished with the ABI PRISM BigDye Primer v3.0 Cycle Sequencing Kit with standard or specific primers. Samples were analyzed with an ABI PRISM 3700 DNA Analyzer.

Alternatively, 500 ng - 1 µg of lyophilized DNA template was sent to MWG-biotech company, together with 10 pmol/µl of primer if sequencing with specific primers was needed.

3.2.11.1. Subcloning approach for sequencing of a Fugu cosmid clone

Genomic clone ICRFc66D2193 from *Fugu* cosmid library no. 66 was treated with various restriction enzymes (HindIII, KpnI, PstI, SalI, XhoI) and the resulting bands were separated on a 1% agarose gel and blotted on nylon membrane. The membrane was hybridized under low stringency conditions with a mouse *Pax9* paired box probe (pb-probe) and the following positive bands were determined, HindIII - 5.5 kb, KpnI >10 kb, PstI 4.5 kb, SalI 6 kb, XhoI >10 kb. The HindIII, PstI and SalI bands were subcloned into pBluescriptKS (FuguH, FuguP and FuguS). The positive clones were identified by restriction analysis and hybridization with the pb-probe. The three clones were subsequently ordered in a short contig by restriction analysis (see Figure 3) and thoroughly sequenced starting with standard primers T3 and T7 from the pBluescript and then walking inside the inserts with specific primers. In order to speed up the sequencing procedure, two SalI fragments from FuguH were separately subcloned (FuguS1 and FuguS2).

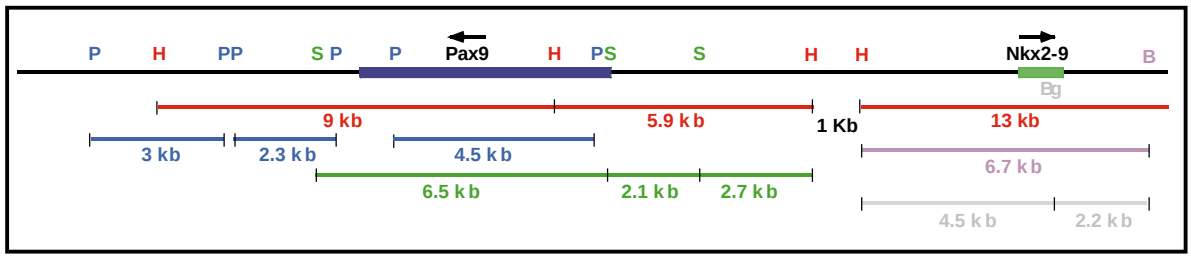


Fig. 3 . Subcloning strategy for sequencing of the *Fugu* cosmid clone.
Pax9 and *Nkx2-9* positions are shown; arrows indicate transcription orientation. Letters indicate restriction sites (B=BamHI, Bg=BglII, H=HindIII, P=PstI, S=SalI). Subcloned fragments of given size are marked with the same color as the corresponding restriction sites.

A similar approach was applied for isolation and sequencing of the portion of the same cosmid clone containing the *Nkx2-9* gene. The same membrane as above, in addition to

one with a BamHI and one with an EcoRI digestions of the cosmid, was hybridized with a mouse *Nkx2-9* homeobox probe (hb-probe) (PCR table 4, no. 5). A 13 kb HindIII positive band was subcloned into pBluescriptKS as above (FuguNK). Cosmid vector sequence was trimmed off the clone by excising a 6.5 kb BamHI fragment and self-ligating the remaining portion (FuguHB). FuguHB was further subdivided after BamHI digestion into two smaller subclones, FuguHBB and FuguBB.

The sequencing gap between FuguH and FuguNK was filled by PCR, using primers lying on the facing ends of the two cosmid fragments and directing externally (PCR table 4, no. 10). The resulting 2 kb band was cloned and sequenced.

A further portion of the cosmid on the 3' side in respect to *Pax9* was subcloned for sequencing as follows. A 3' probe was amplified from the most 3' known sequence of the cosmid (PCR table 4, no. 11) and used to hybridize the membrane with the digests described above. Only two new bands HindIII (9 kb) and PstI (2.3 kb) were identified and cloned in pBS-KS (FuguH3' and FuguP3').

From the new sequence two new primers were designed and a FuguH3' probe was amplified (PCR table 4, no. 12). A new hybridization was performed. This led to the identification of a new PstI band (3 kb), which was cloned in pBS-KS and sequenced.

Here follows the list of primers used for sequencing of the *Fugu* cosmid clone.

Fugu-P1 CACATACGGACATACAAGCAGAG
Fugu-H1 CCTCGCCGAAGGCAGGTTC
Fugu-P2 GATTTCTCGTCCTGGCAACG
Fugu-S1 CATGAGTTAGAGCGCAGGAG
Fugu-P3 TAAGCATGGCTGCACCGAGC
Fugu-H2 TCTAACTCATGCGTAACTGACAC
Fugu-S2 CACATCCATAATTGGCCATTATAGTC
Fugu-S3 GCCGGGTTTGATGGATGACG
Fugu-H3 CAGGTATGACTCAACCGTCCTC
Fugu-H4 GACATGACAGGTCTACTGATCACC
Fugu-P4 GGCTGGTAGATTAGATGCATCAC
Fugu-P5 TTTGCGGTGATGACAAAACG
Fugu-S4 CTTTCATGCAAAGCGGCTTC
Fugu-NK1 GGGAACAGAGGGATCATTGTG
Fugu-P7 AATTCTCCACAAGTCTCGCGAG
Fugu-H6 AATAACGCCGCTGCTCTTTC
Fugu-P6 ATCTTCCTCTTCGCCGTCTTAC
Fugu-H5 TTACGCCCGGCAGAAATGTG
Fugu-NK2 CCAGGATGGACACAAAGACTGTG
Fugu-H7 TTGGTGCGGGAGGAACTGAG
Fugu-PN1 GCCACTTTGCACATGAATGTCG
Fugu-NK3 CTCTCGAATGGAGTGGCCTC
Fugu-NK4 AGGATGGGCAGGCTAATACCAG
Fugu-P7b CTCAGCGTAACAGACTCGGTTG
Fugu S5 ACGCATCACCATCTCAGAGC
Fugu-S6 CATCAAACCCGGCGAACGAG

Fugu3'-P1 GGTCGAGTGCGAGTCCGCAG
Fugu3'-H1 TGGACCAGGACCCCACCTTG
Fugu 3'-P2 CGCTCTACGGGTCAATCTAATC
Fugu 3'-P3 TGTGATCACCTCAGAGCAGCAG
Fugu 3'-H2 GGGAGGAAGTGGGTCGAGC
Fugu 3'-H5 GCTTTCATGCCGGCTTTGC
Fugu 3'-H6 TCGACGCTCACCCCTCCTC

3.2.12. Large genomic sequence comparison

Alignment and search for homologous regions between large genomic sequences of different species was carried out with the algorithm "Percent Identity Plot" (PIP) available on the web page http://bio.cse.psu.edu/pipmaker/ (Penn State – Bioinformatics Group) (Schwartz et al. 2000).

3.2.13. Construction of conventional transgenes with CNSs

A 1 kb fragment containing the CNS-6 was amplified from BAC DNA (PCR table 3, no. 11) and cloned in the SmaI-linearized pASShsp68lacZpA. Its original orientation with respect to the mouse *Nkx2-9* promoter was maintained in the construct. This was checked in the construct by PCR with universal primer T3 and the CNS-6 internal specific primer mCpG-3' GCTGCAGTCCTACCAAGCGTG.

A 2.5 kb fragment containing the CNS+2 was amplified (PCR table 3, no. 12) and cloned as above. Orientation was checked by NcoI digestion. The correct inserts were confirmed by sequencing with primer T3. The transgenes were linearized and excised from the vector sequence by SalI digestion.

3.2.14. RNA isolation

For all of the methods described below absolute RNase-free and sterile conditions were used. The self-made solutions were prepared with DEPC treated H_2O and sterilized by autoclaving or filtering. Special material and instruments for RNA work only were used, when possible, and handled with most care to avoid RNA degradation due to RNase contamination.

Total RNA was extracted from mouse embryos, organs or cultured cells by means of the RNeasy Mini or Midi Kits (QIAGEN), according to the amount of the starting material, as recommended in the provided handbooks.

Disruption of embryonic tissue was done in the supplied lysis buffer by homogenization through a syringe needle, first several times through a 20-G needle and subsequently several times through a 26-G needle. Tissue disruption of organs of adult mice was performed with a rotor-stator homogeniser. Extraction proceeded as described in the kit protocol. The quality of the extracted RNA was checked on a 1% agarose gel.

The mRNA fraction was separated by using the mRNA isolation kit from Roche. The isolation procedure is based on hybridization of mRNA polyA-tails with a biotin-labeled oligo(dT) probe and subsequent capturing on streptavidin magnetic particle with the use of a magnetic particle separator. The extraction was achieved following enclosed kit instructions. The volume of each solution and the required amount of oligo(dT) probe and streptavidin magnetic particles were determined for each specimen, according to the starting RNA material and the provided indications.

3.2.15. RNA formaldehyde agarose gel and Northern blot

Formaldehyde denaturing gels were prepared with a 1-1.5% agarose concentration, as follows. Agarose was dissolved in H_2O and let cool down to about 50°C before adding 10x MOPS buffer to final 1x concentration (10 mM MOPS pH7, 2 mM EDTA, 5 mM sodium acetate) and formaldehyde to a final concentration of 2.2 M. RNA samples were diluted 1:5 in RNA loading buffer (1x MOPS buffer, 2.2 M formaldehyde, 50% formamide, 0.025% bromophenol blue, 0.5 µg/ml EtBr) and heated up to 65°C for 15 min in order to release RNA secondary structures. About 4 µg of poly(A)-RNA were loaded in each well.

After overday run (8-10 hrs) at 60 V with occasional stirring of the buffer to avoid formation of a pH gradient, the gel was photographed, washed once with H_2O for 15 min, equilibrated in 10x SSC for 15 min and blotted (see description for Southern blot) in 10x SSC overnight. RNA was fixed on the membrane by UV-crosslinking (see Southern blot).

Radioactive hybridization mRNA blotted on nylon membranes could be detected with specific radioactive DNA probes. The procedure for probe labeling, hybridization and exposition was carried out as described for Southern blot hybridization with the following modifications.

Hybridization buffer: 5x SSC, 0.2% N-lauroylsarcosine, 0.01% SDS, 2% Blocking Reagent (Roche), 50% Formamide. Hybridization temperature: 42°C

Washing: 3x 30 min in 0.2x SSC, 0.1% SDS at 50°C.

Three different probes were used for detection of *Pax9* mRNA from mouse embryonic RNA extracts, a paired box probe excised from pP9paired, a 3' UTR probe (HindIII probe) excised from pcPax9-k5 and an exon4 probe amplified from pcPax9-k5 (PCR table 2, no. 1).

As a control, a β-actin probe was used (PCR table 4, nos. 1-2)

Mouse Northern RNA blot-12 major Tissues, from "Origene".

Hybridization with exon4 probe was carried out according to enclosed protocol. Briefly, membrane, previously hybridized with an actin probe, was rinsed with 4x SSC 10 min at RT and pre-hybridized for 4 hrs at 42°C in 10 ml hybridization buffer (0.2% SDS, 5% SSPE, 5x Denhardt's solution, 0.1 mg/ml salmon sperm DNA, 50% deionized formamide, 10% dextran sulfate). Hybridization went on in 5 ml of fresh buffer with labeled probe overnight at 42°C.

Washing 3x 5 min 2x SSC, 0.1% SDS at RT 3x 30 min 0.25X SSC, 0.1% SDS 65°C

Exposition overnight

3.2.16. RT-PCR

RT-PCR was performed on total RNA samples in order to check the expression of a gene and/or to subclone the corresponding amplified cDNA sequence.

For the first strand cDNA synthesis, the SuperScript II (Gibco BRL) was employed together with buffer and reagents supplied with the enzyme, as described in the enclosed protocol. The cDNA synthesis reaction was primed with an oligo(dT) primer (Gibco BRL) or with a hexanucleotide mix (Random primer $p(dN)_6$, Roche).

After retrotranscription, the RNA was removed with 2 units *E. coli* RNase H (Gibco BRL) for 20 min at 37°C.

2 µl of the reaction mix were used for PCR amplification in 50 µl volume.

Alternatively, the "SuperScript One-Step RT-PCR with Platinum Taq" from Gibco BRL was used with the reagents and conditions described in the enclosed protocol.

RT-PCR efficiency was determined by PCR amplification of a house-keeping gene (mouse hypoxanthine phosphoribosyltransferase-hprt) (PCR table 4, no. 3)

5' RACE-PCR

Two different kits were used to perform a 5' RACE-PCR for mouse *Pax9*. The first one was the "5' RACE System for Rapid Amplification of cDNA Ends, Version 2.0" (Gibco BRL). The following specific primers were used.

GSP1-pax9 ATGAGTAAATGTGGTTGTAG, reverse primer for first strand cDNA synthesis

GSP2-pax9 GGTGCTGCTTGTAAGAGTCGTAATG, nested reverse primer for the first PCR amplification

GSP2n-pax9 GCACGTTGTACTTGTCGCACA, nested reverse primer for the nested PCR amplification.

Both PCR amplifications were performed with Taq DNA pol (Gibco BRL)

GSP2n5'-pax9 ATTGCTCTGAGCAGTACACCAAC, oligo probe used for colony hybridization in screening for 5' RACE-PCR products.

Further attempts to extend the 5' sequence were made with primer GSP6 for the first strand synthesis and the following PCR primers, GSP7-pax9 GACACACCCCAAAAGAGGTG (first PCR) and GSP7n-pax9 TGACACACCCCAAAAGAGG (nested PCR) or

ext5' TGCTGGAGTCCAGCGAGCGCTTAGC

and ext5'nested GCGGCCTGAAACCCACTTTTCATTCTCC

The second kit, based on a different principle, was the "GeneRacer Kit" Version B (Invitrogen). Primer GSP2-pax9 was used for retrotranscription.

Mouse *Pax9* transcription start point B (TSS B) was detected with primers

GSP7+ GCGACGACGACGCTGTGGACGAAC (first PCR)

and GSP7n+ CGGCTGTTCAGCCTTCCGCCAGATG (nested PCR). TSS A was detected with primers GSP8+pax9 CCCGAAGGCTGGCTCCATTGCTCTG (first PCR)

and GSP8n+pax9 CCCGGCCCCAGTTCCGCACTC (nested PCR).

All of the PCR reactions were executed with High-Fidelity Platinum Taq Polymerase (Gibco BRL) at 68°C annealing and extension temperature.

A Marathon ready-cDNA kit (Clontech), which contains single stranded DNA pool from mouse E11.5 embryos, tagged at the 5' end with an anchor primer, was used with primers GSP7+ and GSP7n+ to confirm 5' RACE-PCR results.

3' RACE-PCR

The "5' RACE System for Rapid Amplification of cDNA Ends, Version 2.0" (Gibco BRL) was adapted for 3' RACE as follows, a poly(dT) primer (Gibco BRL) was used for first strand cDNA synthesis; the two PCR rounds were performed with primers:

3'-probe-5' CCTCTAACAGAAGTCACTAGG

and anchor-poly(dT) GTGTAGTCATGCAGTGATCGTACAG$(T)_{20}$ (first round);

GSP5-pax9 TAGGAACACATCTAATGTGAAATGG

and anchor GTGTAGTCATGCAGTGATCGTACAG (second nested PCR).

3.2.17. Screening of BAC library RPCI – 23 filters for *Pax9*

Probe labeling. A 850 bp HindIII fragment corresponding to the 3' end of the mouse *Pax9* gene was excised from plasmid pcPax9-k5 (see above) and used as probe for this screening. 100 ng of it were labeled with the Megaprime DNA Labeling Kit (Amersham), using 200 µCi of $\alpha[^{32}P]$dCTP. Total activity: $6x10^7$ cpm

Contemporaneously, a control probe, provided with the filters as orientation marker, was labeled. Total activity: $6.6x10^7$ cpm.

Pre-hybridization. The filters were initially rinsed in 6x SSC, 0.1% SDS, twice for 10 min at RT and then pre-hybridized as described above (see radioactive hybridization) in 150 ml solution for 3 hrs at 65°C.

Hybridization. 20 ml fresh solution per filter supplemented with the two probes, 16 hrs at 65°C. Note: this screening was performed together with a mouse *Pax1* probe.

Washing. Filters were rinsed briefly in 2x SSC, 0.1% SDS (150 ml per bottles) and then washed as follow.

2x 30 min at 65°C in 2x SSC, 0.1% SDS

4x 30 min at 65°C in 0.1x SSC, 0.1% SDS in hybridization oven (150 ml/ bottle)

2x 30 min at 65°C in 0.1x SSC, 0.1% SDS 2 l in a box in water bath

again 2x as above in 1 l solution.

Final rinsing in 0.1x SSC

Exposition overnight at –80°C and a shorter exposition 4 hrs at –80°C for membranes with high background.

Filters that still showed high background were further washed 2x 30 min at 65°C in 0.1x SSC, 0.1% SDS and exposed again.

23 specific signals were identified as duplicate spots as described in the supplied instructions and the respective bacterial clones were ordered from Research Genetics.

These clones will be from now on described as BAC 1 to 23.

In order to discriminate the *Pax9* from the *Pax1* clones, a *Pax9* specific PCR was executed on the BAC DNA samples after preparation (PCR table 1, no. 1).

3.2.18. Cloning of BAC ends

BAC clones were digested with six different restriction enzymes (BglII, HindIII, KpnI, NheI, PstI, XbaI), which would as well cut inside the BAC vector pBACe3.6. Reactions were carried out with 5 µl miniprep DNA and 20 units of enzyme in 25 µl volume.

2x 10 µl of each of the six restriction sets were run in duplicate on a 0.8% agarose gel and blotted on nylon membranes. 5 µl were kept for later use.

The membranes were subsequently non-radioactively hybridized with common left and right end probes, which consisted of about 350 bp long fragments amplified and subcloned from the vector sequences on both sides of the BAC insert (PCR table 1, nos. 2 and 3).

Considering the size of the hybridization signal from each restriction digestion of each BAC clone with either end probe and considering the position of each restriction site on the vector, it was possible to calculate the size of the ends generated by the six restriction enzymes in all of the BAC clones. For each BAC clone, the restriction enzymes were chosen that would generate an end fragment between 0.5 and 2.5 kb in size on either side. The selected end fragments are reported in the table below. The remaining 5 µl of each of the listed digestions were purified (QIAQuick) and incubated with 1 unit of T4 DNA ligase (Roche) for 1 hr at 37°C (self ligation). The self-ligated BAC ends were then amplified by inverse PCR using the respective common end primer (left or right) and a specific primer from the vector sequence adjacent to the restriction site as shown in the Figure. Primers are listed in PCR table 1 nos. 4 and 5. The PCR products were subcloned into the pCR2.1TOPO vector and used as probes for BAC end mapping on blotted EcoRI digests of the BAC clones.

BAC clone	Left end	Right end
1	BglII (1 kb)	HindIII (1.5 kb)
2	HindIII (1 kb)	PstI (1.2 kb)
3	PstI (1.2 kb)	PstI (0.7 kb)
4	HindIII (1 kb)	HindIII (1.7 kb)
6	HindIII (0.5 kb)	HindIII (1.5 kb)
8	PstI (1 kb)	PstI (0.7 kb)
9	BglII (1 kb)	HindIII (1.5 kb)
15	XbaI (0.8 kb)	HindIII (2.5 kb)
16	BglII (0.5 kb)	BglII (0.4 kb)
17	HindIII (0.6 kb)	BglII (1.5 kb)
18	as BAC17	as BAC17
21	XbaI (1.8 kb)	HindIII (2 kb)
22	PstI (1.5 kb)	BglII (1.2 kb)

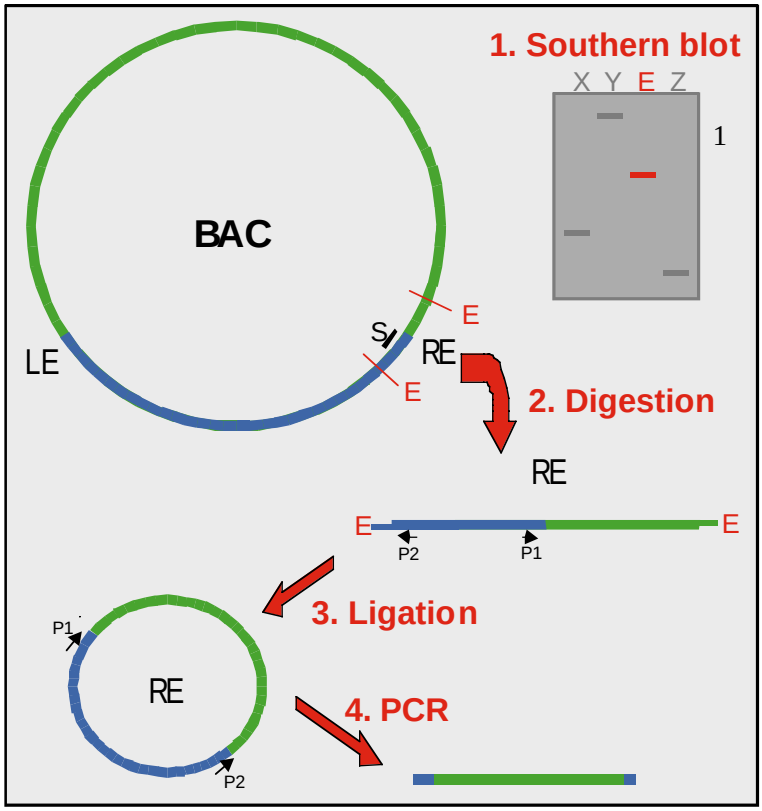


Fig. 4. Cloning of BAC ends
1. Restriction enzyme E is selected for cloning of the right end fragment (RE).
2. RE is excised and
3. self-ligated into a circular form.
4. BAC end fragment is amplified by inverted PCR using common primers (P1 and P2) from the vector sequence.

The amplified BAC end can be subcloned and used for further applications.

3.2.19. BAC modification through homologous recombination in *E. coli*

This method allows modifications of large DNA constructs, like BACs, PACs and cosmids, where the normal cloning techniques based on the use of restriction enzymes cannot be applied. It is based on implemented targeted homologous recombination in *E. coli* in order to introduce insertions, deletions or point mutations.

Since the normally used *E. coli* strains are deficient in the recombination machinery, an exogenous recombination system has to be imported. For this purpose several different approaches have been established (Yang et al. 1997; Jessen et al. 1998; Muyrers et al. 1999; Yu et al. 2000; Lalioti and Heath 2001; Swaminathan et al. 2001). Two of them were employed in this work for the modification of BAC clones.

3.2.19.1. RecA-mediated BAC modification (Yang et al. 1997)

The principle of this method is described in Figure 5.

Briefly, a targeting cassette is constructed, containing recombination arms at both ends (as short as 500 bp), which are homologous to the DNA sequences flanking the target locus. This cassette is cloned in a temperature sensitive RecA expressing plasmid, which confers tetracycline resistance (pSV1-RecA). The RecA-mediated homologous recombination of the plasmid into the target DNA through one of the two recombination arms will generate a circular co-integrate (Fig. 5B). The co-integrates will be selected on tetracycline at non-permissive temperature (43°C), in order to eliminate the free non-integrated plasmid, and the correct recombination event is verified by PCR and Southern blot analysis. Co-integrates are then resolved through a second recombination event, which results in excision of the pSV1-RecA. Treatment with fusaric acid will favor these resolved constructs by selecting against the clones that still have tetracycline resistance. If the second recombination involves the same homologous arm that recombined the first time, the target DNA will result unmodified (Fig. 5 C-D1). If the second homologous arm recombines, the excision of the targeting vector will leave a modified locus as a result (Fig. 5 C-D2). An appropriate screening will lead to the identification of the recombinants.

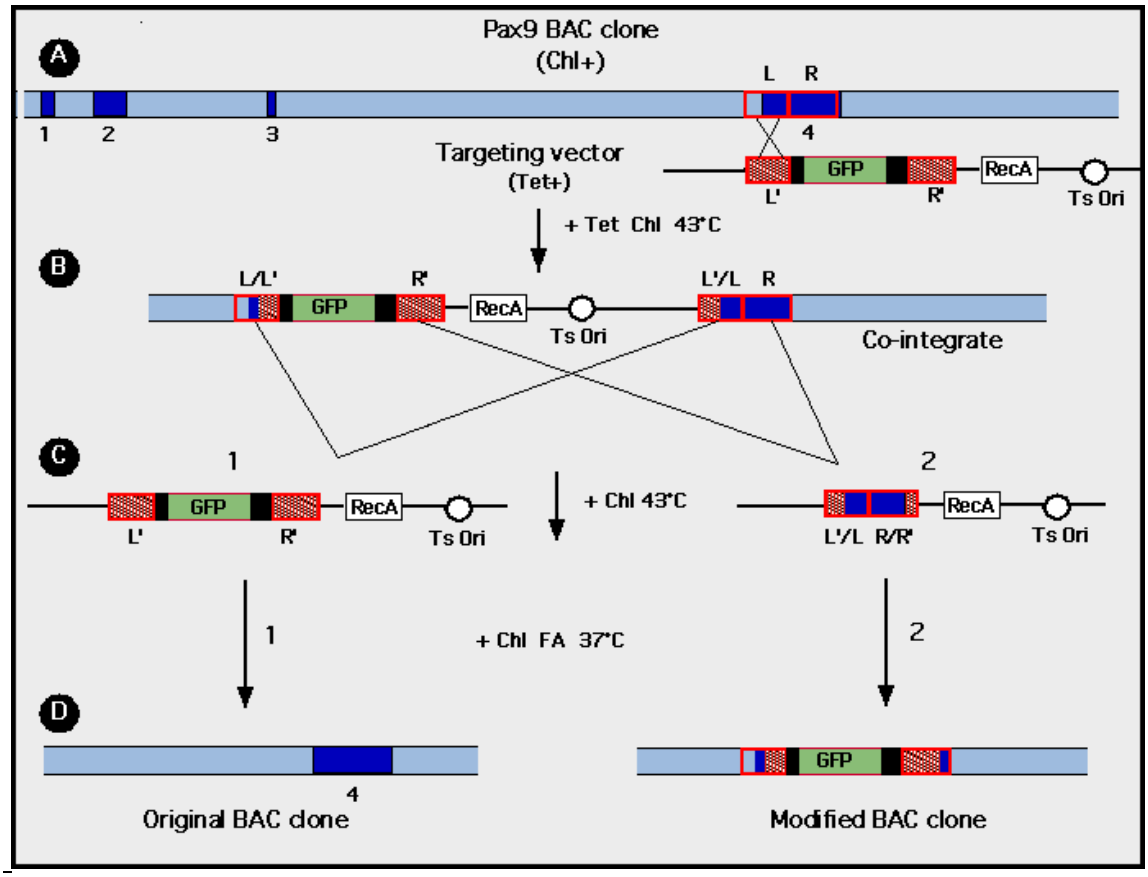


Fig. 5. BAC modification by RecA mediated recombination (see text above).
(A) BAC clone with Pax9 gene (4 exons) and targeting vector. Note: both DNAs are circular. Recombination through left arm (L in target DNA, L' in targeting vector) is represented as an example. Recombination through right arm (R and R') is also possible. Ts Ori - temperature sensitive replication origin, Chl - chloramphenicol, Tet – tetracycline.
(B) Co-integrate and second recombination event. L/L' - L'/L: recombined left arms.
(C-D) Excised fragment (C) and final product (D): 1) wrong resolving eliminates entire targeting vector and leaves BAC unmodified; 2) correct resolving eliminates undesired vector sequence and results in the final modification. FA – fusaric acid.

Insertion of an Ires-GFPneo cassette in the *Pax9* BAC clone 17

1. Targeting vector construction.

The targeting cassette was initially constructed in the pBSKS vector.

Left and right recombination arms (LA and RA), each about 1 kb long, were amplified from the 3'-UTR sequence of the mouse *Pax9* gene (PCR table 1, nos. 6 and 7). LA reverse primer contained an EcoRI site, RA forward primer carried a XbaI site and RA reverse primer a SacI and a more internal SalI sites. LA PCR fragment was cut with SalI (endogenous site) and EcoRI and cloned into the pBSKS corresponding sites (right orientation verified with AvaII digestion). RA fragment ends were in turn cut with SacI

and XbaI and inserted into the same sites of LA-pBSKS (right orientation verified with KpnI digestion).

The Clontech pEGFP-C1 vector was digested with XhoI and SalI and self-ligated, in order to remove the SalI site. The vector was transferred in the DM1 *dam*$^{-}$ *E. coli* strain to activate the XbaI restriction site and linearized with BamHI and XbaI. The Neo^{R} gene was excised from the pGT1.8IresβgeO with the same enzymes and ligated into the prepared vector, producing an in-frame EGFPneo fusion gene. A 1.7 kb EcoRV/BsrGI fragment containing the IRES sequence from the encephalomyocarditis virus and the EGFP coding sequence was excised out of vector pIRES-EGFP and ligated into the NheI/BsrGI linearized pEGFPneo construct, after blunting of the NheI end.

The IRES-GFPneo cassette was cut with EcoRI and XbaI out of the construct, prepared from DM1 *E. coli*, and ligated to the same sites of the LA-RA-pBSKS.

Finally, the whole targeting cassette was moved into the SalI site of the pSV1-RecA plasmid in order to make the targeting vector (pTV-GFP). The right orientation was verified with EcoRI digestion.

2. First recombination and co-integrate formation

BAC17 cells were made chemically competent and transformed with pTV-GFP. Transformants were selected on chloramphenicol (Cm, 12.5 µg/ml) and tetracycline (Tet, 10 µg/ml) at permissive temperature (30°C). Four colonies were picked and dispersed each in 1 ml LB. 100 µl were plated out on Cm + Tet at 43°C to select for co-integrates. Only two of the plates showed normal sized colonies (10 each) on a background of satellites. The colonies were screened by PCR (table 1, nos. 8 and 9) for left or right arm recombination with one internal and one external primer. Six clones turned to be positive (three from either side), but only three (one LA recombinant and two RA recombinants) grew after 2 days liquid culture with Cm + Tet at 43°C. EcoRI digests of these clones were separated on 0.8% agarose gel, blotted and hybridized with exon4 probe (for left arm) and HindIII probe (for right arm).

3. Resolution of co-integrates

The recombinants were streaked out on Cm plates and let grown at 43°C in order to let the second recombination take place. Three colonies from each plate were again streaked

out on TB* plates, containing NaH_2PO_4 (72 mM) and Fusaric acid (12 µg/ml). After three days at 37°C, eight colonies were picked and expanded in liquid culture. Southern blot analysis was carried out as above.

3.2.19.2. BAC modification by ET-cloning

The ET-cloning method was proposed for the first time by Stewart and co-workers (Zhang et al. 2000). In this work, after several unfruitful attempts with the classical approach, the improved version, established by Ioannou and co-workers (Narayanan et al. 1999), was successfully fulfilled (Figure 6).

A targeting cassette is constructed as already described for the RecA-mediated method, although the recombination arms can be considerably shorter, as short as 50 bp. A selectable marker is inserted into the targeting cassette, flanked by loxP or FRT sites, that enable the eventual excision by respectively Cre or Flip recombinase. The bacterial clone containing the target DNA is transformed with a plasmid (pGETrec), which expresses the *E. coli* recombination factors RecE and RecT together with the phage λ Gam factor (an *in vivo* inhibitor of the RecBCD complex) under the control of an arabinose inducible promoter. The linear targeting cassette is introduced by electroporation in the RecET expressing cells and recombinants are selected for the marker. Positive clones are then confirmed for correct recombination by PCR and/or Southern blot analysis. The selectable marker is subsequently removed by transient expression of the appropriate recombinase from a temperature sensitive plasmid (706pMJ-tet).

Insertion of an Ires-βGeo cassette in the *Pax9* BAC clones 17 and 15

1. Targeting cassette construction.

The targeting cassette was built up using the pBSKS as a vector backbone. A 4.5 kb XbaI fragment containing the Ires-βGeo cassette was excised from pGT1.8Iresβgeo and inserted into pBSKS. As a selectable marker an FRT-flanked kanamycin cassette was amplified from pGK-FRT with primers containing respectively SalI and XhoI cloning ends (PCR table 1, no. 10). Since the direct cloning procedure did not succeed, the PCR product was initially cloned into pCR2.1-TOPO vector and the cassette was excised with

* TB: 1% Bacto tryptone, 0.5% Yeast extract, 0.5% Glucose, 0.8% NaCl, 50 nm ZnCl2, 50 µg/ml Chlorotetracycline

SalI and XhoI and cloned into the corresponding sites of the pBSKS/Ires-βGeo, taking care of the correct orientation.

Initially a targeting cassette was constructed, which carried 70 bp long recombination arms, synthesized as linkers from *Pax9* 3'-UTR sequences and inserted in the XbaI/NotI sites (left arm) and in the XhoI site (right arms) of the targeting construct.

Left arm linker: forward oligonucleotide 5'GGCCGCTGTAACTTCCCTTTTCCAGGAAACCT GGCATAACTTTAGGATTTAAAAACAAAAGCAACTCTAAAGGT3'

and reverse oligonucleotide 5'GGCCACCTTTAGAGTTGCTTTTGTTTTTAAATCCTAAAGTTATG CCAGGTTTCCTGGAAAAGGGAAGTTACAGC3'

Right arm linker: forward oligonucleotide 5'TCGATGGAATGAGGCATTTGTGTTGCCCGCACA CTGTTTTAACACAGAGAAGAAACCTATCCCCCTCAAAGGGC3' and reverse oligonucleotide 5'TCGAGCCCTTTGAGGGGGATAGGTTTCTTCTCTGTGTTAAAACAGTGTGCGGGCAACACAA ATGCCTCATTCCAC3'.

The several unsuccessful attempts with this cassette were compensated when the homologous arms were extended to around 250 bp. The new right arm was amplified with SalI and XhoI cloning ends (PCR table 1, no. 12) and it replaced the old short arm in the XhoI site of the construct (XhoI and SalI generate compatible ends). The orientation was checked by XbaI/XhoI digestion. The short left arm could not be exchanged, for XbaI/NotI excision would have disrupted the kanamycin cassette. A 200 bp extension for the left arm was amplified from the directly downstream sequence of the 3'-UTR with restriction sites for EagI and NotI, which as well generate compatible ends (PCR table 1, no. 11). The PCR product was inserted into the NotI site of the construct and its orientation was checked by KpnI/NotI digestion. The final construct was prepared in big amount and 20 µg were used to extract the recombination cassette by XhoI/NotI digestion. The 7.5 kb long fragment was once gel purified, treated again with the same restriction enzymes plus ScaI (which only cuts inside the pBSKS backbone) and gel-purified again. This laborious purification procedure ensures the elimination of any undigested plasmid, which could lead to transformation background. The recombination fragment was finally concentrated in 10 µl H_2O after EtOH precipitation.

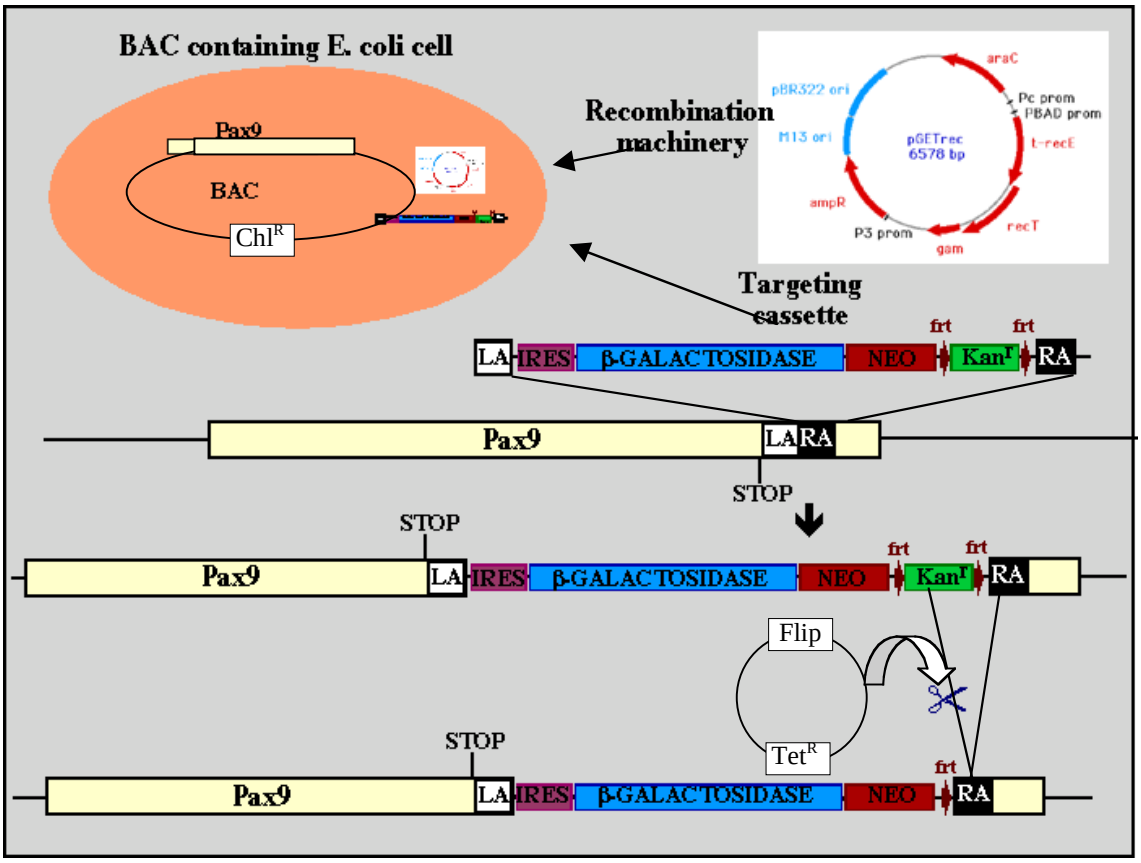


Fig. 6. BAC modification by ET-cloning (see text)

2. Preparation of RecET expressing BAC cells

BAC clones were made chemically competent and transformed with the pGETrec plasmid. pGETrec containing BAC cells were subsequently made electrocompetent as follows.

250 ml LB with Cm and Amp (the latter for pGETrec selection) were inoculated with 5 ml overnight culture and let grown at 37°C to OD_{600} = 0.2. L-arabinose was added to 0.2% final concentration and the RecET system expression was induced at growing conditions for 1 additional hour. The growth was then stopped by placing the culture on ice for 15 minutes. The cells were pelletted at 4000 rpm (Haereus) for 10 min at 4°C. The pellet was thoroughly resuspended in 250 ml ice cold water and centrifuged as above. The same procedure was repeated again twice but with ice cold 10% glycerol instead of water and after the last centrifugation step, the final pellet was resuspended in the

residual supernatant solution in a total volume of 600-700 µl. The cells could be immediately used for electroporation or stored at -80°C for future use in 50 µl aliquots.

3. ET-mediated recombination

2-3 µl of targeting cassette (100-400 ng) were used to transform a 50 µl aliquot of electrocompetent recombination-competent BAC cells. The electroporation was performed with a Bio-Rad Gene Pulser in 0.1 cm gap cuvettes at 2.5 kV, 25 µF with pulse controller set to 100 ohms. Bacteria were resuspended in 1 ml SOC and incubated shaking for 90 min at 37°C. 100% of the cells were plated out on one Cm/Kan plate and placed 24 hrs at 37°C. Colonies were screened by PCR using an internal and an external primer on both sides (PCR table 1, nos. 13 and 14). PCR positive clones were expanded in liquid culture on Cm/Kan selection and further checked by Southern blot analysis with Exon4 and HindIII probes on EcoRI digests.

4. Removal of the kanamycin cassette

Recombinant clones were made chemical competent and transformed with the 706pMJ-tet. Transformants were selected on Cm/Tet plates at the permissive temperature of 30°C for 2 days. Four to six colonies were inoculated in 2 ml LB + Cm and incubated overday at 37°C to allow expression of the Flip-recombinase. Each culture was then streaked out on Cm plates and incubated overnight at 37°C. Single colonies were again picked and streaked individually on Cm plates and in parallel on Kan plates. The clones that only grew on Cm and did not on Kan had lost the Kan cassette. The correct excision was checked on EcoRI digests of these clones with a Kan probe.

Deletions of three large intergenic sequences from modified BAC17

1. Targeting cassette construction

The FRT sites of the kanamycin cassette were replaced with mutated FRT5 sites as follows.

The 200 bp SalI/AgeI fragment of the pCR2.1/FRT-Kan (see above), containing the 5' FRT site, was removed and an FRT5 linker (S-FRT5-A) with compatible ends was inserted at its place. Similarly, on the 3' side a ~ 400 bp KpnI/HindIII fragment was replaced with an FRT5 linker (K-FRT5-H). The correct insertion of the two mutated FRT5 sites was confirmed by restriction analysis and sequencing. FRT5 sequence was taken from Schlake and Bode 1994).

S-FRT5-A linker:

forward oligo Frt5 5'L TCGACGAAGTTCCTATTCTTCAAAAGGTATAGGAACTTCA

reverse oligo Frt5 3'L CCGGTGAAGTTCCTATACCTTTTGAAGAATAGGAACTTCG

K-FRT5-H linker:

forward oligo Frt5 5'R CGAAGTTCCTATTCTTCAAAAGGTATAGGAACTTCA

reverse oligo Frt5 3'R AGCTTGAAGTTCCTATACCTTTTGAAGAATAGGAACTTCGGTAC

The new kanamycin cassette was then transferred into the pBSKS through SalI/HindIII digestion and ligation and was used as a basic construct for the further construction of three targeting cassettes, each of which was created as described below.

The three right recombination arms (RA1, RA2, RA3) were amplified with primers carrying HindIII and XbaI cloning ends (PCR table 1, nos. 16, 19 and 22) and were directly ligated into HindIII/XbaI linearised pBS/FRT5-Kan. The correct cloning of each insert was verified by restriction analysis and sequencing.

The three left arms (LA1, LA2, LA3) were amplified with linked SalI and XhoI cloning ends (PCR table 1, nos. 15, 18 and 21), but direct cloning of the PCR products was not successful. An intermediate cloning step into pCR2.1-TOPO vector was then necessary. After excision with SalI and XhoI, the three inserts were cloned in the respective constructs (e. g. LA1 in pBS/FRT-Kan-RA1 and so on). The resulting constructs bearing the targeting cassettes for the BAC deletions were called pTCΔ1, pTCΔ2, and pTCΔ3.

Sequencing of the these constructs showed that probably due to a sequence error in the cloning ends of the primers, the left arms had been excised from the XhoI site of the vector instead of their own. This resulted in the presence of pCR2.1-TOPO multicloning site sequence from XhoI to EcoRI at the 5’ ends of the left arms.

The targeting cassettes would be excised with XhoI and XbaI. In the case of pTCΔ1 only an XbaI digestion was necessary due to the presence of an internal XbaI site in the RA1. In the case of pTCΔ2 and pTCΔ3 after XhoI and XbaI digestion, part of the additional vector sequence on the left arm side was removed with NotI. Cassette purification was performed as described above.

2. Preparation of RecET expressing BAC cells

This step was realized exactly as described above, with the difference that *lacZ*mBAC17 was used as a target DNA.

3. ET-mediated recombination

See above for the procedure description. Recombinants were checked by PCR using one internal and one external primer on both sides for each deletion. PCR with two external primers from both sides allowed to amplify through the modified locus (PCR table 1. nos. 17, 20, 23)

4. Removal of the kanamycin cassette

Same procedure as before. The molecular evidence of the cassette excision was accomplished by PCR amplification with a pair of external primers. Comparison of the PCR products before and after Flip-recombinase treatment showed the expected size difference.

The reduced size of the three deleted BAC clones was confirmed by PFGE after linearisation with NotI.

3.3. Cell culture

Cells were cultured in DMEM (Dulbecco Modified Eagle Medium, Gibco BRL) supplemented with 2 mM L-Glutamine or DMEM with Glutamax (Gibco BRL) without L-Glutamine addition, 10% Fetal Calf Serum (PAA) and Penicillin-Streptomycin Solution (Sigma). Culturing conditions were 37°C and 5% CO_2. Cells were grown to confluence in 75 cm^2 culturing flasks. Confluent cells were detached from the flask bottom with 1 ml trypsine/EDTA (Gibco BRL), after removing old medium and rinsing with PBS (Gibco BRL), diluted 1:10 - 1:20 and seeded again with 10 ml medium.

3.3.1. Transfection with plasmid DNA and luciferase assay

This method is used to test in a cell line system the promoter or regulatory activity of candidate sequences extrapolated from the genomic regions upstream or around the gene of interest. By cloning the test fragment in an expression vector that carries the coding sequence for the firefly luciferase, it is possible to measure the expression activation or modulation of the luciferase gene as level of luminescence activity. The co-expression of the Renilla luciferase gene as an internal control allows to normalize the data for transfection, cell lysis and assay efficiency. Cells were seeded the day before transfection in 6- (or 12-) multi-well plates at around 10^4 cells/cm^2 density. The transfection was performed with the Lipofectamine-Plus Reagent (Gibco BRL) using 1 µg (or 0.7 µg) construct DNA together with 10 ng of pRL-SV40 per well and transfection procedure was carried out as described in the enclosed protocol.

Cells were harvested 40 hours later and lysed with the lysis buffer supplied by the Dual-Luciferase Reporter Assay System (Promega). The crude protein extracts were used to measure the double luciferase activity. The kit supplied as well the substrates for the reactions, which were performed in polypropylene tubes by means of a luminometer (AutoLumat LB953, HG&G berthold).

The data were processed with the Microsoft Excel application program. Firefly luciferase activity values were normalized to the corresponding Renilla luciferase activity values and multiplied by 10. The relative activity values were represented in a histogram chart. Results from different experiments were compiled together and a mean was calculated taking into account the variable assay efficiency.

3.3.2. Construct preparation

Promoter constructs

About 2 kb long fragments, corresponding to Promoter A and Promoter B, were amplified by PCR (PCR table 3, nos. 1 and 2), using 5'-primers containing an XbaI restriction site (pGL3-promA and pGL3-promB). PCR products were digested with XbaI and ligated into the NheI/SmaI linearized pGL3-basic vector, producing XbaI and NheI compatible ends. Similarly, the Promoter B' construct (pGL3-promB') was generated (PCR table 3, no. 3).

Promoter B deletion constructs (pGL3-promBΔK and pGL3-promBΔS) were obtained by excising the 1 kb KpnI fragment or the 1.5 kb SacI fragment from pGL3-promB and self-ligating the remaining linear DNA.

Enhancer constructs

Test fragments were amplified as shown in PCR table 3 (nos. 3 to 10) with primers carrying SalI cloning ends and cloned into the SalI linearized pGL3-promB. The insert orientation was checked for each single construct by appropriate restriction analysis.

3.4. Methods for experimentation on animals

In this work embryos and organs of the inbred *Mus musculus* strain C57BL/6 and of the outbred strain CD1 were used and analyzed.

Moreover the knockout line *Pax9*^{lacZ}, generated formerly in our laboratory by Dr. Heiko Peters (Peters et al., 1997), was used as a reference for *Pax9* expression during embryonic development and for rescue experiments in crossbreeding with BAC transgenic lines.

The presence of the *Pax9*^{lacZ} allele in the mice produced from the breeding of this line was determined by PCR on crude DNA extract from tail biopsies (PCR table 2, no. 10).

3.4.1. Preparation of mouse embryos

In order to collect mouse embryos from a specific embryonic stage, a daily vaginal inspection of the female in each mating pair was accomplished and the day in which a

vaginal plug was observed was considered as day 0.5 of embryonic development (0.5 day post coitus (dpc) or E0.5).

At the desired day of development (usually between E10.5 and E14.5) the mother was sacrificed and the uterus was extracted. Embryos were carefully pulled out of the uterus in PBS (150 mM NaCl, 15 mM Na-phosphate buffer pH7.3) under a stereomicroscope. Yolk sacs were collected as DNA source if genotyping was required.

3.4.2. X-Gal staining of mouse embryos

Mouse embryos carrying the *lacZ* gene could be stained in order to observe its expression pattern.

Freshly prepared embryos were rinsed in PBS and fixed in solution B (100 mM potassium phosphate buffer pH7.4 (KPP), 5 mM EGTA, 2 mM $MgCl_2$) made 0.2% glutaraldehyde (GA). Fixation was carried out on a shaker at room temperature according to the size of the embryos for a minimum of 15 min (for E10.5 embryos) up to 90 min (for E14.5 embryos). Embryos were washed three times in solution C (solution B + 0.02% Nonidet P-40, 0.01% Na desoxycholate) for minimum 15 min on a shaker. Staining was performed in solution D (solution C + 10 mM $K_3[Fe(CN)_6]$, 10 mM $K_4[Fe(CN)_6]$, 0.5 mg/ml X-Gal) in the dark overnight at 37°C.

After staining samples were washed three times in PBS, post-fixed in 4% paraformaldehyde (PFA)/PBS overnight at 4°C and stored in the same solution.

Clearing of X-Gal stained embryos to observe staining of internal tissues.

After post-fixation embryos were dehydrated through graded steps from 25%, 50%, 75% into 100% methanol (10-30 min per step according to the size of the embryos). Subsequently they were transferred into 1:1 benzoate/benzyl alcohol in a glass dish and cleared in this solution as long as necessary watching every now and then under a stereomicroscope.

Treated embryos were stored in methanol at RT and re-cleared if required for subsequent observations.

3.4.3. Sectioning of stained embryos with vibratome

Post-fixed embryos could be alternatively sectioned to observe more precisely at the tissue level the localization of X-Gal staining.

Whole embryos or parts of them were dipped in gelatin-albumin mix (0.44% gelatin, 27% albumin, 18% sucrose in PBS) for 5-10 min and embedded in the same solution rendered 2.5% GA in a small plastic box. The gelatin-albumin was let solidify and small blocks of embedded material were then dug out and sliced up with a vibratome to 100 µm thickness. Sections were preserved in PBS at 4°C.

Whole embryos and sections were observed on a stereomicroscope (Leica M7 Apo) and photographed with a FUJIX HC-2000 digital camera system. Pictures were imported and edited with Adobe Photoshop 3.5 application.

3.4.4. Whole-mount in situ hybridization

This method is used to detect tissue specific expression of specific genes in a whole-mount embryo by hybridization of a labeled probe on intracellular mRNA. Specific RNA antisense probes are labeled with digoxigenin-UTP and after hybridization are detected with an AP-conjugated anti-digoxigenin antibody. A chromatic reaction for the alkaline phosphatase activity with a specific substrate reveals the localization of the target RNA.

3.4.4.1. Preparation and labelling of RNA probes

The cDNA sequence intended for use as a template for the synthesis of the riboprobe was cloned in a vector containing promoter sequences for initiation of transcription on both sides of the insert (pBSKS or pCRII-TOPO). An antisense probe was generated by 3'-5' oriented transcription of the cDNA and it was used for detection of mRNA by hybridization. A sense probe was transcribed with a 5'-3' orientation from the other side of the insert and used in parallel as a negative control.

Two separate aliquots of 10 µg of plasmid DNA containing a specific cDNA were linearized with two different restriction enzymes that would cut on either end outside the template sequence. The linearized DNAs were purified through a QIAquick spin column, eluted with 50 µl DEPC-H_2O. 10 µl of DNA were transcribed with 2 units T7, Sp6 or T3 RNA polymerase (Roche) (according to the adjacent promoter) in 1x transcription buffer,

40 units RNAse inhibitors (Roche), 1x Dig-RNA Labeling Mix (Roche) in final 20 µl volume, for 2 hrs at 37°C. After the transcription DNA template was removed with 20 units RNAse free- DNAse (Roche) for 30 min at 37°C. The volume was raised to 100 µl and the riboprobe was precipitated by adding 33 µl 7.5 M NH_4Ac, 1 µg tRNA (as carrier) and 400 µl 100% EtOH and incubating 1 hr at -80°C. The pelletted RNA sample was then resuspended in 100 µl DEPC-H_2O plus 40 units RNase-inhibitors. 5 µl of the probe were checked by agarose gel electrophoresis.

Preparation of a mouse *Pax9 in situ* probe

This probe was prepared from plasmid pcPax9-WM as described in Neubüser et al. (Neubüser et al. 1995).

Preparation of a mouse *Pax9* exon 0 *in situ* probe

Probe was amplified by RT-PCR from 11.5 dpc mouse embryo total RNA extract (PCR table 2, no. 9) and cloned in pCRII-TOPO (Invitrogen).

Antisense probe was transcribed with SP6 RNA polymerase on XhoI-linearized template.

Sense probe was transcribed with T7 RNA polymerase on BamHI-linearized template.

Preparation of a mouse *Nkx2-9 in situ* probe

Probe was amplified by RT-PCR from 11.5 dpc mouse embryo total RNA extract (PCR table 4, no. 6) and cloned in pCRII-TOPO (Invitrogen).

Antisense probe was transcribed with SP6 RNA polymerase on XhoI-linearized template.

Sense probe was transcribed with T7 RNA polymerase on BamHI-linearized template.

Preparation of a *lacZ in situ* probe

The probe was transcribed with T7 RNA polymerase from the plasmid pzhsp70-nβgal, linearized with HindIII

Preparation of a zebrafish *Pax9 in situ* probe

A 850 bp EcoRI fragment was excised from pzPax9a and subcloned in pBluescript KS.

Antisense probe was transcribed with T3 RNA polymerase on BamHI-linearized template.

Sense probe was transcribed with T7 RNA polymerase on XhoI-linearized template.

3.4.4.2. Whole-mount in situ hybridization

Mouse embryos were prepared as described above and fixed in 4% PFA/PBS overnight at 4°C. After fixation they were dehydrated through 25%, 50% and 75% methanol (MetOH)

steps in PBS for 10 min each at 4°C and then bleached for 1 hr with MetOH/H_2O_2 (85% MetOH, 15% H_2O_2). H_2O_2 was then extensively removed by washing twice with large volumes of 100% MetOH. Embryos could be stored in MetOH at -20°C.

Before hybridization the embryos were rehydrated in descending MetOH steps (75%, 50%, 25% in PBS) 10 min each at 4°C; then they were washed twice for 10 min with PBT (PBS, 0.1% Tween20) and again 5 min with PBT. A proteinase K treatment followed in proteinase K buffer (20 mM Tris-HCl pH7, 1 mM EDTA) with 20 µg/ml proteinase K for 3 min at 37°C. Embryos were then washed 4x 5 min with PBT.

Hybridizations were performed with an Insitupro robot (Abimed)

All the solutions were prepared with DEPC H_2O and handled in RNAse free conditions. Bottles, tubes and columns were cleaned and sterilized after each hybridization by soaking overnight in 0.1 N NaOH and subsequently rinsing with 100% EtOH and drying up in an oven at 55°C.

Hybridization program was as follows

4x 10 min + 1x 5 min wash in PBT

1x 10 min permeabilization in RIPA (0.05% SDS, 150 mM NaCl, 1% Nonidet P40, 0.5% Deoxycholate, 1 mM EDTA, 50 mM Tris-HCl pH8)

1x 10 min + 1x 5 min PBT

1x 20 min fixation in 4% PFA, 0.1% GA/PBS

1x 5 min + 1x 10 min in PBT

1x 10 min PBT:Hybe buffer (1:1)

1x 10 min in Hybe buffer (50% formamide, 5x SSC, 50 µg/ml heparin, 0.1% Tween20, pH6 with 1 M citric acid) + tRNA (100 µg/ml)

3 hrs pre-hybridization at 65°C in Hybe buffer + tRNA

16 hrs hybridization at 65°C in Hybe buffer + tRNA + DIG-labeled probe (about 0.25 µg/ml previously denatured at 80°C for 3 min)

1x 5 min + 2x 30 min wash at 65°C in Hybe buffer

1x 10 min Hybe buffer:RNAse buffer (1:1)

1x 5 min at 37°C in RNAse buffer (0.5 M NaCl, 10 mM Tris-HCl pH7.5, 0.1% Tween20)

1 hr incubation at 37°C with 100 µg/ml RNAse A in RNAse buffer

1x 15 min in RNAse buffer:SSC/FA/T (1:1)

2x 5 min + 3x 10 min + 8x 30 min at 65°C in SSC/FA/T (2x SSC, 50% formamide, 0.1% Tween20)

1x 10 min at RT in SSC/FA/T:TBST (1:1)

2x 10 min at RT in TBST (0.8% NaCl, 0.02% KCl, 25 mM Tris-HCl pH7.5, 1% Tween20)

2x 10 min at RT in MABT (0.1 M maleic acid, 0.15 M NaCl, pH7.5 with NaOH, 0.1% Tween20)

2 hrs blocking at RT in MABT + 2% blocking reagent (Roche)

12 hrs incubation at RT in MABT + 2% blocking reagent + AP-conjugated anti DIG antibody (1:5000)

3x 5 min + 8x 1hr wash at RT in TBST

Embryos were then taken out of the robot and washed further in TBST overnight at 4°C.

Before staining embryos were equilibrated 2x 5 min at RT in alkaline phosphate buffer (100 mM NaCl, 50 mM $MgCl_2$, 100 mM Tris-HCl pH9.5, 0.1% Tween20, 2 mM Levamisol). Staining was performed in BM purple AP substrate solution (Roche) with 0.1% Tween20 and 2 mM Levamisol at 4°C in the dark. The reaction was carried on for a minimum of one overnight up to a couple of days observing the embryos occasionally on a stereomicroscope until they showed a clear localized blue staining.

Samples were washed 3x 10 min with PBS, post-fixed with 4% PFA in PBS overnight at 4°C and stored in the same solution.

3.4.5. Whole mount ISH on zebrafish embryos

Preparation of embryos

Zebrafish embryos were collected from mating boxes soon after laying and allowed to develop until the desired stage, when they were fixed in 4% PFA for 24-48 hrs at 4°C. The chorion was carefully removed on the microscope.

Embryos were rinsed 2x 5 min in PBT (see above) and then dehydrated in MetOH series (MetOH/PBT) as described for mouse embryos, 2-3 min per step. Embryos up to 24 hours post fertilization (hpf) are transparent and do not need bleaching. For WISH at later

developmental stages (from 48 hpf on), an albino strain was used (no pigments). Embryos could be stored in MetOH at -20°C.

Embryos were processed in a 48-well plate.

Before hybridization, embryos were rehydrated through reverse MetOH/PBT series, 2-3 min each step and equilibrated in PBT. Proteinase K treatment followed: 10 µg/ml in PBT at RT 3 min (20 somite stage), 30 min (48 hrs stage) or 50 min (66-94 hpf); wash 2x 2 min in PBT without shaking.

Hybridization was performed according the following protocol

-Post-fixation in 4% PFA 20 min at RT

-Rinse 4x 5 min in PBT with gentle shaking

-Prehybridization 1 hr at 70°C in of hybridization buffer (65% formamide, 5x SSC, 50 µg/ml heparin, 0.5 mg/ml yeast tRNA, 0.1% Tween 20, 9.2 mM citric acid pH 6).

-Hybridization overnight at 70°C in fresh hyb. buffer + 1.5 µl DIG-probe (about 0.25 µg/ml)

- Rinse in hyb. buffer/2X SSC series (75%/25%, 50%/50%, 25%/75%) 10 min 70°C each step + 10 min 70°C in 2x SSC

- Wash 2x 30 min at 70°C in 0.05x SSC

- 5 min RT in 0.05x SSC/50% PBT (1:1)

- 2x 5 min at RT in PBT

- Block 1 hr at RT in block buffer (PBT, 2% normal goat serum, 2 mg/ml BSA)

- Antibody incubation 2 hrs at RT in block buffer + 1:5000 preadsorbed AP-conjugated anti DIG antibody (preadsorbation in block buffer 1:100 overnight at 4°C)

- Rinse 3x 5 min + 6x 10 min in PBT at RT with shaking and overnight at 4°C

-Equilibration 3x 10 min in NTMT (see above) at RT

- Revelation with NBT/BCIP (225 µg/ml NBT, 175 µg/ml BCIP in NTMT) in the dark.

The reaction was let proceed for several hrs at RT until staining appeared. Stained embryos were washed 4x 5 min in PBT in the dark and stored in 80% glycerol/20% PBT at 4°C.

3.4.6. Generation of transgenic mice

Transgenic mice were generated by pronuclues injection of linearized BAC transgene into fertilized egg cells and subsequent embryo transfer into the oviducts of a pseudopregnant foster mother.

Founder analysis by X-Gal staining was directly performed on transient transgenic embryos at 10, 11 or 12 dpc. Yolk sacs were used for DNA preparation for PCR genotyping (PCR table 3, no. 12).

Generation of BAC transgenic mice was carried out in the laboratory of Prof. Keichi Yamamura (University of Kumamoto, Japan) under the supervision of Dr. Kunjia Abe.

A bacterial clone containing the modified *lacZ*-BAC17 was prepared as agar stab and shipped over to Kumamoto where BAC DNA was prepared. A 190 kb BAC transgene was excised by NotI digestion and purified as described above for BAC DNA gel extraction.

An offspring of 24 animals was obtained from one injection cycle and at around 1 month of age the mice were genotyped for the presence of the transgene as shown in PCR table 1, no. 14 and 24. Founders were delivered from Japan and bred in our mouse facility. The same PCR screening was applied to characterize the succeeding generations.

3.5. PCR Tables

PCR Table 1 (BAC work)

PCR product	primer pair	Polymerase	PCR program
1. mPax9 Intron 1 ~700 bp screening for Pax9 BAC clones	int1 5' CTGATGGGGGACGTTGTCAG int1 3' ACACAGAACGCGCCACAACG	Taq pol (Gibco, BRL)	Ta 62°C extension 1' 72°C 30 cycles
2. Left arm probe for BAC end isolation (~350 bp) Cloning into pCR2.1TOPO	LA- Fw AAACATGAGAATTGGTCGACGG LA-Rev GCGGATCCTCTCCCTATAGTGAG	Taq pol (Gibco, BRL)	Ta 60°C Ext. 1' 72°C 30 cycles
3. Right arm probe for BAC end isolation (~350 bp) Cloning into pCR2.1TOPO	RA-Fw GCGGATCCTTCTATAGTGTCACCTAAA RA-Rev Tttttaatcgttgattgatcgaattga	Taq pol (Gibco, BRL)	Ta 60°C Ext. 1' 72°C 30 cycles
4. BAC end cloning BAC left ends, various size fragments amplified from BAC clones	Common-L CATCGGTCGAGCTTGACATTG PstI/HindIII-L AACCGTAACCGATTTTGCCAG or XbaI/BglII-L TTGGGGTTATCCACTTATCCACG	Pfx DNA pol	Ta 55-58°C Ext. 1-2' 72°C 30 cycles
5. BAC end cloning BAC right ends, various size fragments amplified from BAC clones	Common-R GATCCTCCCGAATTGACTAGTGG PstI-R GCATACAAAGAAACGTACGGCG or HindIII-R TTAACAAAGCGTACTACGGCGG or KpnI/BglII-R AGCAAGTGGGCTTTATTGCATAAG	Pfx DNA pol	Ta 58-60°C Ext. 1-2' 68°C 30 cycles
6. BAC modification (RecA) Left arm for targeting vector Rev primer EcoRI site	BACmod-L5' AAAAGGCAAAAGTTAGCAAGTG BACmod-L3' GGCAGGCATTCAGAATTCCAGA CCTACAATGTTCCATAAGC	Pfx DNA pol	Ta 55°C Ext. 1' 68°C 35 cycles
7. BAC modification (RecA) Right arm for targeting vector 5' primer XhoI site 3' primer SalI site	BACmod-R5' GGCAGGCATTCATCTAGAGTCT TTGGTAAAACCACACCTG BACmod-R3' GGCAGGCATTCAGAGCTCGTC GACTAAATAGCATTTGTATTTCTGATGC	Pfx DNA pol	Ta 55°C Ext. 1' 68°C 35 cycles
8. BAC modification (RecA) Left side screening	BACmod-L5' AAAAGGCAAAAGTTAGCAAGTG Irbgeo-rev GCTTCGGCCAGTAACGTTAG	Taq pol	Ta 55°C Ext. 1' 72°C 25 cycles
9. BAC modification (RecA) Right side screening	Irbgeo-for TCTCATGCTGGAGTTCTTCG 3'RA-screen AGCAGCAGAAGGAATGCAG	Taq pol	Ta 55°C Ext. 1' 72°C 25 cycles
10. BAC modification (ET)-Iresβgeo 1900 bp Kan-FRT cassette (XhoI-SalI)	frt/kan2-5' CACGGAATCTGGGTCGACTCTGC AAACCCTATGCTACTCCGTCG frt/kan2-3' CACGGAATCTGGCTCGAGTCCC GGCGGATTTGTCCTACTCAGGAGAGCG	Pfx DNA pol	Ta 58°C Ext. 2' 68°C 30 cycles
11. BAC modification (ET) - Iresβgeo Left arm for targeting cassette (XbaI-NotI)	LAl-5' GGCAGGCATTCATCGGCCGAAAA AAAAAAAAAATATTTAAGTGTCATC LAl-3' GGCAGGCATTCATGCGGCCGCAAGG CAGCCATTCTGTGACC	Pfx DNA pol	Ta 58°C Ext. 1' 68°C 30 cycles
12. BAC modification (ET) - Iresβgeo Right arm for targeting cassette (XhoI-SalI)	RAl-5' GGCAGGCATTCATCTCGAGGTCGAC TGGAATGAGGCATTTGTGTTGC RAl-3'b GGCAGGCATTCATCTCGAGCTCGAG ACCTTATTTGATTAGAGCATACCAC	Pfx DNA pol	Ta 58°C Ext. 1' 68°C 30 cycles
13. BAC modification (ET) - Iresβgeo Left side screening	exon4 5' GCAGTTTCGTCTCAGCATC Irbgeo-rev GCTTCGGCCAGTAACGTTAG	Taq pol	Ta 57°C Ext. 1' 72°C 30 cycles
14. BAC modification (ET) - Iresβgeo Right side screening and genotyping of mBAC17 transgenic mice	Irbgeo-for TCTCATGCTGGAGTTCTTCG GSP5r-pax9 AGCATTTGTATTTCTGATGCCAAC	Taq pol	Ta 57°C Ext. 1' 72°C 30 cycles
15. BAC modification (ET) - BAC17Δ1 Left arm for targeting cassette (XhoI-SalI)	17del1la 5' TCGCACACATTCCTCGAGGCAT GCTGAAAACCAACCAC 17del1la 3' TCGCACACATTCGTCGACTGGG GCTACCACAGTCTGTC	Pfx DNA pol	Ta 58°C Ext. 1' 68°C 30 cycles

16. BAC modification (ET) - BAC17Δ1 Right arm for targeting cassette	17del1ra 5' TCGCACACATTCAAGCTTGCAG CTGCAATCAGCCCTG 17del1ra 3' TCGCACACATTCTCTAGACCAA ATGGAAGGCAACTCCC	Pfx DNA pol	Ta 58°C Ext. 1' 68°C 30 cycles
17. BAC modification (ET) - BAC17Δ1 Screening with external primers	Del1 ext-5' AAGGAAGATGGACTCCAGACC Del1 ext-3' CCTCGTGAAGTGCTTTACAGC	Taq pol	Ta 55°C Ext. 1' 72°C 30 cycles
18. BAC modification (ET) - BAC17Δ2 Left arm for targeting cassette	17del2la 5' TCGCACACATTCCTCGAGCACC TTCACTCCCCGTAGAAC 17del2la 3' TCGCACACATTCGTCGACATCA GCACGAGTTGAGGGAG	Pfx DNA pol	Ta 58°C Ext. 1' 68°C 30 cycles
19. BAC modification (ET) - BAC17Δ2 Right arm for targeting cassette	17del2ra 5' TCGCACACATTCAAGCTTGACA CCGGCTGGACCTAGG 17del2ra 3' TCGCACACATTCTCTAGACACA TGGTCGGTGGCTCAC	Pfx DNA pol	Ta 58°C Ext. 1' 68°C 30 cycles
20. BAC modification (ET) - BAC17Δ2 Screening with external primers	Del2 ext-5' CCAGCTACTGGGTCAACCTAAC Del2 ext-3' GCTGGCTGCTCTTCCTAGAGG	Taq pol	Ta 55°C Ext. 1' 72°C 30 cycles
21. BAC modification (ET) - BAC17Δ3 Left arm for targeting cassette	17del3la 5' TCGCACACATTCCTCGAGTTCCAG GACAGAAAACAGCC 17del3la 3'TCGCACACATTCGTCGACGCTAGAGG CAATCTCACCACC	Pfx DNA pol	Ta 58°C Ext. 1' 68°C 30 cycles
22. BAC modification (ET) - BAC17Δ3 Right arm for targeting cassette	17del3ra 5' TCGCACACATTCAAGCTTCTCAAAA GTCATGCTTGTGTTTAGC 17del3ra 3' TCGCACACATTCTCTAGACACTTTAGTC TTTCACTTTCCTGGC	Pfx DNA pol	Ta 58°C Ext. 1' 68°C 30 cycles
23. BAC modification (ET) - BAC17Δ3 Screening with external primers	Del3 ext-5' GCCTTTAGCTCTCTTGCAGAGG Common-L CATCGGTCGAGCTTGACATTG	Taq pol	Ta 55°C Ext. 1' 72°C 30 cycles
24. Genotyping of mBAC17 transgenic mice	BAC-end V AGGTGACACTATAGAAGGATCCG BAC-end I GGTGCTAGCTCAACTGGTGG	Taq pol	Ta 58°C Ext. 1' 72°C 30 cycles

PCR Table 2 (Pax9 gene)

PCR product	primer pair	Polymerase	PCR program
1. ~600 bp from mPax9 cDNA; exon4 probe for Northern and Southern blot hybridization	exon4 5' GCAGTTTCGTCTCAGCATC exon4 3' CTCAACAATTGCACGTTTCG	Taq pol	Ta 57°C Ext. 1' 72°C 25 cycles
2. mPax9 Intron2 ~2.5 kb	int2-5' GGGCATCCGCTCCATCACC int2-3' GTAGGGGGAGCTGTCGCTC	Pfx DNA pol	Ta 60°C ext. 3' 68°C 25 cycles
3. mPax9 Intron3 ~9.5 kb	int3-5'b GATTGGAGAAGGGAGCCTTG int3-3'b GGATGCTGAGACGAAACTGC	Elongase (Gibco,BRL)	Ta 58°C ext. 10' 68°C 35 cycles
4. 300 bp from mPax9 cDNA colony PCR of 5' RACE-PCR clones for screening	GSP2n-pax9 GCACGTTGTACTTGTCGCACA GSP3-pax9 GAGGAGTGTTCGTGAACGGAAG	Taq pol	Ta 60°C Ext. 1 72°C 25 cycles
5. RT-PCR (700 bp) to confirm new mPax9 exon0 after 5' RACE.	GSP2-pax9 GGTGCTGCTTGTAAGAGTCGTAATG 5'RACE CTCCAGCACTGGCAATCTCG	Taq pol	Ta 55°C Ext. 1' 72°C 35 cycles
6. RT-PCR (600 bp) Pax9 cDNA fragment exon1-exon2	9581 TTC AGC CGG GCA CAG ACT TCC GSP2-pax9 GGTGCTGCTTGTAAGAGTCGTAATG	Taq pol	Ta 55°C Ext. 1' 72°C 35 cycles

7. RT-PCR (800 bp) Pax9 cDNA fragment exon2-exon4	int2-5' GGGCATCCGCTCCATCACC exon 4-3' CTCAACAATTGCACGTTTCG	Taq pol	Ta 55°C Ext. 1' 72°C 35 cycles
8. RT-PCR (850 bp) Pax9 cDNA fragment exon0-exon2	ext5'nested GCGGCCTGAAACCCACTTTTCATTCTCC GSP2n-pax9 GCACGTTGTACTTGTCGCACA	Taq pol	Ta 59°C Ext. 1' 72°C 35 cycles
9. RT-PCR (350 bp) Pax9 in situ probe exon0-exon1	ex0 probe 5' CTCCCCTGGCCTGGGAAG ex0 probe 3' GTGGGCTGGGCTGAGCAG	Taq pol	Ta 57°C Ext. 1' 72°C 35 cycles
10. Genotyping of Pax9^{lacZ} mice	9580 CGA GTG GCA ACA TGG AAA TCG C 9581 TTC AGC CGG GCA CAG ACT TCC 9582 GCT GGT TCA CCT CCC CGA AGG	Taq pol	Ta 60°C Ext. 1' 72°C 52 cycles

PCR Table 3 (promoter analysis)

PCR product	primer pair	Polymerase	PCR program
1. Promoter A 2 kb fragment including Pax9 TSS-A	prom1-5' GGCAGGCATTCATTCTAGAGTGAAAAGATCGGTGCTTGG prom1-3' GCGCAGCCAGAAACTTCAG	Pfx DNA pol	Ta 58°C Ext. 2' 68°C 35 cycles
2. Promoter B 2 kb fragment including Pax9 TSS-B	prom2-5' GGCAGGCATTCATTCTAGAAAGGGAGGTGTGCGACAGC prom2b-3' TCTCACTGAGCCGGCCTG	Pfx DNA pol	Ta 58°C Ext. 2' 68°C 35 cycles
3. Promoter B' Promoter B with a 3' end 150 bp deletion	prom2-5' GGCAGGCATTCATTCTAGAAAGGGAGGTGTGCGACAGC prom2-3' GTGGGCTGGGCTGAACAG	Pfx DNA pol	Ta 58°C Ext. 2' 68°C 35 cycles
4. Enhancer construct CNS-5 (~1100 bp)	CNS-5 5'S AAATTCTCTGGTCGACAATGCTTATGGGAGTGTGTGC CNS-5 3'S AAATTCTCTGGTCGACGAATGGAAGTCCCCACACAG	Pfx DNA pol	Ta 58°C Ext. 2' 68°C 35 cycles
5. Enhancer construct CNS-4 (~2300 bp)	CNS-4 5'S AAATTCTCTGGTCGACCCAGTTACATCCGTGCCCTG CNS-4 3'S AAATTCTCTGGTCGACGTTTCCCACCAGTGTGCTGC	Pfx DNA pol	Ta 58°C Ext. 2' 68°C 35 cycles
6. Enhancer construct CNS-3 (~3000 bp)	CNS-3 5'S AAATTCTCTGGTCGACCTGGCTGGCTGTAACACTCC CNS-3 3'S AAATTCTCTGGTCGACCAGCTCCTGTCCACATTTGG	Pfx DNA pol	Ta 58°C Ext. 2' 68°C 35 cycles
7. Enhancer construct CNS-2.1 (~1800 bp)	CNS-2.1-5'S AAATTCTCTGGTCGACCAGGTTCAGAGCAATCGTGC CNS-2.1-3'S AAATTCTCTGGTCGACCAGCACCATGTGAACCACAC	Pfx DNA pol	Ta 58°C Ext. 2' 68°C 35 cycles
8. Enhancer construct CNS-1 (~2200 bp)	CNS-1 5'S AAATTCTCTGGTCGACCCCTGCTAGGAGCATACTGG CNS-1 3'S AAATTCTCTGGTCGACGAGGGAACGTGCAATGATTTAC	Pfx DNA pol	Ta 58°C Ext. 2' 68°C 35 cycles
9. Enhancer construct CNS+1 (~1200 bp)	CNS+1-5'S AAATTCTCTGGTCGACCACAAATGTGCTTCCAAATGC CNS+1-3'S AAATTCTCTGGTCGACTGAAGGGCTGGTTGGAACTC	Pfx DNA pol	Ta 58°C Ext. 2' 68°C 35 cycles
10. Enhancer construct CNS+3 (~2500 bp)	CNS+3 5'S AAATTCTCTGGTCGACGGACCAGGCCTTTGTATAAGGC CNS+3 3'S AAATTCTCTGGTCGACTGATTGTGACCCCTGGTTTAGC	Pfx DNA pol	Ta 58°C Ext. 2' 68°C 35 cycles
11. Transgene construct CNS-6 (1000 bp)	mCpG-5'b CATTTTGCCAGAGGCAGAGG mCpG-3'b AAGGGACAGTGAGCGGTCTG	Pfx DNA Pol	Ta 61°C Ext. 1' 68°C 30 cycles
12. Transgene construct CNS+2 (2500 bp)	CNS+2 5' GGACCAGGCCTTTGTATAAGGC CNS+2 3' TGATTGTGACCCCTGGTTTAGC	Pfx DNA Pol	Ta 58°C Ext. 1' 68°C 30 cycles
13. Genotyping for transient transgenesis	Lac2 CAAGGCGATTAAGTTGGGTAACG Hsp68 ACCTCGAAGCGGCCGCTTC	Taq pol	Ta 58°C Ext. 1' 72°C 30 cycles

PCR Table 4 (other PCRs)

PCR product	primer pair	Polymerase	PCR program
1. RT-PCR first round amplification for mouse actin probe	act5' ATGGATGACGATATCGCTGC act3' GTGGTGGTGAAGCTGTAGC	Taq pol	Ta 56°C Ext. 1' 72°C 35 cycles
2. RT-PCR nested amplification for mouse actin probe (600 bp)	act5'-nest GACAACGGCTCCGGCATG act3'-nest TCCCGGCCAGCCAGGTC	Taq pol	Ta 56°C Ext. 1' 72°C 30 cycles
3. 657 bp of the mouse hypoxanthine phosphoribosyltransferase (hprt) cDNA	Hprt-F: ATGCCGACCCGCAGTCCCAGCGT Hprt-R: TTAGGCTTTGTATTTGGCTTTTCC	Taq pol	Ta 65°C Ext. 1' 72°C 35 cycles
4. 600 bp probe from Kanamycin/Neomycin resistance gene	Kana5' CCGGTTCTTTTTGTCAAGACC Kana3' ATATTCGGCAAGCAGGCATC	Taq pol	Ta 56°C Ext. 1' 72°C 30 cycles
5. Mouse Nkx2-9 Southern blot probe	Nkx2-9 for AACCCTGGACACTCCCGACTGCGG Nkx2-9 rev GCCTCACCAGTTCCAGGAGACCAG	Taq pol	Ta 56°C Ext. 1' 72°C 30 cycles
6. RT-PCR for a mouse Nkx2-9 in-situ probe	Nkx2-9pb5' TCCTCGGACGAGAGCGGCCTGG Nkx2-9pb3' GCCTCACCAGTTCCAGGAGACCAG	Taq pol	Ta 56°C Ext. 1' 72°C 35 cycles
7. Inter-exon PCR for mouse Nkx2-9	Nkx-int5' AACCCTGGACACTCCCGACTGCGG Nkx-int3' GCCTCACCAGTTCCAGGAGACCAG	Pfx pol	Ta 58°C Ext. 1' 68°C 30 cycles
8. mouse Nkx2.1. Synthesis of a genomic probe (350bp)	Nkx2.1-5'b ATGTCGATGAGTCCAAAGCAC Nkx2.1-3'b CGCATGGTGTCCTGGTAAG	Taq pol	Ta 56°C Ext. 1' 72°C 30 cycles
9. Amplification of a Fugu paired box probe with degenerated primers	Fugu-deg 5' GAGCANAC(AG)TN(CT)GGGGA(AG)GTGAACCA Fugu-deg 3' CGGATCTCCCA(AG)GC(AG)AAGATGC	Taq pol	Ta 50°C Ext. 1' 72°C 35 cycles
10. 2000 bp gap between Pax9 positive fragment and Nkx2-9 positive fragment	Pax/Nkx 5' TTCAAAACATTCCGGTCATGAG Pax/Nkx 3' ACAGCCACAATGATCCCTCTG	Pfx pol	Ta 57°C Ext. 2' 68°C 30 cycles
11. Fugu3' probe	FuguS5 ACGCATCACCATCTCAGAGC FuguS6 CATCAAACCCGGCGAACGAG	Taq pol	Ta 56°C Ext. 1' 72°C 30 cycles
12. FuguH3' probe	Fugu3'-H5 GCTTTCATGCCGGCTTTGC Fugu3'-H6 TCGACGCTCACCCCTCCTC	Taq pol	Ta 56°C Ext. 1' 72°C 30 cycles

4. RESULTS

A previous approach for the identification of *Pax9* regulatory elements had been attempted before the beginning of this work by conventional transient transgenesis. Around 15 kb of genomic region directly upstream of the published 5' end of mouse *Pax9* cDNA were attached in front of a *lacZ* reporter gene. The construct was used to generate transgenic mice, which were analyzed at a transient state at the embryonic stage. None of six different transgenic embryos showed a *lacZ* expression pattern upon X-Gal staining which resembled the endogenous *Pax9* expression. The *lacZ* positive domains were mostly ectopic and irreproducible (H. Peters, personal communication). Together with this observation, it has to be mentioned that the attempt to rescue *Pax1* knockout phenotype by transgenesis with two different BAC clones failed, suggesting that the *Pax1* regulatory elements reside scattered in a very large genomic region (B. Wilm, unpublished results). Since the high homology between *Pax1* and *Pax9* both at the level of the coding sequences and with respect to their expression patterns (Neubüser et al. 1995), it was inferred that, if the homology corresponds also to a similarity in gene structure, *Pax9* regulatory elements might as well be located very far apart from the transcribed region. This made necessary to extend the analysis to much wider genomic sequences.

4.1. Determination of *Pax9* gene structure

The first step towards the identification of the promoter region and *cis*-regulatory elements of the mouse *Pax9* gene was the characterization of the gene structure. At the beginning of this work only limited information was available about *Pax9* structure in the mouse. That was the sequence of two overlapping cDNA clones that came up to a total of 2.5 kb transcribed region (Neubüser et al. 1995) and the position of the first intron roughly described within the strategy for the construction of the knockout targeting vector (Peters et al. 1998b).

4.1.1. Establishment of a BAC contig encompassing mouse Pax9

In order to have material for a complete analysis of the gene structure and subsequently for a search for promoter and distal *cis*-regulatory elements, a mouse BAC library was screened for clones containing the *Pax9* gene. This mouse BAC library (RPCI–23 C57BL/6J) consisted of partial EcoRI digests of approx. 200 kb each that covered the mouse genome 11.2 folds. A specific probe corresponding to the last 800 bp of the 3'-UTR of the gene was used (HindIII probe). The library was screened together with a Pax1 probe, so that the total 23 positive clones had to be further distinguished as *Pax1* or *Pax9* clones. This was accomplished by PCR analysis using specific primers for the amplification of the *Pax9* first intron. Twelve of these clones turned out to be *Pax9* positive by this first PCR analysis (Fig. 7a). The length of the insert of each of these clones was roughly estimated by PFGE analysis after NotI digestion (Fig. 7b).

A different PCR amplification with primers lying on the 3'-UTR of the *Pax9* gene allowed to identify an additional *Pax9* BAC clone (clone n°21), which upon the first PCR analysis resulted negative both for *Pax1* and for *Pax9*. The BAC clone 21 contains only the 3' end of the *Pax9* gene and its most 5' end maps probably inside the third intron.

a

Clone n°	*RPCI-23 BAC*	**Size (kb)**
1	**29 B 9**	**190**
2	**36 F 9**	**190**
3	**78 N 16**	**120+95**
4	**76 K 18**	**<220**
6	**136 M 3**	**<220**
8	**181 N 4**	**120+60**
9	**178 A 19**	**190**
15	**327 I 21**	**160+60**
16	**306 P 6**	**220**
17§	**384 O 22**	**190+5**
18§	**384 O 21**	**190+5**
21*	**431 H 15**	**200**
22	**419 A 21**	**150+30**

b

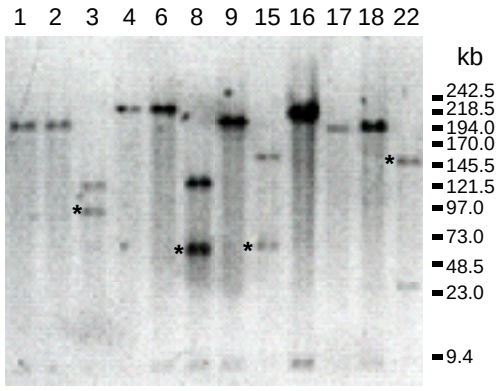


Fig. 7. Pax9 BAC clones. a) List of *Pax9* BAC clones and estimated molecular weights of respective NotI fragments b) PFGE of the BAC clones upon NotI digestion. In case of two bands, a star indicates the *Pax9* positive bands after hybridization with the HindIII probe.

§ the 5 kb band of clones 17 and 18 was observed on a normal agarose gel

* clone 21 is not included on this gel

The second step in the analysis of these BAC clones was the arrangement of a contig around the *Pax9* locus. In order to do that and to determine the relative position of the clones with respect to each other, a BAC end mapping approach was carried out. BAC end fragments were isolated as described in "Materials and Methods" and used as probes on EcoRI digests of the BAC clones themselves. This series of hybridization experiments produced the results showed in table 2 and, together with the presence of an internal NotI site in a subset of the clones (Fig. 7b), it helped establish the contig shown in Figure 8. The position of the *Pax9* gene inside this BAC contig was determined by hybridization with the HindIII probe on NotI digests of the clones (not shown).

Additional information was achieved by comparing the complete restriction pattern of the BAC clones after EcoRI digestion and the size of the BAC ends produced by cutting with different restriction enzymes (data not shown). This comparison allowed to identify clones which shared one common end on either side or which turned out to correspond exactly to the same genomic fragment, like clones 1 and 9 and clones 17 and 18.

	1		***3***		***4***		***6***		***17***		***21***	
	L	***R***	***L***	***R***	***L***	***R***	***L***	***R***	***L***	***R***	***L***	***R***
1	+	+	+	-	+	-	-	+	+	-	+	-
2	+	-	+	-	+	-	-	+	+	-	+	-
3	-	+	+	+	+	-	-	+	-	+	+	-
4	+	-	+	-	+	+	+	+	+	-	+	-
6	+	+	+	-	+	-	+	+	+	-	+	-
8	-	+	-	+	+	-	-	+	-	+	+	-
9	+	+	+	-	+	-	-	+	+	-	+	-
15	-	+	-	+	+	-	-	+	-	+	+	+
16	-	-	+	-	+	-	-	+	+	-	+	-
17	-	+	+	-	+	-	-	+	+	+	+	-
18	-	+	+	-	+	-	-	+	+	+	+	-
21	-	+	-	+	+	-	-	+	-	+	+	+
22	-	+	+	-	+	-	-	+	-	+	+	-

Table 2. BAC end mapping.
Results of hybridisations with left (L) and right (R) end probes of six BAC clones (1, 3, 4, 6, 17, 21) on the complete BAC clone set. Positive matches (+) or no matches (-) are indicated.

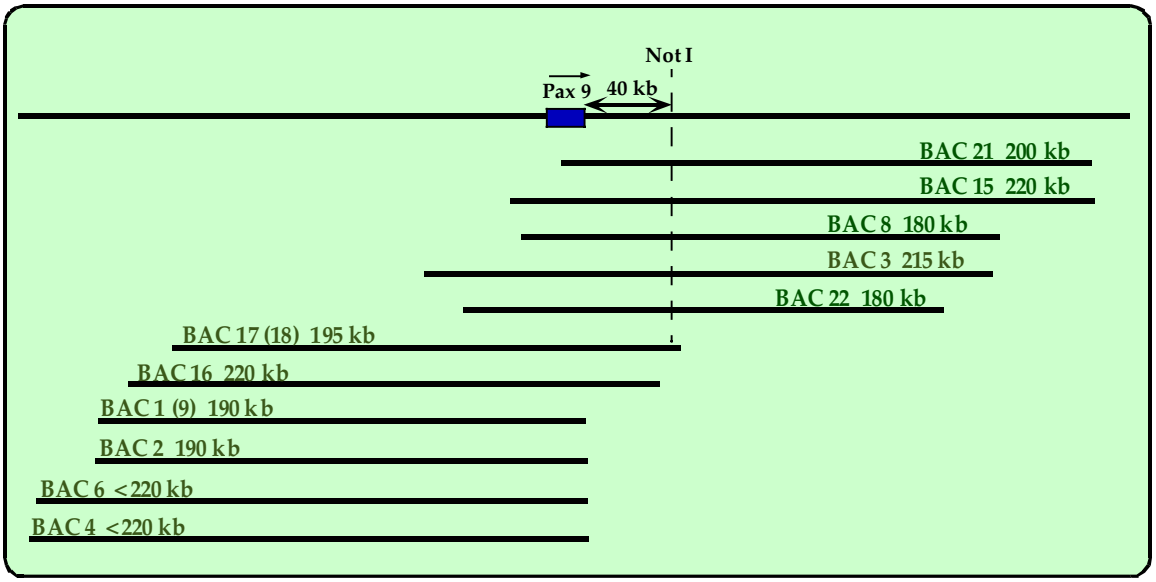


Fig. 8. BAC contig around the *Pax9* locus
The position of the NotI site is mapped around 40 kb from the *Pax9* 3' end
BAC17 and BAC18 coincide, so do BAC1 and BAC9 (see text)

4.1.2. Analysis of Pax9 genomic region

The sequence information of a BAC clone encompassing the human *PAX9* gene was available in the data bank (accession number AL079303). By alignment of this genomic sequence with the published human *PAX9* cDNA sequence (Peters et al. 1997) (Genbank: NM_006194), it was possible to determine the exon-intron boundaries of the gene. From this analysis, it emerged that *PAX9* consists of four exons spanning over about 16 kb sequence (Fig. 9a), in accordance to previous partial observations (Stockton et al. 2000).

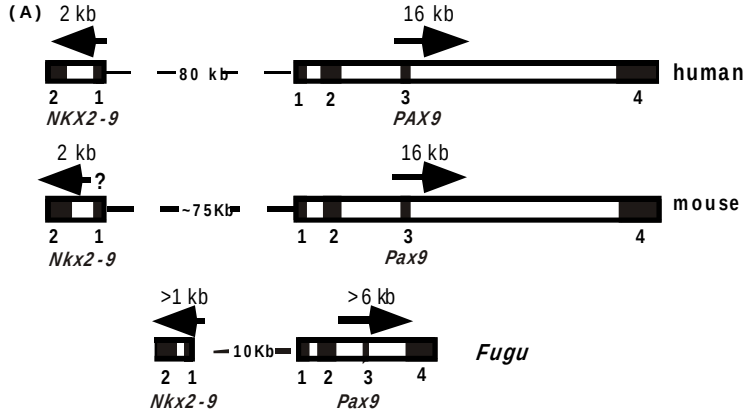
Given the high degree of sequence homology between human and mouse *Pax9* (Peters et al. 1997), it was assumed that the gene structure could as well be quite conserved. An indication for this structural conservation was the similar size and position of intron 1 (Peters et al. 1998b; Stockton et al. 2000) (Fig. 9). Assuming the position and the size of the mouse *Pax9* gene introns 2 and 3 to be as in the human situation, an inter-exon PCR strategy was adopted using primers lying on the exon sequences adjacent to the predicted exon-intron boundaries. This resulted indeed in the amplification of two bands corresponding to introns 2 and 3, respectively about 2.5 kb and 9.5 kb long. The exact intron-exon boundaries and the presence of splice sites were determined by sequencing of the PCR products (Fig. 9b).

As expected, Mouse *Pax9* consists of four exons with intron sizes similar to those in human (Fig. 9). As in the human gene, both the paired box and the octapeptide domain reside in the second exon (Santagati et al. 2001).

Sequence analysis of the human *PAX9* BAC sequence (about 200 kb) revealed the presence of the gene *NKX2-8* about 80 kb upstream of *PAX9*. *NKX2-8* belongs to the NK-2 family of homeobox transcription factors and it was found to be expressed in fetal liver and hepatocellular carcinoma (Apergis et al. 1998). *PAX9* and *NKX2-8* are oriented head to head (Fig. 9a). It was hypothesized that the same physical linkage might also be found in the mouse genome. By sequence comparison, mouse *Nkx2-9* (Pabst et al. 1998) was found to be the most similar to the human *NKX2-8*. Therefore, a PCR amplification was performed on the *Pax9*-positive mouse BAC clones by using primers from the first exon of mouse *Nkx2-9*. A PCR product with the expected size of 270 bp was obtained from six of the twelve *Pax9*-positive BAC clones. By sequencing, the PCR product was proven to correspond to the *Nkx2-9* exon. The result was confirmed by Southern blot hybridization (Fig. 10). Thus the association between *Pax9* and *Nkx2-9* in the mouse genome was confirmed (Fig. 9a). By inter-exon PCR, it was possible to determine the genomic organization of mouse *Nkx2-9*, as shown in Figure 9b.

In order to estimate the distance between the two genes, a partial restriction map of the BAC contig was established. BAC clones 3 and 17 were digested with a few rare cutters and then hybridized with the respective end probes and with a *Pax9* paired box probe and an *Nkx2-9* intron probe. This series of hybridizations allowed to create the map shown in Figure 11 and to approximately determine the position of the two genes at a reciprocal distance of about 70-75 kb.

Human *NKX2-8* and mouse *Nkx2-9* have never been described to be orthologous. The two genes were described and published almost contemporaneously (Apergis et al. 1998; Pabst et al. 1998), and even later, in further mapping studies, the human counterpart of mouse *Nkx2-9* was not indicated to correspond to the previously published *NKX2-8* (Wang et al. 2000). However, their sequence similarity and conserved association with *Pax9* together indicate that they indeed are orthologous genes. Therefore, hereafter in this work human *NKX2-8* will be referred to as *NKX2-9* (Santagati et al. 2001).



(B)

Gene	Organism	Exon 1	Exon 2	Exon 3	Exon 4
Pax9	human	>394	627	140	1550
	mouse	>194	627	143	1575
	Fugu	?	612	125	>870
Nkx2-9	human	>357	886		
	mouse	>374	851		
	Fugu	>157	>569		

(C)

Gene	Intron	Intron size (bp)	Organism
Pax9	1	aatgg GTgag.....803.....tgcAG agcca	human
		aatgg GTgag.....707.....tgcAG agcca	mouse
		catgg GTgag.....466.....accAG aactt	*Fugu*
	2	ccaag GTaag.....2938.....atcAG tgagc	human
		ccaag GTaag.....~2800.....ctcAG gagtg	mouse
		acaaa GTaag.....1446.....cccAG ttagt	*Fugu*
	3	gtcag GTgag.....9595.....ctcAG gcacc	human
		gccag GTgag.....~9500.....ctcAG gcacc	mouse
		cacag GTaac.....2285.....tgcAG acgcc	*Fugu*
Nkx2-9	1	ccctt GTaag.....768.....cctAG cctcg	human
		cctgt GTgag.....~700.....cccAG cctcg	mouse
		tatct GTgag.....265.....tgcAG cctct	*Fugu*

Fig. 9. (from Santagati et al. 2001)
(A) Genomic organization of *Pax9* and *Nkx2-9* and their physical association in the human, mouse and *Fugu* genomes. Human *NKX2-8* is referred to as *NKX2-9* (see the text). Solid boxes indicate exons. Transcription orientations are shown with arrows. Dashed lines show intergenic regions with sizes in kb. (B) Exon sizes of *Pax9* and *Nkx2-9* sequence given in bp. The size of mouse *Pax9* exon 1 refers to published cDNA sequence (NM_011041). (C) Exon/intron structure of *Pax9* and *Nkx2-9*. Exon and intron sequences are separated by a space. The GT/AG splicing signals are given in upper case.

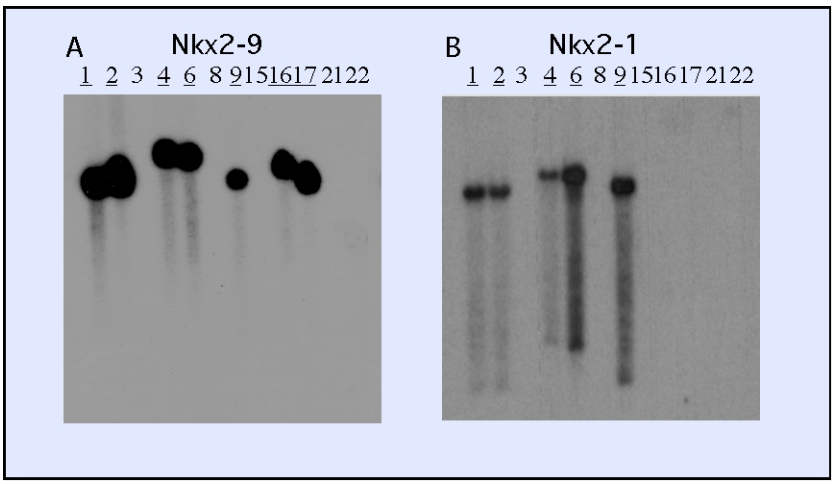


Fig. 10. Identification of *Nkx2-9* and *Nkx2-1* genes on the BAC contig. PFGE of NotI digested BAC clones hybridized with an *Nkx2-9* intron probe **(A)** and with an *Nkx2-1* intron probe **(B)**. Positive clones are underlined

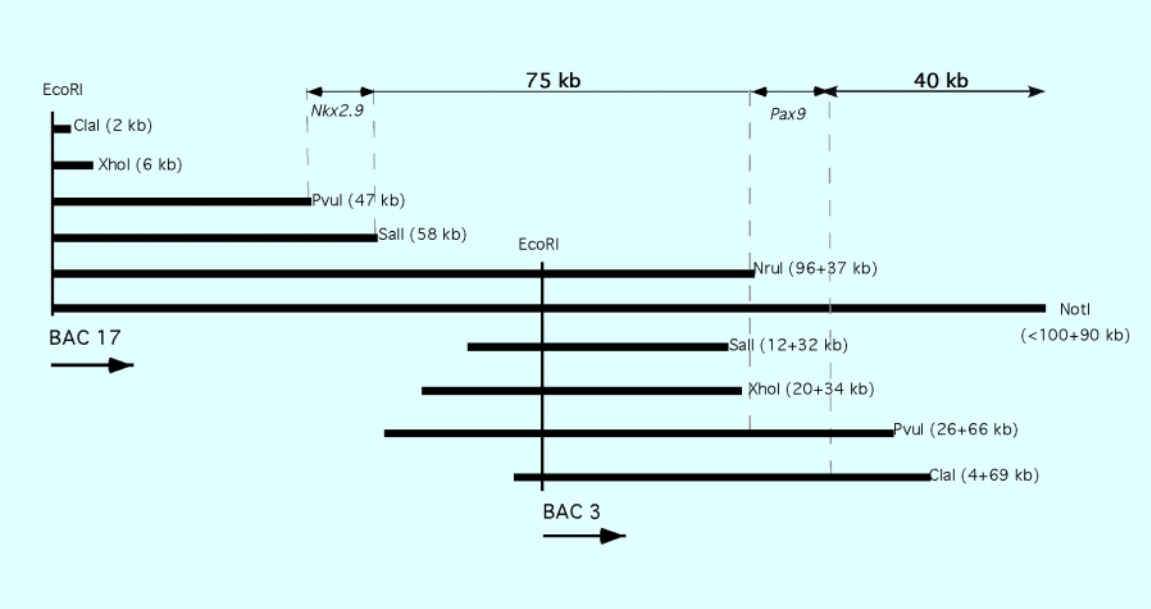


Fig. 11. BAC3 and BAC17 partial restriction map and positioning of *Pax9* and *Nkx2-9*
End fragments of BAC3 and BAC17 were obtained by digestion with NotI alone, or in combination with another rare cutter (ClaI, NruI, PvuI, SalI and XhoI), and they were identified with respectively BAC3 and BAC17 left-end probes. The same digests were as well probed with *Nkx2-9* intron and *Pax9* paired-box sequences, whose positions fall in the intervals demarcated by the double-arrowed lines. 75 kb is the resulting minimal distance between *Nkx2-9* and *Pax9*. 40 kb is the previously determined distance of *Pax9* 5' end to the NotI site.
EcoRI restriction sites mark the positions where the BAC sequences begin. Arrows indicate the direction. In brackets beside each restriction enzyme the corresponding fragment size; the BAC3 left end splits some of the restriction fragments in two parts whose sizes before and after the splitting point are given in brackets.

In Wang et al. a physical linkage between *Nkx2-9* and *Nkx2-1* both in the mouse and in human is described (Wang et al. 2000). Analysis of the human sequence in the databank revealed that *NKX2-1* is located around 65 kb far apart and downstream of *NKX2-9*. In order to check whether the mouse *Nkx2-1* gene was present in the BAC contig, a Southern blot hybridization with an Nkx2-1 3'-UTR specific probe was performed on NotI digests of the BAC clones (Fig. 10). Indeed five of the clones turned out to contain the *Nkx2-1* gene.

4.2. Identification of a conserved syntenic genomic region in *Fugu rubripes*

4.2.1. Identification of Fugu Pax9 gene

As described later on in this work, one of the approaches for the identification of DNA sequences with a functional role in the regulation of the *Pax9* gene was to conduct a comparative sequence analysis among different species. This would allow to find conserved

elements outside the coding sequences that could be taken as candidates for further functional analysis.

The availability of the human sequence on the databank and the ongoing sequencing of the mouse BAC clones through the NIH-funded Mouse Genome Sequencing Network (Trans-NIH Mouse Initiative) brought about the idea of sequencing the *Pax9* genomic region from an out-group animal species, other than a mammal, that could allow a more stringent analysis and the search of widely evolutionarily conserved elements within the vertebrates.

The decision to clone and sequence the *Pax9* gene from the Japanese pufferfish *Fugu rubripes* was taken in the light of the latest research on this species for genomic and evolutionary analysis. The advantage of using *Fugu* as a model species for this kind of genomic studies is that it has a genome with a size of 400 Mb, that is 7.5 times smaller than the human genome. In spite of that, the amount of coding sequence is approximately the same. On the contrary, the *Fugu* genome contains less than 10% of all of the forms of repetitive DNA and it shows a conspicuous shrinkage of inter- and intragenic non-coding sequences, that results in a higher gene density (more genes in a shorter genomic region) and in shorter genes, which are easier to study (Elgar et al. 1996; Brunner et al. 1999; Elgar et al. 1999).

A Fugu cosmid library was screened by hybridization with a probe obtained by PCR from *Fugu* genomic DNA using degenerated primers based on the most conserved sequences of the paired box among different species (see PCR table 4, no. 9). The library screening resulted in 20 positive clones with different signal intensities (Jürgen H. Blusch's personal results). BamHI digests of the clones were further examined by hybridization at low stringency with a mouse *Pax9* paired-box probe. Only one of them (clone ICRFc66D2193) was positive (Fig. 12). In order to confirm the result, the same clone was digested with a different restriction enzyme (EcoRI) and re-probed with the same mouse *Pax9* sequence. Overlapping restriction fragments, positive for this probe, were subcloned and sequenced (GenBank: AF266754). The sequence comparison between the *Fugu* sequence and other vertebrate Pax genes showed the highest homology to *Pax9*, particularly to the type A isoform of zebrafish *Pax9* (Nornes et al. 1996), indicating that the cloned genome sequence encoded *Fugu Pax9*. The remarkable homology to zebrafish *Pax9a* allowed to determine the putative cDNA sequence and the positions of exon/intron boundaries on the *Fugu* gene.

The prediction for the boundaries was supported by the presence of conserved splicing donor and acceptor sites (Fig. 9c). The *Fugu Pax9* gene consists of four exons like the mouse and the human orthologous counterparts, but it extends for a shorter genomic sequence of only about 6 kb in length, being the introns in general reduced in size (Fig. 9c). The deduced amino acid sequence was aligned with the Pax9 protein sequence from other species (Fig. 13). Zebrafish Pax9a showed overall 86% identity and 90% similarity, while both mouse Pax9 and human PAX9 showed 73% identity and 81% similarity with the *Fugu* counterpart. Inside the paired domain the identity increased to 98% among all the species and of the few residue changes only the Gly at position 109 instead of the conserved Val was strictly specific for the *Fugu* gene. The octapeptide domain showed conversely 100% conservation. The whole protein is slightly shorter than the orthologous counterparts in the other species. It consists namely of a total of 332 aminoacids (aa), while the human, mouse, chick and zebrafish proteins contain between 341 aa and 343 aa (Santagati et al. 2001).

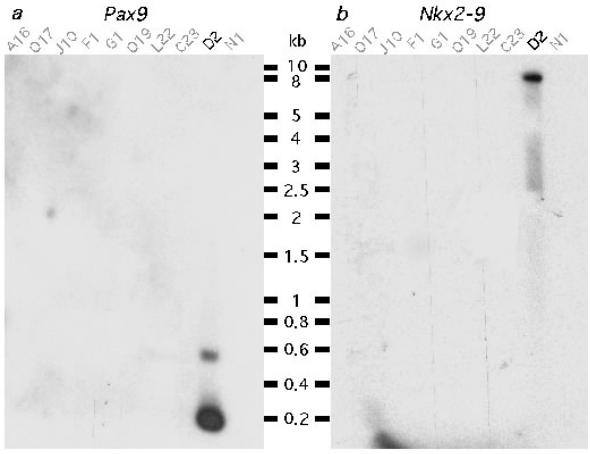


Fig. 12. Identification of *Fugu Pax9* and *Nkx2-9* genes
Southern blot hybridization of 10 of the 20 PCR pre-selected Fugu cosmid clones BamHI digested (see text).
Same membrane hybridized at low stringency with a mouse *Pax9* paired-box probe (**a**) and a mouse *Nkx2-9* homeobox probe (**b**). The same clone (D2) was positive for both probes.
A similar hybridisation with the remaining 10 clones did not produce any positive signal (not shown)

4.2.2. Identification of Fugu Nkx2-9 gene

In order to know whether the physical association between *Pax9* and *Nkx2-9* was also conserved in the *Fugu* genome, the *Fugu Pax9* cosmid clone was hybridized with a mouse *Nkx2-9* homeobox probe. Remarkably, the clone turned out to be also positive for this probe (Fig. 12). By sequencing of positive restriction fragments from the cosmid clone (GenBank: AF267536), it was possible to identify a gene that showed the highest similarity

to mouse *Nkx2-9* and human *NKX2-9*. This sequence similarity and the physical linkage to *Pax9* together strongly suggested that this NK-2 gene is *Fugu Nkx2-9*.

This was the first *Nkx2-9* gene identified in non-mammalian vertebrates arguing against the hypothesis that it could exclusively be a mammalian specific gene (Wang et al. 2000). The distance between *Nkx2-9* and *Pax9* in *Fugu* is about 10 kb, and they are orientated head to head as in the human genome (Fig. 9a). Again as for *Fugu Pax9*, the unavailability of the cDNA sequence of the gene made it necessary to deduce its structure in term of exon-intron boundaries based on the alignment with the mouse and human genes. The demarcation of putative boundaries was supported by the identification of conserved donor and acceptor splicing sites (Fig. 9c).

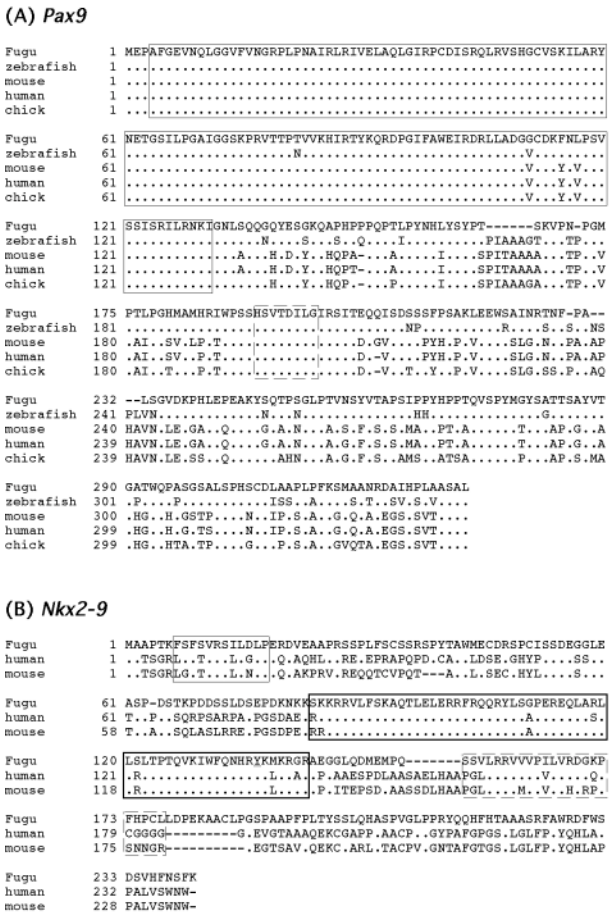


Fig. 13. (from Santagati et al. 2001)
Deduced amino acid sequences of *Fugu* Pax9 (**A**) and Nkx2-9 (**B**), aligned with the respective counterparts from other species (zebrafish Pax9a: ACC60034; mouse Pax9: P472421; human PAX9: NP_006185; chick Pax9: P55166; human NKX2-9: AAC71082; mouse Nkx2-9: CAA75751).
Conserved domains are boxed: Pax9, the paired domain (plane box) and the octapeptide (dashed box); Nkx2-9, the TN domain (plane box), the homeodomain (bold box), and the NK2-specific domain (dashed box). The tyrosine 54 in the homeodomain is underlined. A dot indicates an identical residue as in *Fugu*, and a dash represents a gap.

On the basis of the predicted coding sequence, a corresponding aminoacid sequence was deduced. The alignment with the mouse Nkx2-9 and human NKX2-9 is shown in Figure 13b. The overall homology among the Nkx2-9 protein in the three species is in general very low and it is restricted to the three conserved domains that demarcate the designation to the

NK-2 gene family (Harvey 1996). The NK-2 domain is present in the three species but quite divergent, especially in the N- and C-terminal parts (Harvey 1996). On the contrary, the homeodomain is highly similar among the three species and contains a tyrosine at position 54, unique to NK-2 proteins (Harvey 1996). The third domain is the TN domain at the N-terminus of the proteins. This domain was described to be absent in the mouse Nkx2-9 (Pabst et al. 1998), but Figure 13b shows that it is indeed present and has a fairly good homology to the human and pufferfish counterparts (Santagati et al. 2001).

4.3. Investigations on *Pax9* mRNA

4.3.1. Northern blot analysis

The following step in the characterization of *Pax9* structure was the determination of the transcript size. Two mouse *Pax9* cDNA clones of respectively 1.6 and 2.4 kb in length were previously isolated and covered a total of 2.5 kb sequence (Neubüser et al. 1995). However, Northern blot analysis on various mouse tissues revealed two transcripts, a small one (2.2 kb) and a longer one with an actual size estimated between 4.7 and 5.3 kb, suggesting that a relative large part of the transcribed sequence had not been identified (Neubüser et al. 1995; Peters et al. 1997). Similarly, three transcripts (5.3, 3.5 and 2.1 kb) were detected in total RNA extracts of human esophagus (Peters et al. 1997).

In order to confirm these data and complete the sequence information of *Pax9* transcript, the Northern blot analysis was repeated. A detection of *Pax9* transcript was only possible if at least 2 to 4 µg of poly(A)-RNA were used for the hybridization. The RNA was separately extracted from limb buds and tails of 11.5 dpc mouse embryos of C57BL/6 strain. The result of the hybridization with a paired-box probe is shown in Figure 14a. Two bands could be detected. The highest one about 4.5 kb long probably corresponded to the 4.7 kb band described (Neubüser et al. 1995). An additional band of lower intensity, above the 2.35 kb band of the molecular weight marker, was thought to represent the RNA isoform corresponding to the longer cDNA clone. Nevertheless the stronger intensity of the higher band suggests that the 4.5 kb transcript represents the main RNA isoform of mouse *Pax9 (*Fig. 14). Unfortunately, further attempts of hybridization with two different probes from the 3'-UTR did not produce appreciable results.

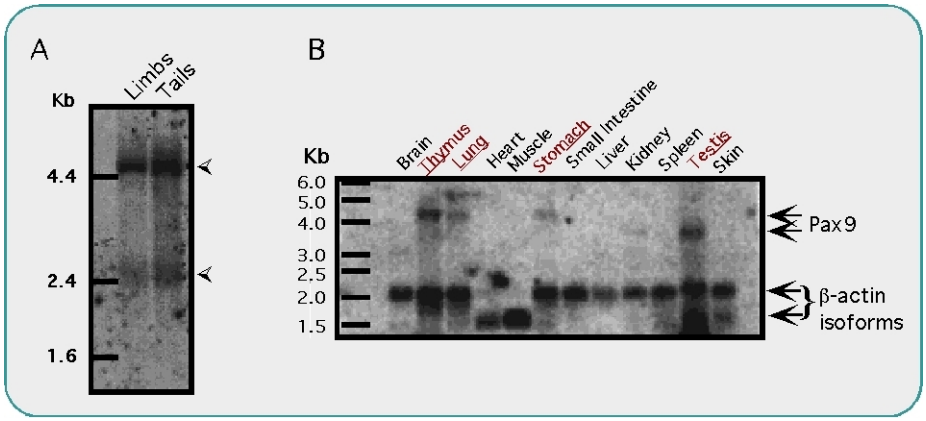


Fig. 14. Northern blot analysis of *Pax9*
(A.) RNA extracts from 11.5 dpc mouse embryo tails and limb buds hybridised with a paired-box probe. Two bands are shown by the arrowheads. **(B)** Origene commercial RNA membrane with adult mouse tissues. Double hybridization with a *Pax9* exon4 probe (positive bands in thymus, lung, stomach and testis) and a β-actin probe (heart and striated muscle samples display a shorter isoform)

Given the remarkable size difference between the detected *Pax9* mRNA and the available cDNA sequence, an additional experiment was required to prove that the 4.5 kb band was no artifact and indeed corresponded to a real transcript.

A hybridization for *Pax9* RNA was carried out on a commercial Northern blot membrane from Origene, which included mRNA extracts from 12 different adult mouse tissues. This time a 600 bp probe was used, which was amplified from part of the fourth exon (exon4 probe), including the end of the coding sequence and the beginning of the 3'-UTR. The hybridization provided consistent indications with the results obtained from the embryonic extracts (Fig. 14b). A band between 4 and 5 kb could be detected on RNA extract from thymus, lungs and stomach, while a shorter band between 3 and 4 kb in length was observed in the testis sample. These data confirmed that the 4.5 kb transcript is indeed specific. *Pax9* expression in the thymus had already been described (Neubüser et al. 1995; Peters et al. 1997) and represented a positive control for the hybridization. On the contrary, no expression in the lungs and in the stomach had been detected so far (Neubüser et al. 1995; Peters et al. 1997). A further confirmation of *Pax9* expression in the lungs of adult mice came from RT-PCR analysis (not shown). The expression in the stomach was not confirmed with additional experimental evidence. However, a contamination of esophagus tissue in this sample, where *Pax9* is highly expressed (Peters et al. 1997), cannot be ruled out.

The lower band from the testis sample (Fig. 14b) was initially thought to represent a new isoform of *Pax9*, possibly generated by tissue specific alternative splicing of the transcript. Nevertheless, RT-PCR analysis on RNA extracts of adult mouse testis did not bring about reproducible and convincing results.

4.3.2. RACE-PCR analysis

Once ascertained that a large portion of *Pax9* mRNA sequence was still unknown, it became necessary to extend the sequence information on both sides of the transcript. The determination of the 5'-end or transcription start site of the gene was required in order to locate the promoter sequence. A first attempt was made by screening the RIKEN database of mouse full-length cDNA clones (Kawai et al. 2001; Bono et al. 2002) by using a specific *Pax9* 3'-UTR sequence. Three clones were pulled out (1110048-E04, 2700046-A17, 2700028-N19). Restriction analysis showed that they represented the same 1600 bp fragment and sequencing of one of the clones revealed that it corresponded to the shorter *Pax9* cDNA clone described in Neubüser et al. 1995. As an alternative, 5'- and 3'-RACE-PCR approach was carried out using total RNA extract from 11.5 dpc mouse embryos as starting material. Two different RACE-PCR systems were used for this purpose as described in "Materials and Methods". The first one was based on the classical method in which the first strand cDNA is synthesized with a specific primer from the mRNA template and then tagged at its 5' end with a linker primer; the tagged sequence is subsequently amplified with gene specific primers. Only one PCR product was clearly visible upon this amplification and it was comparable in size with a fragment obtained as control from the known cDNA clone, suggesting that it did not contain any relevant additional sequence information. However, the whole PCR mix was subcloned, and 30 of the screened colonies turned out to be positive for *Pax9* 5' sequence. The majority of them (29/30) contained the same small PCR product. Sequencing of one of these clones showed that this fragment comprised a 26 bp longer 5' sequence than the published 5' end. This new 5' end of *Pax9* transcript was regarded as a putative transcription start site (TSS-B in Fig. 15). One of the RACE clones contained an even longer 5' sequence than the others, although the corresponding PCR product was invisible on agarose gel. Sequencing of this clone revealed that it contained further 431 bp from TSS-B. Comparison of these 431 base pairs with the

mouse genomic sequence revealed that the first 261 bases originated from a novel exon (hereafter designated as exon 0), and the following 170 bases were from the region immediately upstream to TSS-B (Fig. 15a). Therefore, these 170 bases are considered as part of exon 1. Exon 0 was located about 3.7 kb upstream of exon 1 (Fig. 15a). The presence of the splicing donor site (GT) following the 3' end of exon 0 and the splicing acceptor site (AG) preceding the 5' end of exon 1 was confirmed on the genomic sequence, as shown in Figure 15b and 15c respectively. RT-PCR analysis further confirmed the existence of this additional 5' UTR sequence on *Pax9* cDNA (data not shown). The 5' end of exon 0 was regarded as another putative transcription start site (TSS-A). Judging from the 5'-RACE data, it is likely that the majority of *Pax9* transcripts start around TSS-B carrying a truncated exon 1 and completely lacking exon 0, while only a minor population starts around TSS-A and contains exon 0 and a complete exon 1. In the human *PAX9* and *Fugu Pax9* genome sequences, no homology to mouse exon 0 was found. This suggests either that the sequence of exon 0 might not be conserved because it is only part of the untranscribed region, or that exon 0 might be a mouse specific exon (Santagati et al. 2001). Since the RACE-PCR data were obtained from whole embryo RNA extracts, it was hypothesized that there could be an alternate distribution of the two *Pax9* transcripts in the different *Pax9* expressing domains due to differential promoter usage in a tissue specific manner. Whole mount *in situ* hybridization was accomplished on 10.5 dpc mouse embryos using a specific RNA probe that covered the 431 bp of the new 5' sequence. No difference in the expression pattern was observed in comparison to a hybridization control with a *Pax9* coding sequence RNA probe, suggesting no preferential usage of the transcription start site A in any *Pax9* positive tissue during mouse development (not shown).

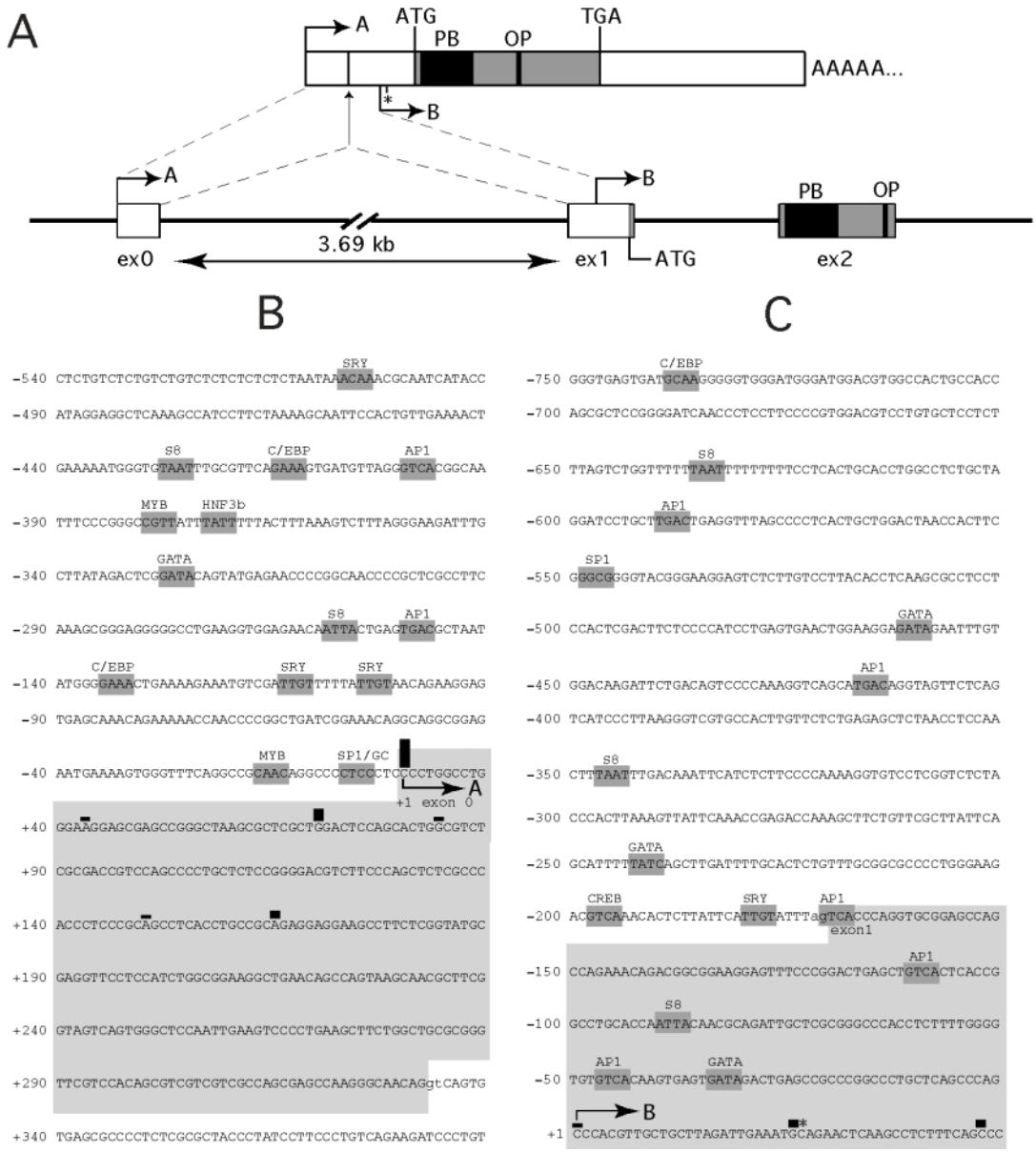


Fig. 15. (modified from Santagati et al. 2001) (A) Schematic representation of the mouse *Pax9* transcript (top) and the genomic organization of 5'-exons (bottom). Coding region is shadowed, and the positions of the start and stop codons are indicated. Both 5'- and 3'- UTRs are shown as open boxes. The arrows A and B indicate the positions of putative transcription start sites (TSS-A and TSS-B). The asterisk (*) indicates the position of the previously published 5'-end. PB: paired box; OP: octapeptide. (B) and (C) Nucleotide sequence and putative transcription binding sites of the mouse *Pax9* promoter regions, including upstream (B) and downstream (C) putative transcription start sites. Black blocks above single nucleotide positions indicate the 5'-ends of different 5'-RACE clones and the height of each block represents the frequency of the respective clone. The bases 5' or 3' from the most upstream transcription start sites (indicated by the arrows A and B) are numbered negatively or positively, respectively. Four bases in the core motifs of putative transcription binding sites identified by MatInspector are marked by shadowed boxes. The splicing signals following exon 0 and preceding exon 1 are shown in lower case.

The extended 5'-UTR elongated the *Pax9* cDNA sequence to about 3 kb, but yet it did not accomplish for the discrepancy in length with the 4.5 kb Northern blot band. For this reason a second 5'-RACE-PCR method was applied, in which the real 5' end of the mRNA is directly tagged with a linker before first strand cDNA synthesis. The tagging occurs only for the mRNA molecules that carry a 5'-cap avoiding in such a way tagging and amplification of false 5' ends produced by RNA degradation.

Briefly, the presence of the two alternative transcription start sites was again confirmed but no further extension of the *Pax9* cDNA sequence was obtained. Likewise, the use of primers designed from the genomic region upstream of the TSS-A produced no results.

Sequencing of several RACE clones showed that in reality no single base position could be identified as precise transcription start point. Rather, different 5'-end positions could be localized for each clone within an interval of about 140 bp in case of the TSS-A and 70 bp in case of TSS-B (Fig. 15b and 15c).

The genomic sequences including either of the putative transcription start sites were regarded as putative promoter sequences. Although MatInspector analysis (Quandt et al. 1995) of these putative promoters detected several potential transcription binding sites (Fig. 15), no typical TATA box or CCAAT box could be identified, suggesting that *Pax9* transcription is driven by TATA-less promoters (Santagati et al. 2001). The absence of a precise transcription start point is in agreement with the absence of a TATA-box. Some TATA-less promoters have been indeed described to drive initiation of transcription from a variegate number of nucleotide positions even within intervals of hundreds of base pairs, instead of defining one single start point (Smale 1997).

Similar RACE-PCR analysis was conducted on the 3' side of *Pax9*. By using specific primers on the fourth exon, as described in "Materials and Methods", two different 3'-ends about 900 bp apart from each other could be amplified. These two cDNA sequences with respectively a short (about 430 bp) or a long (1320 bp) 3'-UTR were already isolated and described in Neubüser et al. 1995. No longer 3' sequence could be obtained. The possibility that these sequences could represent two real transcripts generated by alternative polyadenylation signaling was ruled out. The poly(A) tail of the shorter 3'-UTR corresponds to a 10 adenosine stretch within the 3' genomic region of *Pax9*, suggesting that this cDNA clone is an artifact produced by wrong alignment of the poly(T) primer. On the

contrary, no adenosine stretch was found in the genomic region around the second 3'-end, which has therefore a real poly(A)-tail. However, no canonical polyadenylation signal could be found upstream of the poly(A)-tail, leaving open the possibility that this is not the real 3'-end of the gene.

4.4. *Pax9* promoter analysis in cell culture

The determination of two transcription start sites for *Pax9* was an important landmark for the localization of the promoter sequence(s). The genomic regions encompassing these sites were likely to be bound by the basic RNA polymerase complex for the initiation of transcription. However, an experimental proof was still necessary to definitely assess whether these sequences really performed the putative function.

A cell culture system is very suitable to establish a promoter assay. For this purpose, stable cell lines were sought that would express *Pax9*. A promoter assay in such cell lines would not only allow to define the minimal *Pax9* promoter region, but it would as well conceivably provide a tool for the identification of specific regulatory elements.

4.4.1. Choice of Pax9 expressing cell lines

In Peters et al. a possible role of *Pax9* in the formation of stratified squamous epithelia is discussed. This observation was based on the finding of *Pax9* expression in the squamous epithelium of the esophagus both in the adult mouse and in human (Peters et al. 1997). A remark about a possible role of *Pax9* in the formation of squamous cell carcinomas was added in conclusion of the article, in the light of the fact that other Pax genes had been found to be associated to various types of tumors (see "Introduction"). AT478 is a cell line derived from a spontaneous mouse squamous carcinoma (Guttenberger et al. 1990). Western blot analysis on RNA extract from this cell line had already been successfully performed in the search for *Pax9* expression (H. Peters, unpublished results). Considering this prior investigation, the AT478 cells were chosen as one potential *Pax9* expressing cell line for promoter assay. Another cell line available in the laboratory was as well thought to express *Pax9*. These cells (MLB13myc) were a clonal lineage obtained from 13.5 dpc mouse embryo limb buds after transformation with a v-myc vector (Rosen et al. 1994). Of the different isolated clones, one in particular (clone no. 14) had shown early skeletal

progenitor features and ability to differentiate in a chondroblastic and then in an osteoblastic line, upon treatment with BMP-2. As already described in the introduction section, *Pax9* is known to be expressed in the developing mouse limbs. This made these cells likely to express *Pax9* and suitable for the desired experimental approach.

Before setting up the promoter assay, an RT-PCR analysis was accomplished in order to verify that the two selected cell lines indeed expressed *Pax9*. The Figure 16 shows a panel of PCR amplifications using three different sets of primer pairs from the *Pax9* cDNA sequence. First strand cDNA synthesized from total RNA extracts of these cell lines was used as a template. In addition a positive control was performed in parallel, using total RNA from 11.5 dpc mouse embryo as starting material. It could be observed that both cell lines express *Pax9* mRNA, as expected. However, in neither line the sequence corresponding to the newly discovered exon 0 and extended exon 1 could be amplified, suggesting that only one promoter (promoter B) is active in these cells.

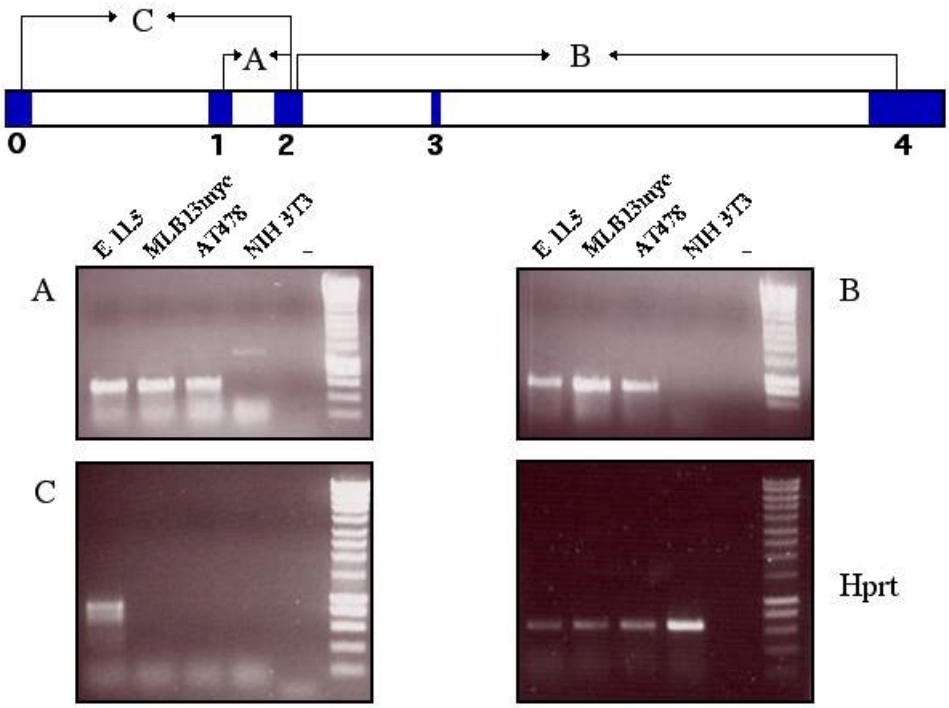


Fig. 16. *Pax9* RT-PCR analysis on cell lines. On top schematic representation of mouse *Pax9* gene structure, exons are blue boxes numbered from 0 to 4. A, B and C indicate the segments covered by the RT-PCR amplifications shown in the agarose gel photos below.
RNA samples were extracted from the following sources
E 11.5: mouse embryo at 11.5 dpc - AT478: mouse squamous cell carcinoma - MLB13myc: mouse limb bud cells (E 13.5) - NIH 3T3: mouse fibroblasts
Hprt: control RT-PCR for the hypoxanthine phosphoribosyltransferase gene

A third cell line was checked for *Pax9* expression. NIH-3T3 cells are transformed embryonic fibroblasts widely used for transfection experiments. These cells resulted clearly negative for *Pax9* mRNA and they were therefore used for further experiments as a negative control cell line.

4.4.2. Luciferase reporter gene based promoter assay

The promoter activity of a genomic sequence can be tested in a cell culture system by cloning the test fragment in front of a reporter gene and then measuring the gene transcription efficiency.

The firefly (*Photinus pyralis*) luciferase gene is commonly used as reporter system for this type of experimental approach. The level of transcription is reported as activity of the luciferase protein, which can be measured as photon emission upon oxidation of the luciferin substrate.

Two genomic fragments of about 2 kb in length and containing either transcription start site were amplified and cloned in the luciferase reporter vector pGL3-basic (see "Materials and Methods" and Fig. 17). The fragment A was called promoter A and it covered the region from position -1710 to position +266 of the TSS-A (+1 in Figure 15b). Similarly, promoter B was the genomic fragment from position -1941 to position +157 of the TSS-B (+1 in Figure 15c).

As shown in Figure 17, promoter A did not have any significant activity in any of the cell lines compared to the SV40 control promoter sequence present in the pGL3-promoter vector. This result was not surprising, considering that the transcription start site was not active in the two *Pax9* positive cell lines as suggested by the RT-PCR data. Moreover, the efficiency of this promoter was observed to be very low also in its physiological context and it was probably undetectable in the used experimental system.

On the contrary, promoter B exhibited quite a good activity when compared to the SV40 promoter (Fig. 17), although no specificity for the AT478 and MLB13myc cell lines was observed. The NIH-3T3 cells, which were proven not to express *Pax9*, displayed an equivalent level of luciferase activity. This result indicated that the 2 kb fragment contained a basal promoter sequence, which could be functioning in any cell type.

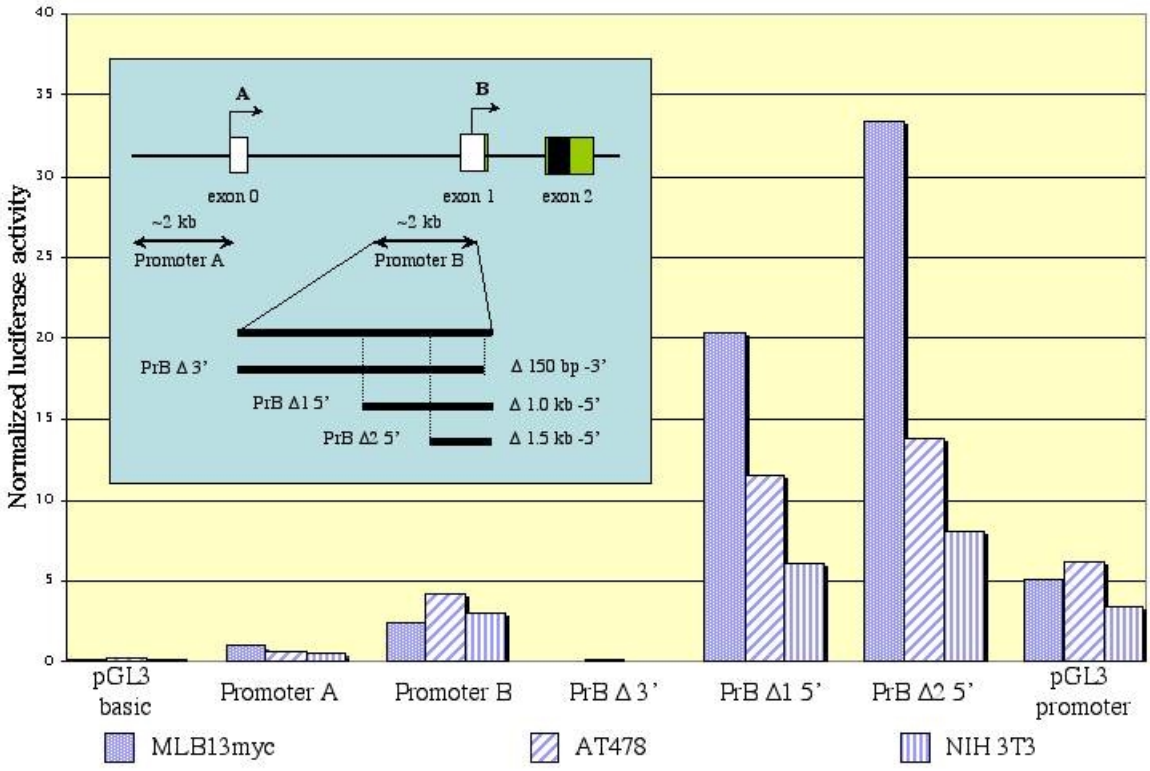


Fig. 17. Histogram of luciferase activity with promoter constructs. In the small square in the top left corner schematic position of the putative transcription start sites A and B and exons 0, 1 and 2. Paired-box domain (in black) is shown inside the coding sequence (in green). Double-arrowed lines represent genomic fragments (promoters A and B encompassing the two TSSs) cloned in front of the luciferase reporter. Promoter B is enlarged below with the series of three deletion constructs (see text).
The graph represents the promoter activity of each construct in the three cell lines estimated as relative firefly luciferase activity, normalized for the renilla luciferase. Negative control (pGL3 basic) and positive control (pGL3 promoter) are included.

In order to demonstrate that the observed promoter activity was not an experimental artifact, a control fragment (Promoter B') was tested. Promoter B' was only 152 bp shorter on the 3'-end compared to the Promoter B fragment, so that the transcription start interval would be deleted. This construct did not show any transcriptional activity, proving that the sequence deprived of its basal structural features had completely lost the promoter function (Fig. 17).

Conversely, deletions of the 5'-end of Promoter B not only did not affect the promoter activity; in fact they even increased it. Construct PromoterB-ΔK carried a 1087 bp 5'-deletion and, compared to PromoterB, it exhibited 10 fold higher activity in the

MLB13myc cell line, nearly 3 fold higher in AT478 and 2 fold in the NIH-3T3. PromoterB-ΔS was generated by deleting 1576 bp from the 5'-end of PromoterB and its activity was 14 fold higher in the MLB13myc, more than 3 fold in the AT478 and 2.5 fold in the NIH-3T3 cells (Fig. 17)

Remarkably, the increase in promoter activity observed with the two 5'-deletion fragments related to an increase in specificity. The *Pax9* positive cell lines, in particular the MLB13myc, displayed a higher responsiveness to the two smaller constructs compared to the negative line NIH-3T3. If this cell line specificity was due to the actual function of *Pax9* specific elements it was not investigated. On the other hand, little attention was given to this observation, since the same specificity was not reproducible with the 2 kb construct.

4.5. Comparative sequencing

4.5.1. Sequence alignment through PIP analysis

As already mentioned before, one of the two approaches employed in this work for the identification of the *Pax9* regulatory elements was based on cross-species sequence comparison.

The identification of evolutionarily conserved non-coding sequences among orthologous genomic regions of different species has been proposed to be one of the most powerful guides for the localisation of functional elements (Koop and Nadeau 1996; Duret and Bucher 1997; Hardison 2000; Wasserman et al. 2000). The advantage of this method is that it can be applied for studies to a genome-scale. Successful applications of this method have already led to the finding of specific enhancer sequences for a number of genes (Göttgens et al. 2000; Ishihara et al. 2000; Bagheri-Fam et al. 2001).

One of the best tools for the alignment of large genomic sequences is the PipMaker program. PipMaker is an automated program on the World-Wide Web (http://bio.cse.psu.edu/) for generating alignments and pips (percent identity plots). A PIP shows the position in one sequence of each aligning gap-free segment and plots its percent identity. The advantages of this server are that it can analyze long sequence files, containing as many as millions of nucleotides, and it is able to compare the complete sequence from one species with an incomplete sequence from a second (Schwartz et al. 2000). A PIP analysis is a solid method for the identification of conserved non-coding

sequences (CNSs) that truly represent gene specific regulatory elements (Hardison 2000; Loots et al. 2000).

In this work, the genomic sequences encompassing the *Pax9* gene of *Homo sapiens*, *Mus musculus* and *Fugu rubripes* were used. The human sequence (GenBank AL079303) consisted of a nearly 200 kb genomic region including the *NKX2-9* gene, from position 20205 to position 18196*, and the *PAX9* gene, from position 99316 to position 115363.

The mouse sequence derived from the combination of two BAC clones, BAC6 and BAC15 in this work corresponding to clones 136M3 and 327I21 respectively of the RPCI-23 Mouse BAC library. BAC6 sequence consisted of a 215 kb gapped sequence of 6 unordered pieces. BAC15 sequence was conversely ordered and completed to final 219 kb in length. From previous BAC clone mapping experiments, it was known that BAC6 contained both the *Nkx2-9* and *Pax9* genes, while BAC15 was more shifted towards *Pax9* and did not bear *Nkx2-9* (Fig. 10). *Pax9* 5'-end (TSS-A) was located at position 6321 of BAC15 and the gene stretched out to position 26086.

PIP analyses were performed using the human sequence as a base template on which the two mouse BAC sequences were alternately aligned. Figure 18 shows the summarizing results of this analysis. Almost the entire overlapping genomic region displayed a rather high overall homology degree, showing a remarkable conservation level even outside of the coding sequences, in the intergenic regions and within the introns. Dense blocks of homologous fragments alternated with long and short stretches of no homology mainly characterized by the presence of various repetitive DNA.

Due to the general abundance of vastly interspersed homologous fragments between the human and the mouse sequence, it was necessary to set a significance threshold for the definition of conserved non-coding sequences (CNSs). Loots et al. already adopted a stringent definition of conservation, requiring an ungapped alignment of at least 100 bp and at least 70% identity (Loots et al. 2000). These parameters allowed the identification of several CNSs throughout the whole region.

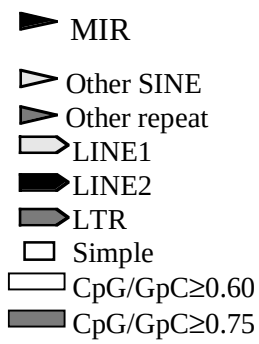
A BLAST search against the EST and non-redundant databanks was performed for each of these elements in order to verify whether they really represented non-coding sequences. Surprisingly, a set of CNSs lying downstream of *Pax9* matched the cDNA sequence of an

* The gene lies in a backward orientation with respect to the sequence annotation

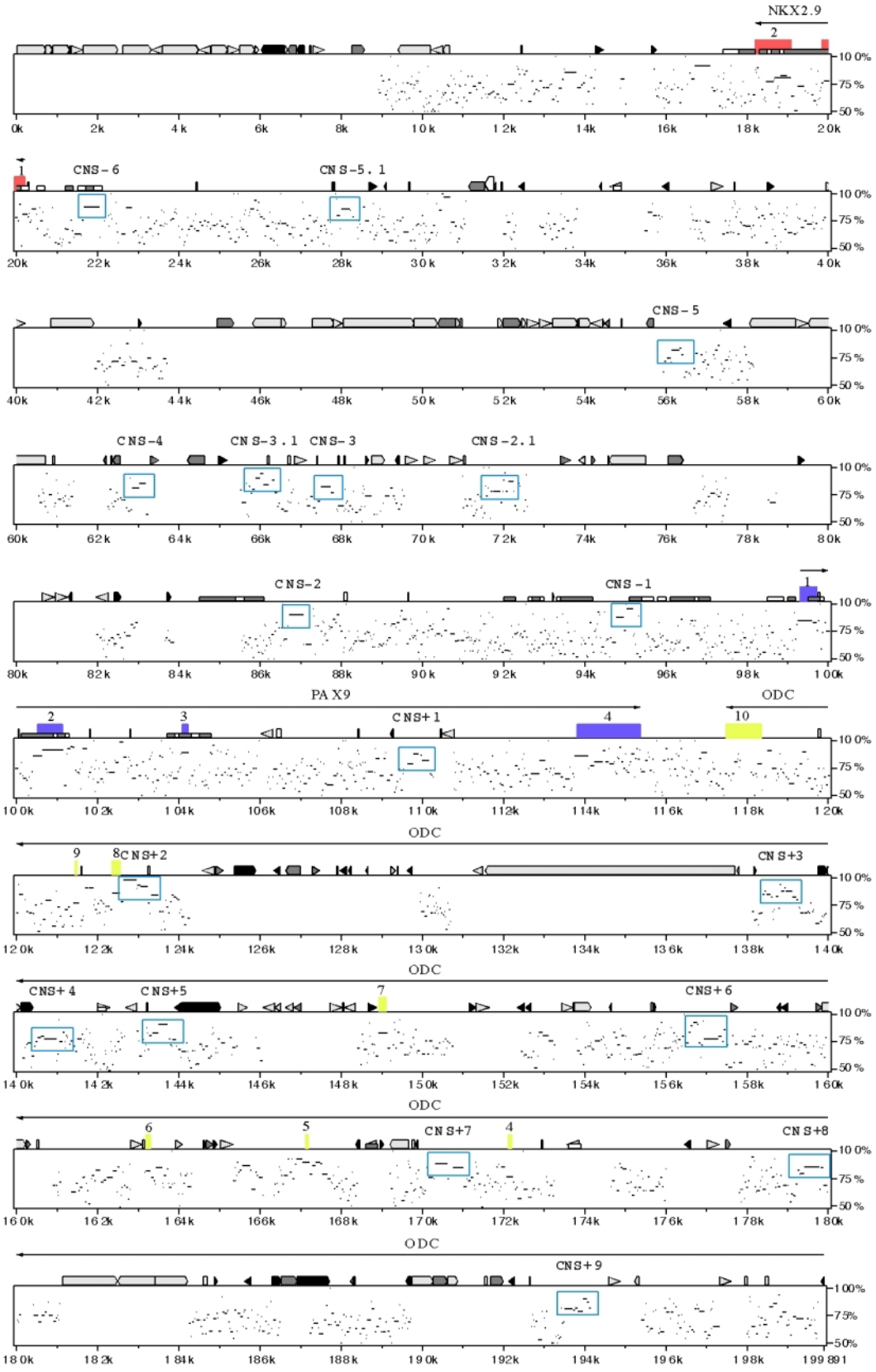
additional gene. This gene encodes the mitochondrial oxodicarboxylate carrier (ODC), a conserved ubiquitous protein localized in the inner mitochondrial membrane and performing a central role in aminoacid metabolism (Fiermonte et al. 2001). No mapping data were available for the gene at the time of this finding, until they were uncovered by a recent publication (Das et al. 2002). The gene lies only approximately 2 kb from the 3'-end of *Pax9*, in a tail-to-tail orientation, and it consists of ten exons, which are scattered over a vast genomic region of about 500 kb, despite the relative short length of its cDNA sequence (2000 bp).

The last seven exons were included in the human sequence of interest and coincided with some of the described CNSs, demonstrating the conserved localization of the mouse *Odc* counterpart. The *ODC* exon sequences were therefore excluded from further investigations, while the remaining set of downstream CNSs were finally located within the introns of the gene (Fig. 18).

Fig. 18 (next page). PIP analysis of the human/mouse alignment. The graph shows about 200 kb of the human genomic sequence (see text) including *Nkx2-9*, *Pax9* and *Odc* genes, whose exons are represented as red, blue, and yellow boxes respectively. Note that the sequence does not include the first three exons of *Odc*. The arrow-lines above indicate the length and direction of the three genes. Homology matches with the mouse sequence (see text) are represented as dots and dashes, where the length corresponds to the length of the matching sequence and the height in the plot to the homology degree (scaling between 50% and 100% on the right side of each row).
CNS elements consist of groups of adjacent homologous fragments outside the coding sequences with a total length of at least 100 bp and an average homology of at least 70% (see text), and they are boxed by open blue rectangles. The CNSs upstream and downstream of *Pax9* have negative and positive numbers respectively.
The human sequence was masked against repetitive DNA with the RepeatMasker from the BCM Search Launcher server.



The legend describes the meaning of the arrows and the boxes above each row, including the various masked repetitive sequences and the CpG rich regions. Note that the regions particularly rich in repetitive elements show the lowest abundance of homology matches.


NKX2.9
2
100%
75%
50%
0k 2k 4k 6k 8k 10k 12k 14k 16k 18k 20k
1
CNS-6
CNS-5.1
20k 22k 24k 26k 28k 30k 32k 34k 36k 38k 40k
CNS-5
40k 42k 44k 46k 48k 50k 52k 54k 56k 58k 60k
CNS-4
CNS-3.1
CNS-3
CNS-2.1
60k 62k 64k 66k 68k 70k 72k 74k 76k 78k 80k
CNS-2
CNS-1
1
80k 82k 84k 86k 88k 90k 92k 94k 96k 98k 100k
PAX9
ODC
2
3
CNS+1
4
10
100k 102k 104k 106k 108k 110k 112k 114k 116k 118k 120k
ODC
9
8
CNS+2
CNS+3
120k 122k 124k 126k 128k 130k 132k 134k 136k 138k 140k
ODC
CNS+4
CNS+5
7
CNS+6
140k 142k 144k 146k 148k 150k 152k 154k 156k 158k 160k
ODC
6
5
CNS+7
4
CNS+8
160k 162k 164k 166k 168k 170k 172k 174k 176k 178k 180k
ODC
CNS+9
100%
75%
50%
180k 182k 184k 186k 188k 190k 192k 194k 196k 198k 199891

The pufferfish sequence was achieved by direct sequencing of the Fugu cosmid clone ICRFc66D2193 (see above). The over 24 kb sequence included the *Nkx2-9* gene from position 2877 to position 1887 (in backwards orientation) and the *Pax9* gene from position 13180 to position 17904.

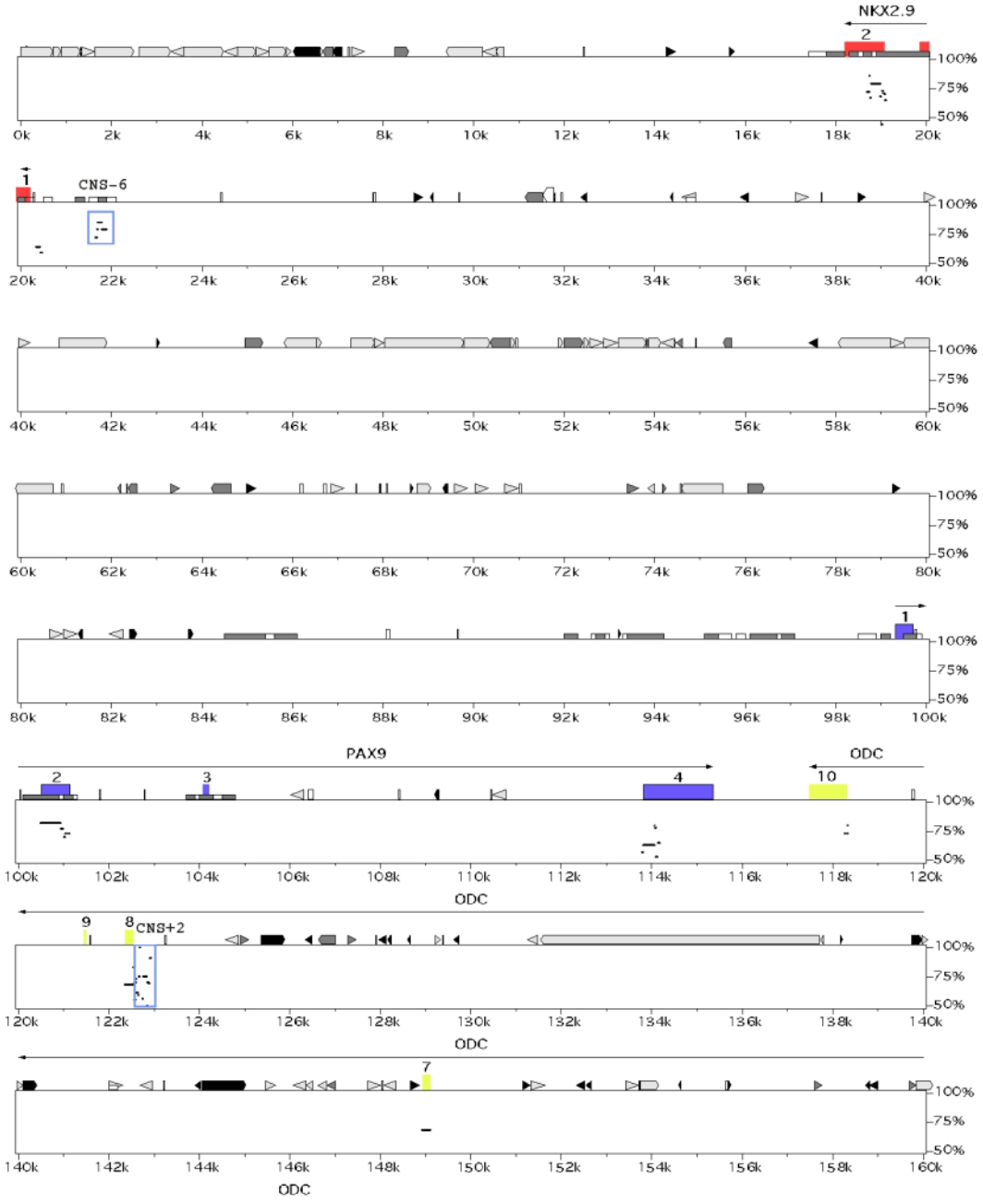


Fig. 19. PIP analysis of the human/*Fugu* alignment. The same 200 kb of human genomic sequence as in Figure 18 aligned against the 24 kb of *Fugu* cosmid sequence (see text). Only 160 kb of human sequence are shown, because the *Fugu* sequence did not reach farther to the exon 6 of the *Odc* gene. See legend of Figure 18 for explanation.

Alignment of the *Fugu* and the human sequences and analysis through the PIP algorithm produced the results shown in Figure 19. It was immediately obvious that the high species divergence (900 million years) corresponded to almost a complete loss of non-coding sequence conservation, despite the conserved locus syntheny. In the intergenic region between *Pax9* and *Nkx2-9* only one single homologous segment was detected, coinciding with the CNS-6 of the mouse-human identity plot (Fig. 18 and 19). This conserved element was less than 2 kb away from *Nkx2-9* 5'-end and included a CpG island. CpG islands are known to be important promoter structures for epigenetic mechanisms of gene regulation (Ioshikhes and Zhang 2000).

Downstream of *Pax9*, three distinct hits were individualized by the PIP analysis. Interestingly, these conserved elements fell into three exons of the *ODC* gene in the corresponding human sequence. Although no complete gene identification and characterization was accomplished, this was sufficient evidence for the presence of the same gene in the pufferfish sequence. This finding was very intriguing, because it extended the region of conserved syntheny among the three vertebrate species to an additional gene.

Apart from that, another conserved non-coding sequence was detected inside the *ODC* intron 7 in the close vicinity of exon 8. This element coincided with the mouse-human CNS+2 (Fig. 18 and 19)

The scarceness of sequence conservation in the intergenic region of the *Pax9* locus between human and pufferfish could be explained either by a loss of sequence homology, due to low selective evolutionary pressure, or by a diversification of the gene function in the mammal and teleost lineages.

4.5.2. Pax9 in situ hybridization on zebrafish embryos

No *Fugu* gene expression pattern data are available. However, a general insight of gene expression in the teleosts can be achieved by studies on zebrafish. The zebrafish *Pax9* gene had been already isolated and it was described to be expressed in the sclerotomal compartment of the somites of the developing embryos (Nornes et al. 1996).

A *Pax9* in-situ hybridization on zebrafish embryos was performed in order to look for further expression domains and to allow a more complete comparison with the mouse data. Different stages of zebrafish development were analyzed. At 20-somite stage (approx. 18

hours after fertilization-hpf), the main *Pax9* expression domain appeared to be in the ventral portion of the somites, corresponding to the sclerotome, in agreement with the published observations (Fig. 20a and Nornes et al. 1996). At this stage, a faint stripe of *Pax9* positive cells could be already noticed laterally on both sides at the level of the developing pharyngeal structures.

The staining in the pharynx region was more evident at 48 hpf of development, when it represented the prevalent *Pax9* expression domain and it expanded even more in the later stages. Conversely, the somite staining had disappeared, maybe due to inaccessibility to those structures in later developmental stages or to a real down-regulation of the gene. Two dots of *Pax9* expressing cells could be noticed in the ventromedial area of the face. The nature of these structures could not be investigated in details (Fig. 20b). Starting from 82 hpf a most anterior expression domain began to delineate. This appeared as a pairwise element, coinciding with the most ventral margin of the olfactory organs. The expression was maintained also in the following stages (Fig. 20d and 20e). Also the expression in the pharyngeal endoderm persisted and outlined the shape of the pharyngeal gills (Fig. 20 e). The comparison of the available *in situ* data from zebrafish and mouse embryos suggested that a few embryonic domains expressed *Pax9* in both species, including the somites, the pharyngeal endoderm and the nasal region. Other structures, like the developing fin buds (corresponding to the limb buds in the mouse) did not show *Pax9* expression. This suggested that an investigation of *Pax9* elements through comparative sequencing of a mammal and a fish genome could be only accomplished for the common positive domains. Moreover, it has to be said that the facial expression in the mouse is much more complex and diverse than in the fish. Several cranio-facial elements like nasal processes, palate processes, maxillary processes are not easy to associate to homologous features in the fish and thus, this would narrow the possibility to identify specific regulatory elements through a simple sequence comparison.

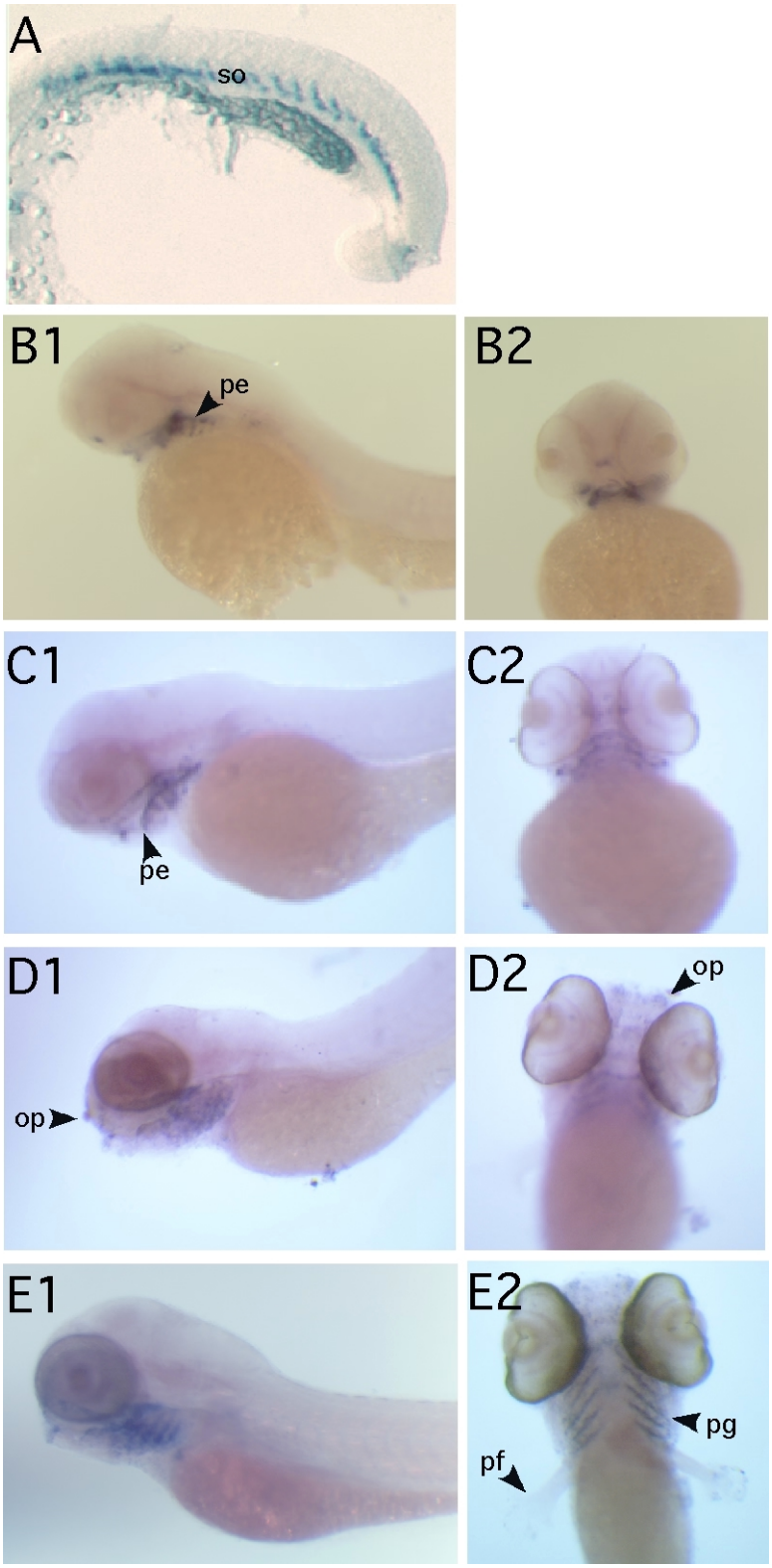


Fig. 20. Pax9 expression pattern in zebrafish. A, B1, C1, D1, E1 left lateral views, anterior is on the left. B2, C2, D2 and E2 ventral view of respectively B1, C1, D1 and, E1, anterior is on top. (A) 20 somite stage (18 hpf). Pax9 is expressed in the ventral part of the somites (so) corresponding to the sclerotome. The expression in the pharyngeal region at this stage is very faint. (B1-2) 48 hpf Pax9 expression is now mainly concentrated in the pharyngeal endoderm (pe) and no staining in the somite is detectable. (C1-2) 66 hpf. The pattern resembles that of 48 hpf with a clear expansion of the expression in the anterior and pharyngeal endoderm. (D1-2) 82 hpf. Most anteriorly *Pax9* starts to be detected in the ventral margin of the olfactory organs (op) (E1-2) 94 hpf. Expression in the olfactory organs and in the five pairs of pharyngeal gills (pg). Notice the lack of expression in the pectoral fin buds (pf).

4.5.3. Cell culture assay with CNSs

An initial functional screening of the conserved elements shared by the mouse and human sequences was attempted in the cell culture system. The availability of two *Pax9* expressing cell lines would offer an easy system where the identified CNSs could be tested for cell-specific enhancer activity. Obviously, this approach would only lead to the identification of enhancer elements specifically acting in these cell lines.

Genomic fragments of 1 to 3 kb in length, which included the closest CNSs to the Pax9 gene, were subcloned into the Promoter-B construct that had shown to be able to perform a basal promoter activity (Fig. 17). The series of constructs was again tested for ability to drive expression of the luciferase reporter gene in the AT478 and MLB13myc cell lines. Unfortunately, two independent experiments, each with a doubled sampling for each CNS-construct, showed that neither specific nor reproducible data could be obtained with this approach. None of the constructs seemed to have an enhancer effect on the promoter sequence of more than 2-2.5 fold in either cell line, making the detectable difference not enough for further investigations. Moreover, these variations of luciferase activity driven by some CNS-constructs were not reproducible in the two separate experiments (data not shown). The failure of this assay could be due to the fact that perhaps none of the tested CNSs represented a real enhancer sequence for the two cell lines or the experimental system was inadequate for the detection of such a function.

4.5.4. Transient transgenesis with CNSs

An alternative way to test the potential regulatory activity of the selected CNSs was an in-vivo assay through transient transgenesis. The general design of the experiment was to place the test genomic fragments upstream of an *hsp68*-promoter-*LacZ*-poly(A) cassette. This cassette contains the bacterial *lacZ* gene as a reporter gene under the control of a minimal promoter from the mouse heat-inducible gene *hsp68*. The *hsp68* minimal promoter is suitable for this kind of experimental approach because it has no detectable basal activity in transgenic mouse embryos (Kothary et al. 1989) but it can be activated in distinct patterns by heterologous enhancer elements (Logan et al. 1993; Sasaki and Hogan 1996).

Moreover, the use of a non-specific promoter would allow to identify enhancer sequences not only for *Pax9* but also for its neighboring genes (see later).

The transient transgenesis method was based on the production of transgenic mouse embryos using CNS-reporter cassette constructs. Founder transgenic embryos would be directly analyzed in mid-gestation by *lacZ* staining. The reproducible staining of the same specific structures in different embryos would provide evidence for the related enhancer activity of the test fragment. The advantage of this experimental procedure is that an in-vivo model reduces the possibility of artificial non-physiological results. In addition to that, while in a cell system only specific enhancers for that particular cell line can be identified, the use of transgenic animals allows to detect the reporter gene expression in a variety of tissues and cell types.

The CNS-6 was taken as a first test fragment for the enhancer assay. Although it represented the most distant element to the *Pax9* gene among the ones identified, some features made it the best candidate for a regulatory function and thus for a positive testing of the transgenic approach. The CNS-6 lay only about 1.5 kb from the *Nkx2-9* 5'-end, making it very likely to be part of the proximal promoter elements of this gene. As a matter of the fact, the CNS-6 included a CpG island associated to this promoter and that represented a further clue for a possible regulatory role. Furthermore, this was one of the only two CNS elements, which had been shown to be conserved also in the *Fugu* genome (Fig. 19). The human-mouse CNS-6 consisted of a 422 bp sequence with an overall 85% identity between the two species. Within this sequence a 244 bp segment had 80% conservation with the pufferfish DNA (Fig. 21).

Five independent 10.5 dpc mouse embryos were found to carry the CNS-6 transgene by genomic DNA PCR analysis. All of them showed β-galactosidase expression upon *lacZ* staining, even though at different grades and with a variety of expression patterns including irreproducible ectopic staining in diverse structures. Nevertheless, all of the samples shared a common positive domain in the ventral neural tube. In only one sample the neural tube expression was very spotted and circumscribed at the hindbrain level. In general it was detectable along the whole body axis, from the most anterior part of the telencephalon and down to the extreme posterior neural tube in the tip of the tail. Only a short interruption could be frequently observed at the mid-hinbrain boundary level. The ventral staining was

in some samples splitting in two parallel stripes, lateral to the floor plate. In some cases the transgene expression was extending laterally, especially in the brain, and in most cases it marked a significant number of cells in the roof plate (dorsal neural tube) (Fig. 22)

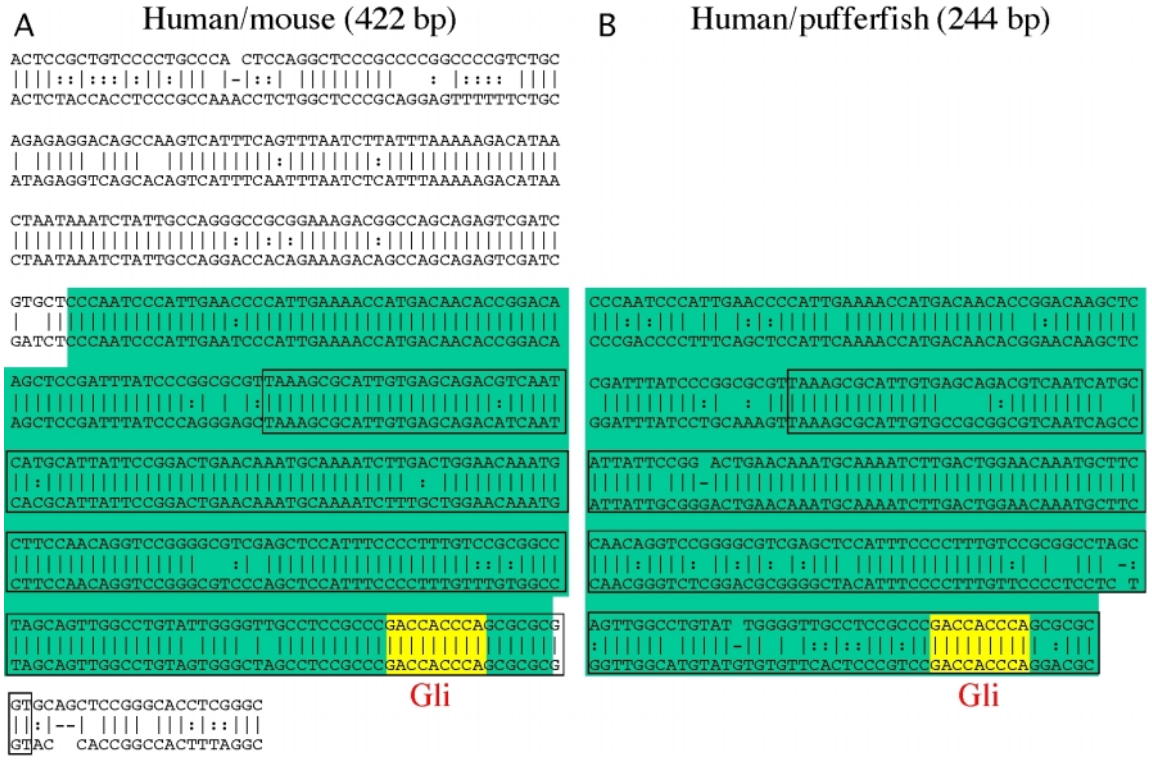


Fig. 21. CNS-6 sequence and interspecies alignment. Human/mouse (A) and human/*Fugu* (B) sequence alignment of CNS-6 element, extrapolated from the PIP analysis shown in Figures 18 and 19. In the alignment a vertical bar indicates a perfect match, while two dots indicate a transition. The green area in A corresponds to the homologous part to Fugu shown in B. Gli binding sites are highlighted in yellow (see text). The boxed sequence corresponds to the human CpG island found in the databank (Z56225).

Nkx2-9 had been shown to be expressed in the ventral half of the neural tube of developing mouse embryos (Pabst et al. 1998). The similar behavior of the transgene accredited the CNS-6 element as the cis-acting DNA sequence that regulates *Nkx2-9* expression in the neural tube. A whole mount *in-situ* hybridization with an *Nkx2-9* RNA probe was performed on 10.5 dpc mouse embryos for direct comparison with the transgenic data. *Nkx2-9* expression at this developmental stage is only restricted to the brain and to the caudal part of the neural tube, while most of the trunk is free of *Nkx2-9* mRNA (Pabst et al. 1998 and Fig. 23). *Nkx2-9* expression in the whole neural tube is only transient and could be observed in younger embryos (9.5 dpc) (Fig. 23), as described by Pabst et al (Pabst et al. 1998).

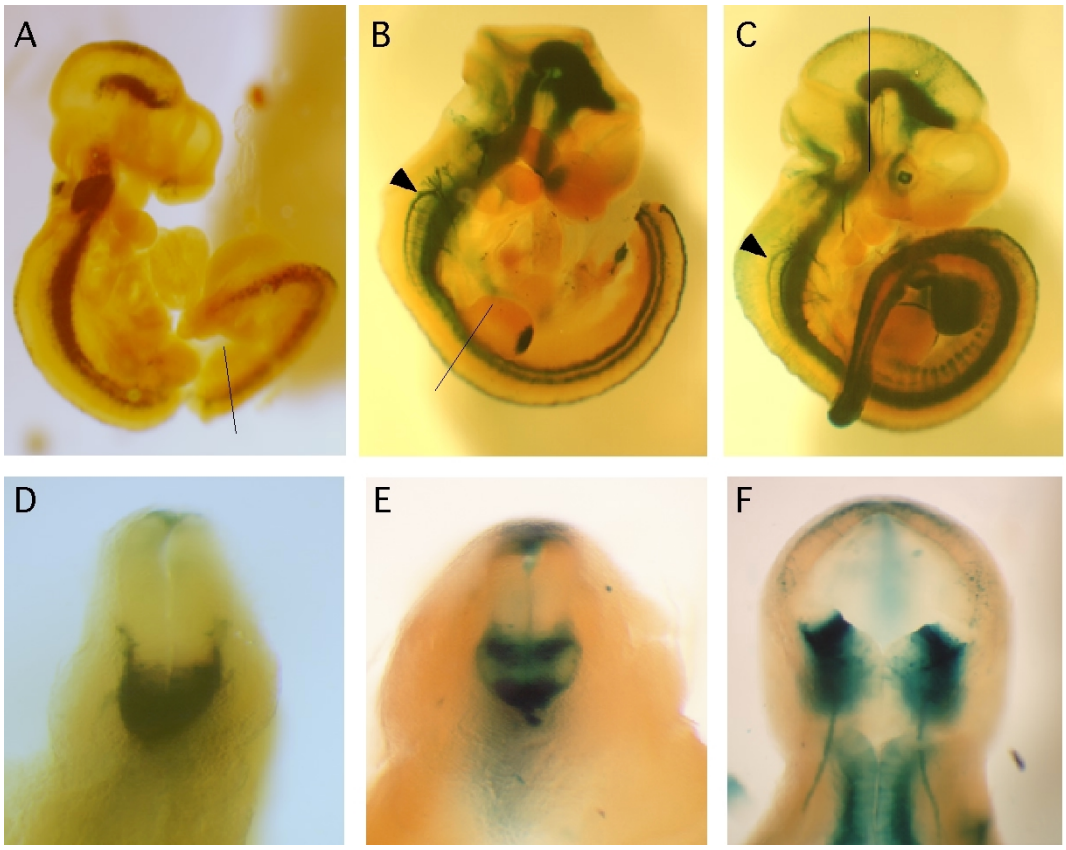


Fig. 22. X-gal staining of CNS-6 transgenic embryos. (A-C) Whole mount overview of three of the five independent transient transgenic embryos at developmental stage E10.5-E11. *lacZ* expression is clearly visible along the whole neural tube. Arrowheads in B and C indicate emerging neurons. On the bottom, D, E, and F cross-sections of A, B, and C respectively. The section level is shown in the top panel as a transversal line across the embryo body axis. X-gal staining is in the ventral half of the neural tube. At midbrain level (F) the staining splits in two parallel ventral stripes. Notice the ectopic staining also in the dorsal neural tube, especially in E.

The *lacZ* expression still observed in 10.5 dpc transgenic embryos could be due to residual β-galactosidase activity, which is likely to be longer detectable than the *Nkx2-9* mRNA. However, it cannot be ruled out that the CNS-6 contains only spatial and no temporal information for *Nkx2-9* expression. On the other hand, no *Nkx2-9* expression was described in the floor plate as in the transgenic model, suggesting that the enhancer sequence, deprived of its genomic context, can mislead the positional information in the regulation of the gene.

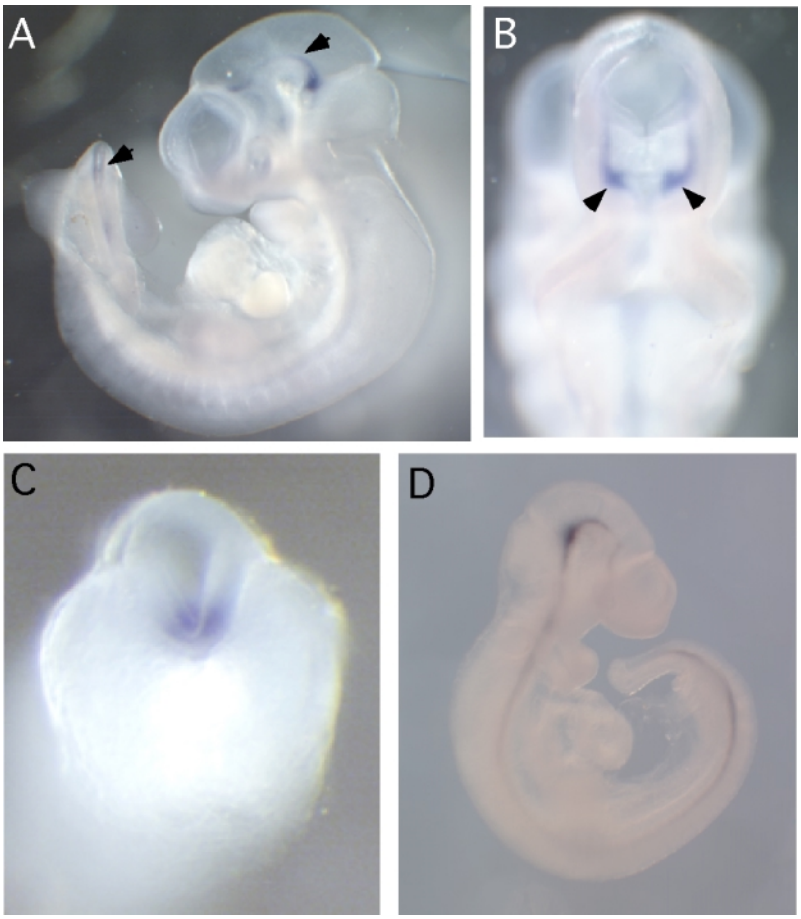


Fig. 23. *Nkx2*-9 whole mount *in situ* hybridization on mouse embryos. (A) 10.5 dpc embryo. Nkx2-9 RNA is detectable in midbrain and caudal neural tube (arrowheads). (B) Dorsal view of the midbrain region of the same sample. Nkx2-9 pattern forms two parallel stripes (arrowheads). (C) Cross-section of same embryo in the caudal region; Nkx2-9 is expressed in the ventral part of the neural tube. (D) 9.5 dpc embryo. Nkx2-9 is distributed along the whole neural tube axis.

In the same publication Pabst et al. discuss a possible involvement of Sonic Hedgehog (Shh) signaling from the floor plate in the regulation of *Nkx2-9* (Pabst et al. 1998). The Gli transcription factors can be induced upon Shh signaling and represent the final effectors of its intracellular cascade at the DNA level (Sasaki et al. 1997). A Gli binding element was identified in the floor plate enhancer of the Hepatocyte Nuclear Factor-3β gene (*Hnf-3β/Foxa2*), whose expression was already known to be Shh-dependent (Sasaki et al. 1997). Interestingly, a fully conserved consensus for Gli binding site could also be recognized within the CNS-6 sequence (Fig. 21).

The second fragment used in this transient transgenesis assay was the CNS+2, which was the other conserved non-coding sequence found in the comparative sequencing between mammals and *Fugu*. This element was located very close to the exon 8 of the *Odc* gene in the three species inside the preceding intron (Fig. 19 and 24). In the human and *Fugu* sequences, the homology through the *Odc* exon continued into the CNS+2 element without

a major interruption. In the mouse, 60 bp of intervening sequence separated the exon from the intronic conserved element (Fig. 22). Excluding the exon sequence, the human/mouse homologous region was 594 bp long with an overall identity of 95%, 300 bp of which were homologous to the pufferfish sequence with 64% identity (Fig. 24).

However, about 2.2 kb of the mouse sequence including the Odc exon 8 were used to make the transgenic construct with the *lacZ* reporter in case the surrounding sequence context was important for the putative regulatory function. Three transgenic mouse embryos (11.5 dpc old) were produced with this construct. Apart a series of ectopic X-Gal staining in several domains which did not match with the *Pax9* expression pattern, probably due to a positional effect from the region of transgene integration (data not shown), a very specific expression was observed in all of the embryos in the ventro-medial region of the medial nasal processes. This domain nicely overlapped with the two stripes of *Pax9* expression in the same structures, which could be reproduced by *in situ* hybridization on embryos from a similar developmental stage (Fig. 25), as previously described in Neubüser et al. (Neubüser et al. 1995). However, the transgene expression did not extend to the more internal medial nasal processes where *Pax9* could be also detected.

Interestingly, the *Pax9* in situ data on zebrafish showed a similar expression in the ventral margin of the olfactory organs, just above the oral region (Fig. 20d and 20e). The sequence conservation of the CNS+2 between mouse and pufferfish strongly suggests that this could as well represent a cis-regulatory element for the *Pax9* expression in the olfactory organs in the fish. This hypothesis will be tested with a similar transgenic approach in zebrafish embryos.

Fig. 24 (next page). CNS+2 sequence and interspecies alignment. Human/mouse (A) and human/*Fugu* (B) sequence alignment of CNS+2 element, extrapolated from the PIP analysis shown in Figures 18 and 19. In the alignment a vertical bar indicates a perfect match, while two dots indicate a transition.

The green area in A corresponds to the homologous green area in *Fugu* shown in B. CNS+2 starts from this green area and its size is given above in brackets. Downstream of the common (green) homologous region no further homology with the *Fugu* sequence was detected, as in the mouse. Conversely upstream of the CNS+2, homology continues into the exon 8 of the *Odc* gene, whose start is highlighted in yellow. Notice that the CNS+2 sequence is shown in opposite orientation in respect to the *Odc* gene (arrows).

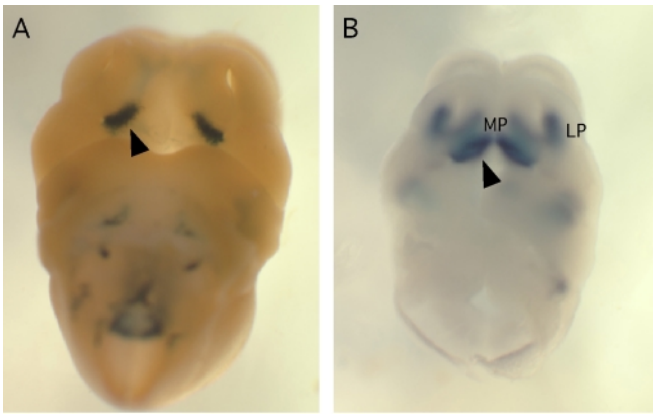


Fig. 25. X-gal staining of a CNS+2 transgenic embryo. (A) Ventral view of the head of a transient transgenic embryo at developmental stage E11.5. X-Gal staining is visible in the oral edge of the medial nasal processes (arrowhead). (B) Ventral view of the head of a E11.0 embryo after *in situ* hybridization for *Pax9*. The arrowhead indicates a similar region as in (A). *Pax9* is expressed also in the dorsal part of the medial nasal processes (MP) and in the lateral nasal processes (LP). Notice that the transgenic embryo (A) is at an older developmental stage (compare the size) and that the nasal region tends to enlarge, bringing the two stained domains apart.

4.6. BAC transgenesis

The second approach in the pursuit of regulatory elements for *Pax9* was based on the creation of a transgenic mouse model. The principle of this experimental design was to generate a BAC transgenic mouse, using one of the isolated BAC clones, in the intent to reproduce the complete *Pax9* expression pattern. The identification of a large genomic sequence that entirely exerts this regulatory function would be the first step to successively narrow it down to its single components. For this purpose, a series of deletions of the original BAC clone would be in a second time similarly tested by the establishment of transgenic mice.

The BAC clone 17 was regarded as the most appropriate candidate to start this analysis. The Figure 26 shows again the BAC contig around the *Pax9* locus, including the identified neighboring genes. The BAC17 was one of the few clones that covered the genomic sequence encompassing the *Pax9* gene and reached up to the two flanking genes, *Nkx2-9* on the 5' side and the *Odc* gene on the 3' side. Thus, considering the two genes as boundaries of the *Pax9* region, the BAC17 was believed to be the most likely to hold all the necessary regulatory elements. In addition to that, the presence of the NotI site favored a more precise localization of this clone in respect to the *Pax9* gene.

The strategy for the transgenic approach was to insert a reporter gene in the chosen BAC clone, so that it would be located in the *Pax9* locus and would be expressed in a similar temporal and spatial manner as the *Pax9* gene itself. The expression distribution of the reporter gene would constitute a landmark to follow up the transcription regulation specifically driven by the elements in the BAC.

4.6.1. BAC modification

It was decided to introduce the reporter cassette in a way that it would not disrupt the *Pax9* gene. The presence of a completely functional gene in the BAC sequence would allow rescue experiments, in which the transgene is crossed into a *Pax9* knockout background and let complement for the lacking function. That would be a further proof for the complete functionality of the BAC regulatory sequences.

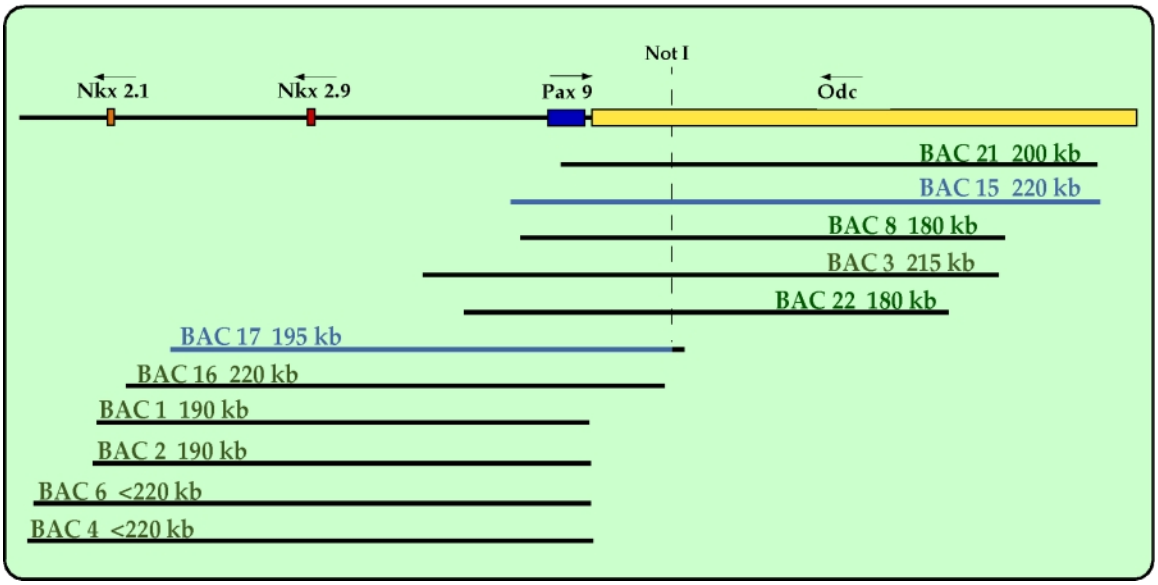


Fig. 26. BAC contig around the *Pax9* locus and the neighboring genes. Gene orientations are indicated by the arrows. BAC clones 15 and 17 (in blue) were modified with the introduction of a *lacZ* reporter gene. Only BAC 17 was used for transgenesis.

The Figure 27 shows a representation of the insertion of the reporter gene in the *Pax9* locus. The reporter cassette was introduced in the *Pax9* 3'-UTR, preceded by an internal ribosomal entry sequence (IRES) that allows internal initiation of translation of a messenger RNA. The transgene would produce in this way a bicistronic transcript that codes both for *Pax9* and for the reporter protein. The reporter gene was additionally fused to the neomycin-resistance gene. This would add a third function to the transgene. The expression of a selectable marker would be useful in future applications for the isolation and in vitro culture of embryonic cells from tissues that specifically express *Pax9*.

Two different reporter cassettes were employed, an IRES-βGeo cassette that carried the *E. coli lacZ* gene, coding for the β-galactosidase, fused to the Neo^R marker, and an IRES-GFPneo cassette, which instead encoded the green fluorescent protein (GFP) as a reporter.

In order to insert these cassettes into the target sequence, a BAC modification approach was applied. This was based on homologous recombination in *E. coli* cells between the target sequence and the insertion cassette. Two different methods were assayed. Both methods and the respective experimental procedures are described in the "Materials and Methods" section.

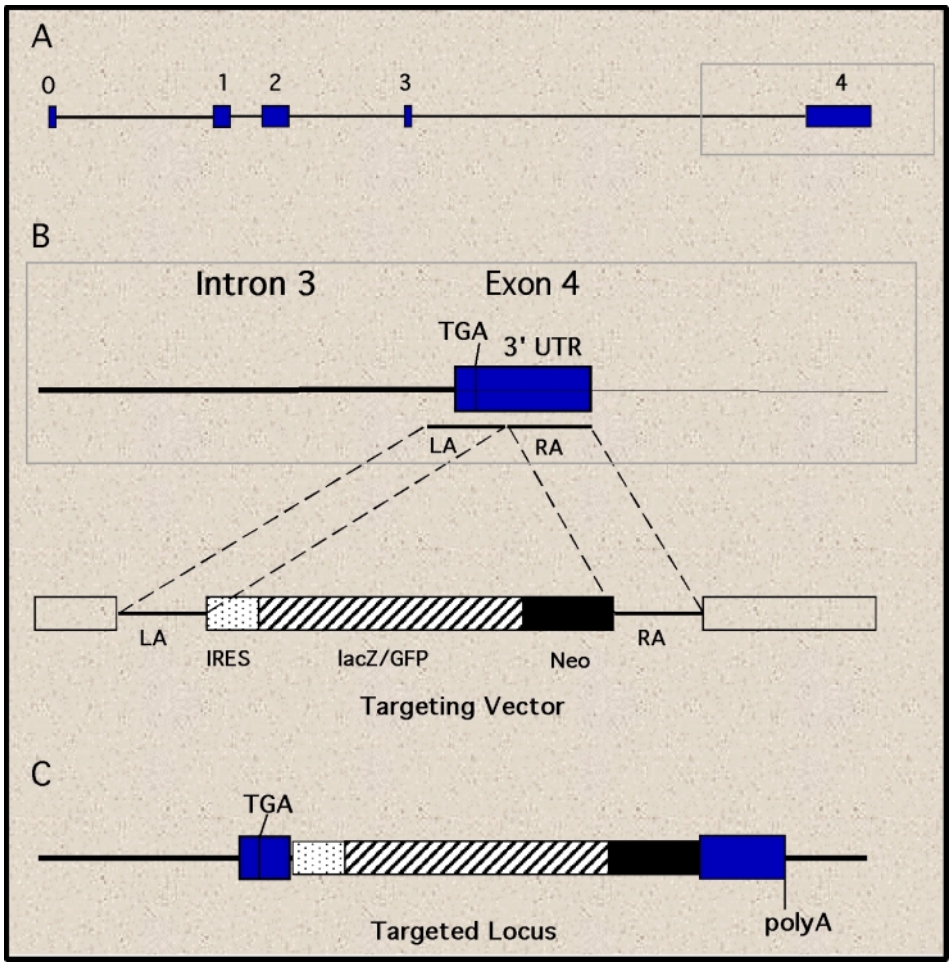


Fig. 27. Strategy for reporter insertion into *Pax9* locus. (A) *Pax9* gene structure. The gray box encloses exon 4, enlarged in (B). (B) Exon 4, including the stop codon (TGA) and the 3'-UTR, is target of homologous recombination. Left and right recombination arms of targeting vector (LA and RA respectively) flank an insertion cassette containing a reporter gene (*lacZ* or GFP) fused on the 3' side to the G418 resistance gene (Neo) and preceded by an IRES element (see text).
(C) Final appearance of targeted locus: the reporter cassette is inserted in the 3'-UTR of the gene.

Initially, the RecA based system was used for the insertion of the IRES-βGeo cassette. However, even though the method was previously described for insertion of a *lacZ* reporter in BAC clones (Yang et al. 1997), it resulted instead in a high recombination frequency of the targeting cassette into the *E. coli* genome, presumably in the endogenous *lacZ* locus, interfering with the desired recombination. Conversely, this approach was successful for the insertion of the IRES-GFPneo cassette, which carried no homologous sequence to the *E. coli* genome. The results of this modification are not shown here, since the construct was not used for further applications.

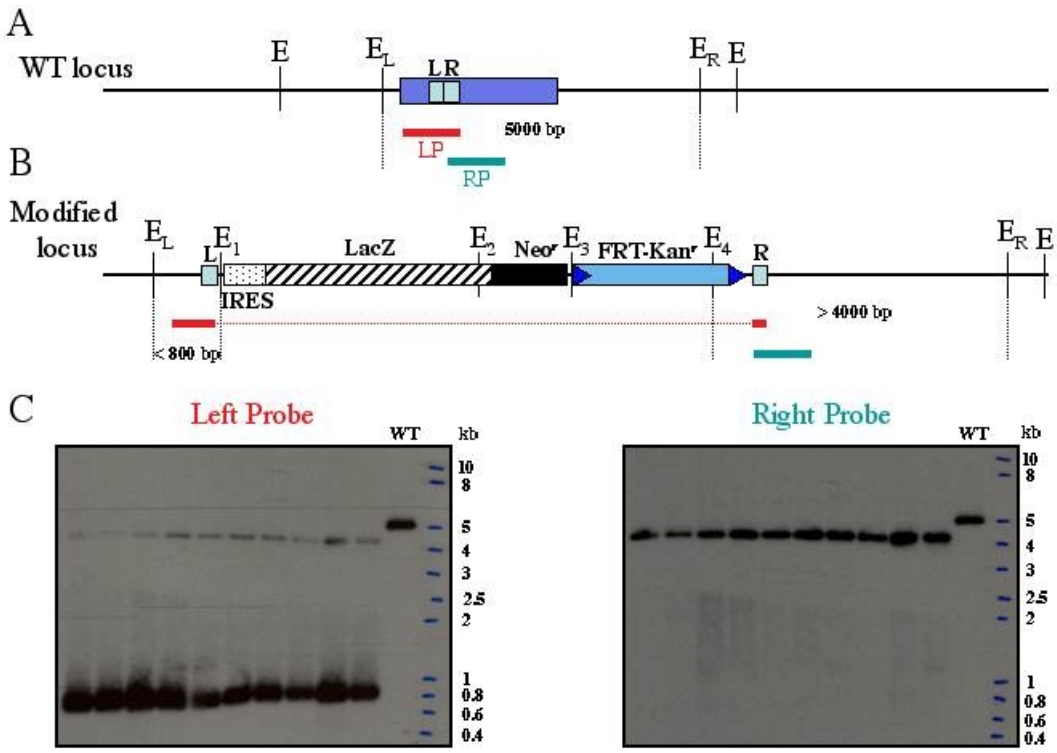


Fig. 28a. ET-cloning: Southern blot analysis of modified BAC clones. (A) wild type (WT) target locus with EcoRI sites (E). Blue rectangle is exon 4, L and R boxes are left and right recombination sequences used for targeting vector, LP and RP are left and right probes used for the analysis. A 5 kb band (E_L-E_R) is detected with both probes after EcoRI digestion (WT in Southern blot (C)). (B) EcoRI restriction map after insertion of reporter cassette. LP detects now an 800 bp band (E_L-E_1) and weakly a 4 kb band (E_4-E_R), while RP only the 4 kb band. (C) 10 distinct kanamycin resistant clones show the correct pattern with the two probes.

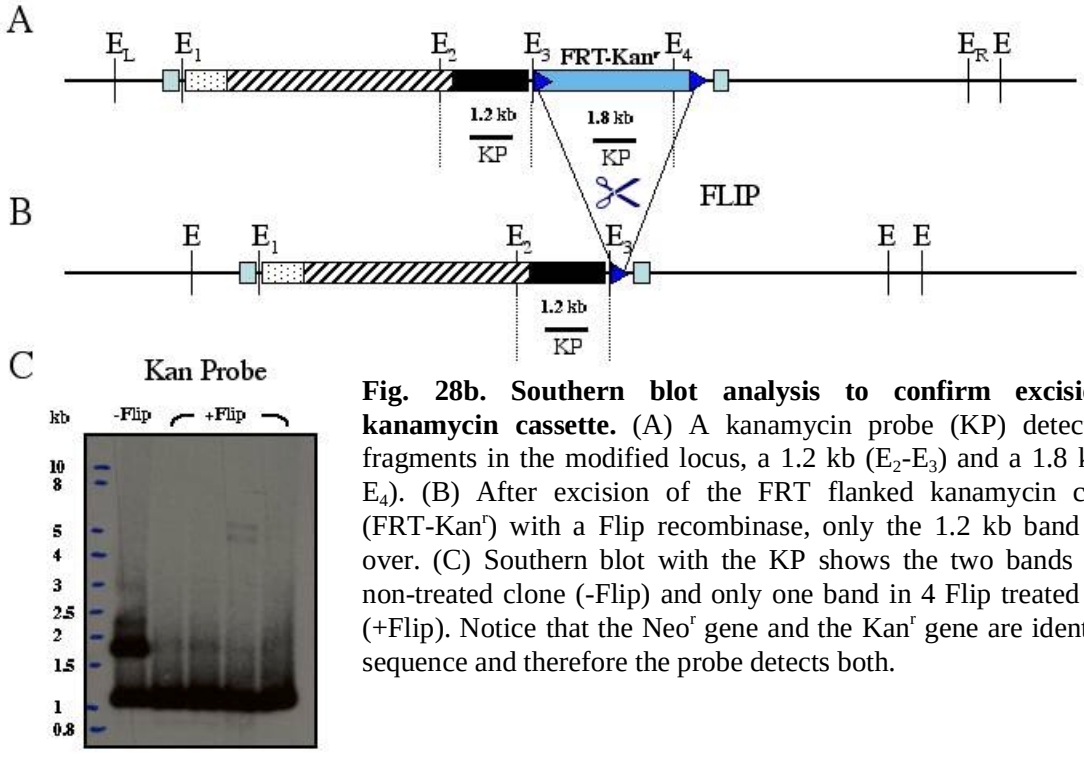


Fig. 28b. Southern blot analysis to confirm excision of kanamycin cassette. (A) A kanamycin probe (KP) detects two fragments in the modified locus, a 1.2 kb (E_2-E_3) and a 1.8 kb (E_3-E_4). (B) After excision of the FRT flanked kanamycin cassette (FRT-Kan^r) with a Flip recombinase, only the 1.2 kb band is left over. (C) Southern blot with the KP shows the two bands in one non-treated clone (-Flip) and only one band in 4 Flip treated clones (+Flip). Notice that the Neo^r gene and the Kan^r gene are identical in sequence and therefore the probe detects both.

The modification with the IRES-βGeo cassette could be accomplished using the ET-cloning method, in which the recombination into the target sequence occurs only through the lateral arms of the cassette and the presence of internal homologous sequences does not cause any undesired result (Fig. 28).

4.6.2. Generation and analysis of BAC-transgenic mice

The IRES-βGeo-BAC17 (mBAC17-*lacZ*) was used for the generation of the first transgenic mouse line. The construct was linearized and excised from the vector sequence with NotI, which released a 195 kb fragment containing almost entirely the BAC insert sequence. The pronucleus injection of the BAC construct in fertilized oocytes was carried out by a collaborating partner group in the University of Kumamoto, Japan, under the supervision of Dr. Kunyia Abe. Three transgenic mice were produced after one round of injection. The screening was conducted by PCR using specific primer pairs from the transgene sequence (PCR table 1, nos. 14 and 24). One of the transgenic mice carried a truncated form of the transgene with only the 3'-end and lacking the modified *Pax9* locus. The other two transgenic mice, one male (mBAC17-04) and one female (mBAC17-01), seemed to contain both ends of the BAC transgene.

Both animals were bred in order to found the respective lines and the analysis of the reporter gene expression was performed in the offspring at different embryonic stages.

The mBAC17-04 line could be directly analyzed by mating the founder mouse with wild-type females that were sacrificed during pregnancy for embryo preparation. On the contrary, for the line mBAC17-01, the impossibility to sacrifice the founder female made it necessary to wait for the next generation before starting the analysis.

The expression of the reporter gene in the transgenic embryos could be observed by X-Gal staining and compared to the endogenous *Pax9* expression using the $Pax9^{lacZ}$ line as a control. These mice carry the *lacZ* gene replacing the *Pax9* coding sequence and were generated for the knockout experiment (Peters et al. 1998b). Heterozygote $Pax9^{lacZ}$ mice are fully viable and X-Gal staining on developing embryos exactly reproduces the *Pax9* expression pattern.

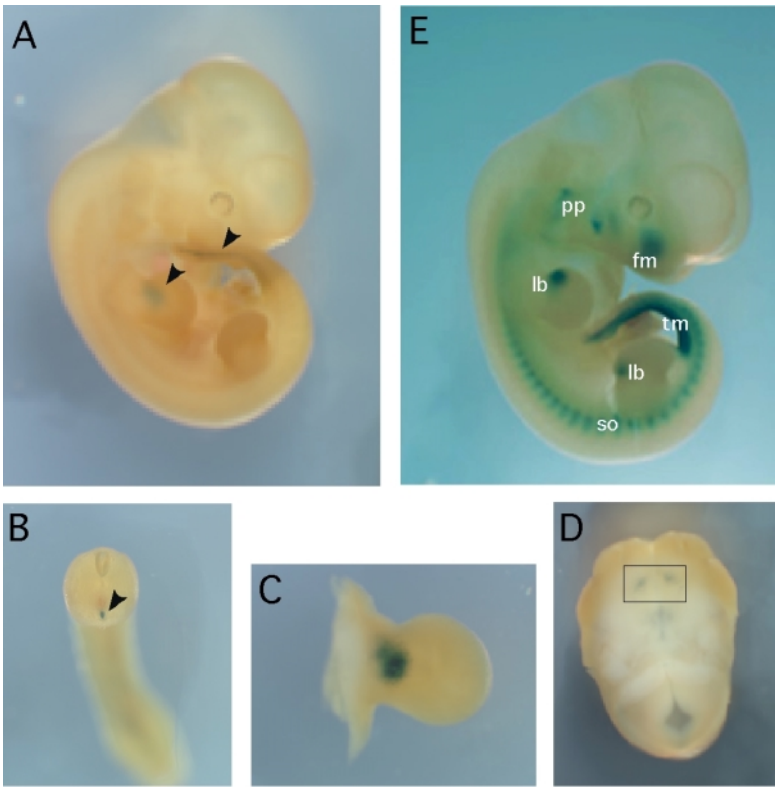


Fig. 29. X-gal staining of mBAC17-04 transgenic embryo at E11.5. (A) Whole mount overview. Staining is only visible in the tail and in the ventral side of limb buds (arrowheads). (B) Cross-section of the tail. Staining in the tail endoderm (arrowhead). (C) Ventral side of forelimb bud. (D) Ventral view of maxillary region after cutting at the level of the mouth. Stained medial nasal processes are boxed. (E) Control X-gal staining of a Pax9^{lacZ} E11.5 embryo. Staining in the somites (so), tail region (tm), limb buds (lb), pharyngeal pouches (pp), and facial mesenchyme (fm)

It was immediately obvious that the *lacZ* expression in the transgenic embryos of the mBAC17-04 line did not as well reproduce the *Pax9* expression pattern. Analyses at different developmental stages showed staining of only few of the *Pax9* domains. At 11.5 dpc, the transgene was only visibly expressed in few domains, in the terminal end of the tail gut, very faintly in the oral edge of the medial nasal processes and in the limb buds limited to the anterior ventral area of the hand and foot pad, whereas the *Pax9* gene is also expressed in the corresponding dorsal region. The expression in the somites, in the pharyngeal pouch endoderm and in the facial mesenchyme was totally absent and it could not be detected even after extensive staining of several days (Fig. 29).

Only starting from 12.5 dpc, a faint staining in the facial region could be observed. This became more evident at later stages (13.5-14.5 dpc), when it was necessary however to clear the embryos after X-Gal staining in order to look at the deeper structures (Fig. 30). Only few of the numerous facial elements that normally express *Pax9* could be stained.

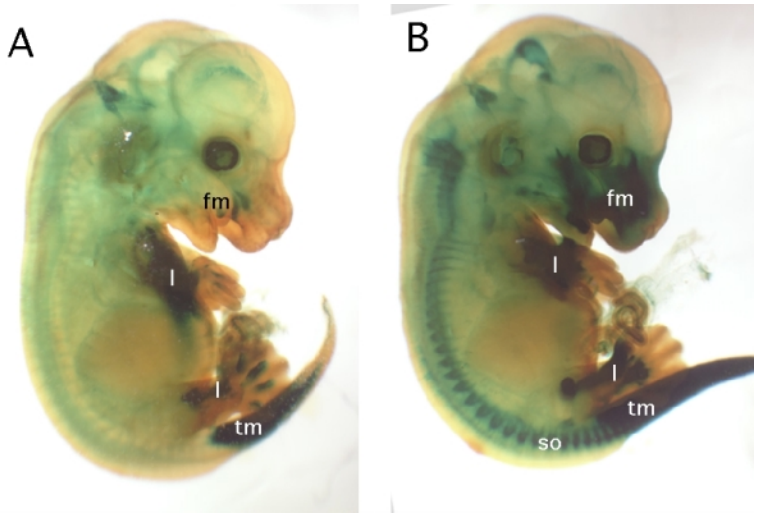


Fig. 30. X-gal staining of mBAC17-04 transgenic embryo at E13.5. (A) cleared e13.5 embryo. expression domains are part of the facial mesenchyme (fm), limbs (l), and tail mesenchyme (tm). (B) control x-gal staining of a pax9^{lacz} e13.5 embryo after clearing. notice the additional staining in the somites (so) and the widespread staining in the facial mesenchyme (fm).

As previously described (Neubüser et al. 1995; Peters et al. 1998b) and as shown in Figure 31, the *Pax9* transcript can be normally detected in the mesenchyme of the medial and lateral nasal processes. Starting from 11.5 dpc, *Pax9* is already strongly expressed in the mesenchyme of the nasal primordia and extends later on its positive domain all along the soft mesenchymal tissue between the nasal epithelium and the medial and lateral walls. The transgene staining appeared conversely only in restricted regions. In the nasal mesenchyme, as already mentioned, two faint stripes were seen in the area of the medial processes (Fig. 30). This staining corresponded to the expression reproduced with CNS+2 element in the transient transgenic experiment (Fig. 25). This was expected since the element was contained in the BAC17 sequence.

Later on at around 13.5 dpc, two small staining spots could be observed, one in a latero-ventral domain and the other in a medio-ventral domain, the latter corresponding to the area of the vomeronasal organ or Jacobson's organ (Fig. 31c). This structure grows away from both sides of the ventral nasal epithelium and develops into an independent organ responsible in many animal species for the reception of pheromones. Similarly to the endogenous gene, a strong expression of the transgene could be also noticed anteriorly in the mesenchymal anlagen of the primary palate, which forms medially in a single structure at the base of the nasal septum.

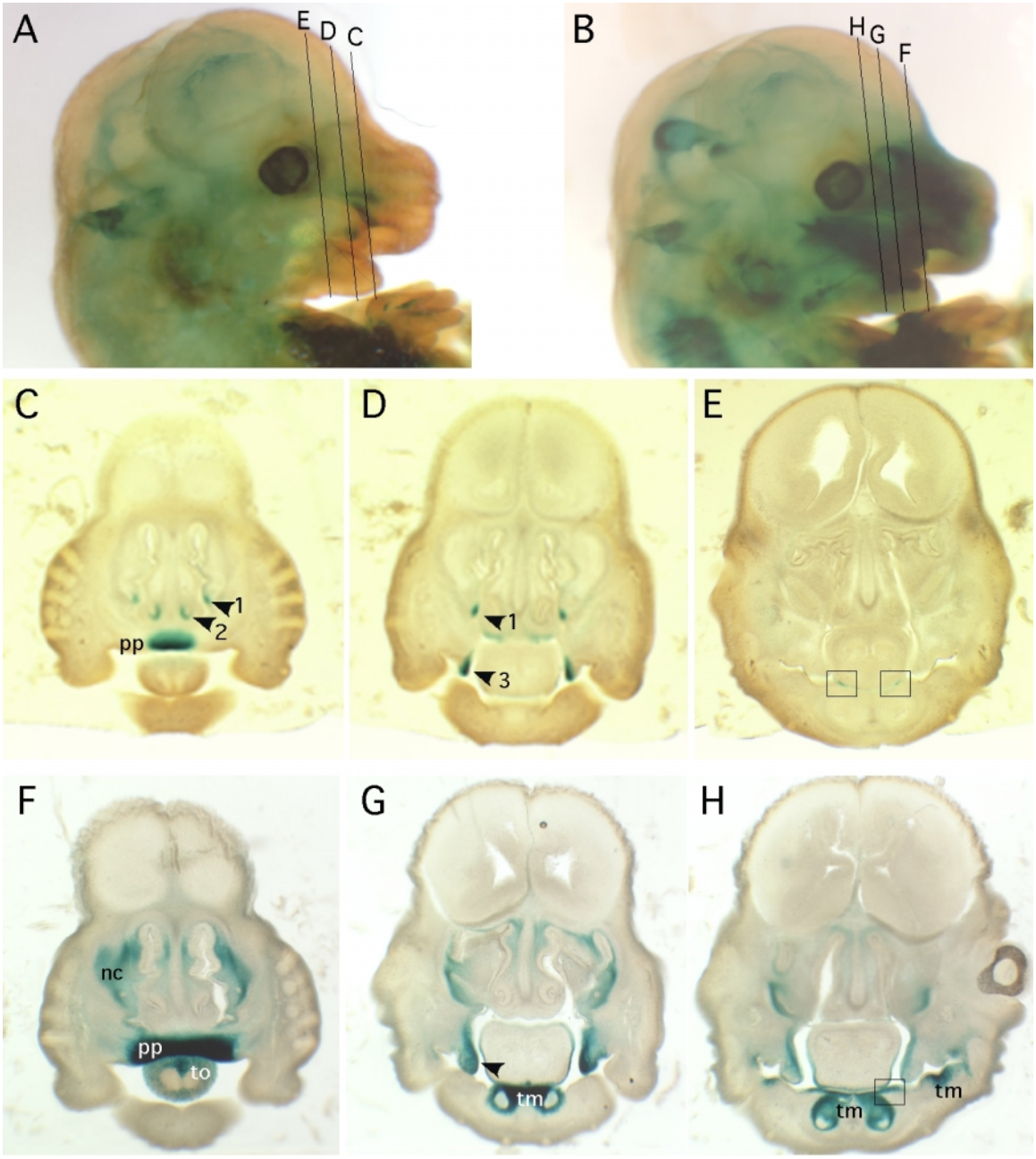


Fig. 31. Transgene expression in the cranio-facial structures. (A) Right lateral view of mBAC17 transgenic embryo head at E13.5. Section levels C, D, and E of respective pictures below are indicated. (B) Right lateral view of Pax9^{lacZ} embryo head at E13.5. Section levels F, G, and H of respective pictures below are indicated. (C-E) Coronal sections of sibling embryo of A. BAC-transgene is expressed in primary palate anlagen (pp) and in restricted spots of the nasal mesenchyme (arrowheads 1and 2 in C and D). Arrowhead 2 points in particular to the vomeronasal organ. Projecting tips of secondary palate processes are as well positive (arrowhead 3 in D). Staining in the sublingual oral epithelium is boxed in E. (F-H) Coronal sections of sibling embryo of B. *Pax9* is normally expressed extensively in the nasal capsule (nc), primary palate (pp), tongue epithelium (to), tooth mesenchyme (tm), secondary palate processes (arrowhead in G) and sublingual epithelium (boxed in H). Note the remarkably broader expression in the primary and secondary palate anlagen and in the nasal area.

Also the anlagen of the secondary palate turned out to be positive for the transgene expression. These could be noticed posteriorly as two separate processes extending from the maxillary mesenchyme and sliding down laterally to the tongue, before they bent upwards and met on the midline dorsally to the tongue to fuse and seclude the nasal cavity from the oral cavity. X-Gal staining could be seen only at the tip of the extending processes, while at a later phase, when the processes meet and fuse, no more staining was detected (Fig. 31d). On the contrary, *Pax9* is normally expressed not only in the palatal shelves, but also in the mesenchyme of the mandibular arch facing the palatal shelves and this expression domain can be observed very early in development (from 11.5 dpc on) before the first evaginations of the palate processes form. Nevertheless, the complete *Pax9* expression in the diffuse mesenchyme of the maxillary and mandibular processes and in the related structures, including the mesenchyme surrounding the thickening epithelium of the tooth buds, could not be reproduced by the transgene. Some faint staining was also observed in the sublingual epithelium, maybe corresponding to the terminal opening ends of the salivary ducts (Fig. 31e). This sublingual expression domain was even stronger at later stages (E 15, not shown).

The expression of the transgene in the limbs was quite comparable to the endogenous *Pax9*, but only in the ventral area. As already mentioned above, the earliest *Pax9* transcripts in the limb buds can be detected from 11.5 dpc both in the dorsal and in the ventral side of the proximal region of the developing hand and foot pads. The same situation can be observed one day later and starting from 13.5 dpc, when the shaping of the digits clearly begins, the dorsal expression slowly fades off, remaining only at a low extent in the anterior region at the level of the thumb in continuation with the ventral expression. Later on at around 14.5 dpc, a new dorsal domain appears in the digits, corresponding to the forming joints between the phalanges. The transgene expression was undetectable in every dorsal structure, while all the ventral domains showed quite a good consistent expression pattern in every examined developmental stage. Both the distal and the proximal expressions of the gene in the plant of the middle hand and foot, which was known to correspond to the mesenchyme between the metacarpals and metatarsals, respectively (Peters et al. 1998b), and in the radial (forelimbs) and tibial (hindlimbs) regions, corresponding to the forming tendons, were faithfully reproduced (Fig. 32). However, the expression in the digits presented some

differences compared to the Pax9^{lacZ} staining. As the digits elongate, *Pax9* is clearly detectable in the joint mesenchyme between the phalanges. These structures were only very faintly stained in the transgenic samples, if compared to the normal staining intensity in the rest of the positive tissues. Moreover, a certain transgene expression was noticed in the interdigital mesenchyme, in a region characterized by the high apoptotic activity of the morphogenetic processes (Montero et al. 2001). Endogenous *Pax9* expression in these interdigital regions could be sometimes observed at a lower level as rather more anteriorly shifted horizontal stripes (Fig. 32).

The expression in the tail region could be discriminated in two separate domains of different origin. At 11.5 dpc an earlier staining in the epithelium of the very distal end of the hindgut was one of the only two characteristic features of the transgene expression (Fig. 30b). This structure could be still observed one day later, but starting from 13.5 dpc no staining could be visible anymore, perhaps due to the normal regression of this terminal part of the intestinal tube. Oppositely, this early tail staining was replaced by a later staining in a different location. From 12.5 dpc on, two stripes of mesenchymal tissue extending laterally along the tail appeared strongly positive for the expression of the transgene and represented a clear landmark for the identification of the transgenic embryos in all the later stages analyzed (Fig. 33). This structure, which was as well observed and previously described for *Pax9*, constitutes the primordia of the connective tissue that will surround the tail muscles (Peters et al. 1998b).

At the end of the analysis, it had to be recognized that yet several *Pax9* expression domains were clearly not represented in the transgenic line. Apart from the mentioned structures, no activity was observed in any somitic element, in the endoderm of the pharyngeal pouches and presumably in its derivatives, in the anterior epithelium of the digestive tract, such as the esophagus, the tongue epithelium, the salivary glands. It was not possible to clarify why so many *Pax9* expression domains could not be reproduced with the transgenic model.

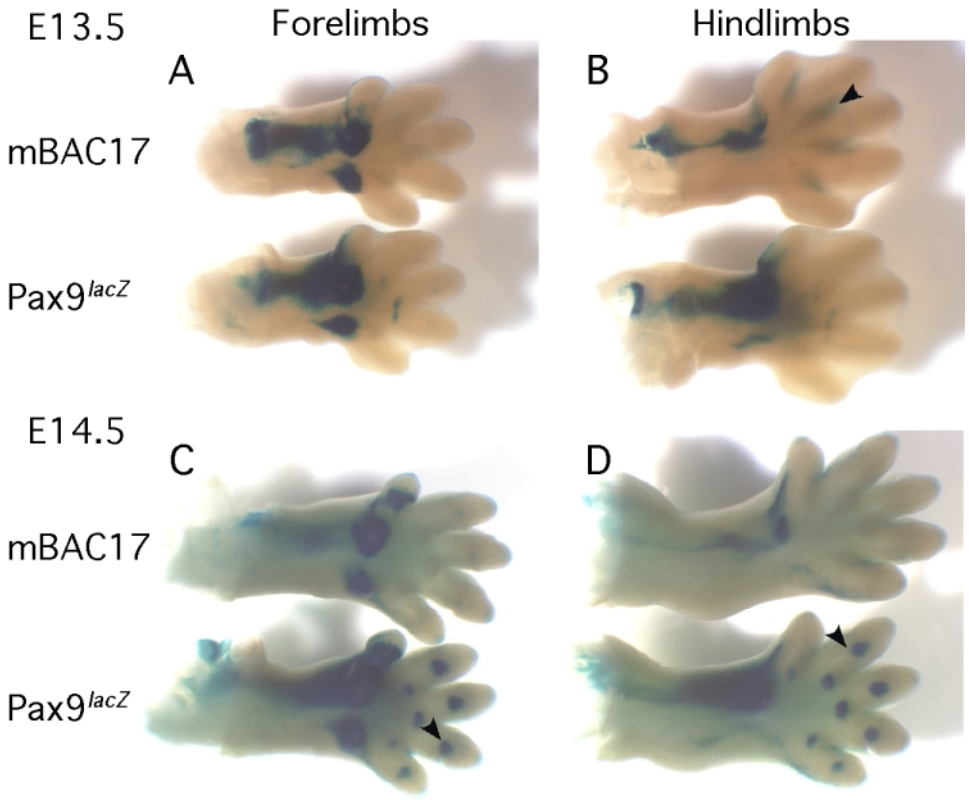


Fig. 32. BAC17-transgene expression in the developing limbs. Ventral view of X-Gal staining in limbs of E13.5 (A and B) and E14.5 (C and D) embryos, respectively; thumb is on top. In each panel mBAC17 embryo is on top and Pax9^{lacZ} on the bottom, as indicated. (A and C) Forelimbs. (B and D) Hindlimbs. Transgenic expression mostly reproduces the endogenous expression in the foot and hand pads and in the mesenchyme at the level of the forming radius and tibia. Expression fails most anteriorly, dorsally respect to the thumb. The transgene is also expressed in the interdigital mesenchyme at E13.5, differently to *Pax9* (arrowhead in B). Conversely, the normal staining in the digits at E14.5 (arrowheads in C and D) is very weak or missing.

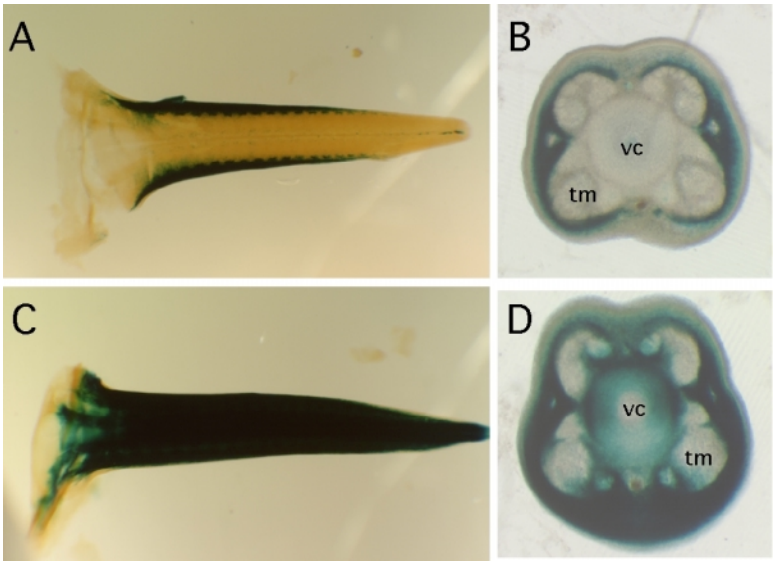


Fig. 33. BAC17-transgene expression in the developing tail. (A and C) Ventral views and (B and D) cross-sections (dorsal on top) of E13.5 mBAC17-embryo (A and B) and Pax9^{lac} embryo (C and D) tails. The transgene is expressed exclusively in the lateral mesenchyme. *Pax9* is expressed in the mesenchyme all around the tail muscles (tm). Staining in the vertebral column (vc) can be as well observed.

4.6.3. In situ analysis of BAC-transgenic mice

It emerged the possibility that the lack of X-Gal staining in many of the expected *Pax9* positive tissues could be to due to an inefficient synthesis of β-galactosidase from the respective mRNA, either because of a low and only tissue-specific functionality of the IRES sequence, that means the mRNA is synthesized but not translated, or because of such a low transcriptional activity that the amount of transcript was not enough to produce detectable enzymatic activity. It has been suggested that an in situ hybridization with a probe for the *lacZ* mRNA could be a more sensitive system and therefore provide more reliable results, because it is able to detect very few RNA molecules per cell (Harafuji et al. 2002).

Hence, an in situ hybridization with a *lacZ* probe was performed on transgenic embryonic specimens and compared to a *Pax9* in situ hybridization conducted on non-transgenic siblings. The outcome, shown in Figure 34, thoroughly confirmed the antecedent results and reproduced exactly the X-Gal staining shown in Figure 30, demonstrating that there was no discrepancy between transcription and translation efficiency of the transgene.

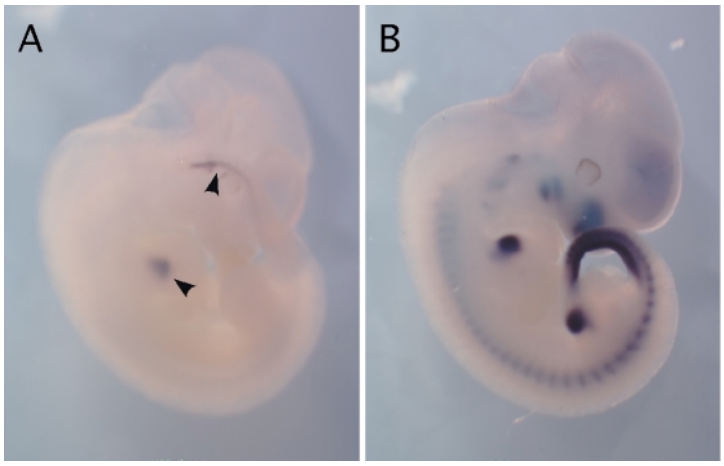


Fig. 34. BAC-transgene expression analysis by whole-mount *in situ* hybridization. (A) Whole-mount *in situ* hybridization of an E11.5 mBAC17-transgenic embryo with a *lacZ* RNA probe. The staining exactly reproduces the X-Gal staining shown in Figure 30. Arrowheads indicate the limb buds and the tail endoderm. (B) Control hybridization with a *Pax9* probe.

Unfortunately, the analysis of the second transgenic line (mBAC17-01) was unsuccessful. Although the founder animal carried both the PCR detectable portions of the BAC construct, its offspring inherited only one fragment, corresponding to the modified *Pax9* locus, while the 5'-end of the BAC was not genetically transmitted. This observation led to the conclusion that the BAC transgene in this line was somewhere truncated and that the founder transgenic mouse had a mosaic distribution of two separate pieces of the construct. The X-Gal staining

of embryos from this line revealed an almost total loss of expression. Only restricted staining spots in the developing limbs and in the face could be detected, revealing the uselessness of this transgenic line for further studies (data not shown). This finding left open the necessity to generate a new line that with the same construct would confirm the described results.

4.6.4. Rescue of Pax9$^{-/-}$ phenotype with BAC transgene

The X-Gal staining pattern observed in the transgenic mice was a clear indication of the reporter *lacZ* gene expression, but it did not reveal whether the *Pax9* gene copy, as well included in the BAC construct, was likewise expressed. This could be only inferred from the fact that the *Pax9* and the *lacZ* genes were transcribed in the same bicistronic mRNA, separated by an IRES element. Thus, the *lacZ* expression was only an indirect clue for the expression of *Pax9*. In order to directly ascertain the expression of exogenous *Pax9* (from the BAC construct), an experiment was performed, in which the BAC17 transgene was introduced in a *Pax9* deficient background trying to rescue the mutant phenotype from the transgenic *Pax9* copy.

Because of the incomplete expression of the transgene, the experiment was also of particular interest to see the effect of a partial rescue of the complex *Pax9* mutant phenotype, limited to the structures where *lacZ* expression was observed.

The generation of *Pax9* mutant mice with BAC17 transgene was accomplished with two series of cross-mating. At first transgenic mice from the mBAC17-04 line were crossed with heterozygous Pax9$^{+/lacZ}$ mice in order to obtain Pax9$^{+/lacZ}$/BAC17 compound mice. The genotyping of the animals was conducted by double PCR analysis both for the presence of the BAC and for the *Pax9* wild-type and mutated alleles. As expected, these mice did not show any apparent phenotype and they were further mated with Pax9$^{+/lacZ}$ mice, in order to obtain a Pax9$^{lacZ/lacZ}$/BAC17 progeny. For the genotyping of this generation, a Southern blot analysis was required because of the impossibility to discriminate between Pax9$^{+/lacZ}$/BAC17 and Pax9$^{lacZ/lacZ}$/BAC17 mice simply by PCR.

All the genotypic combinations of the *Pax9* alleles and the BAC transgene were observed in a normal Mendelian ratio.

The Pax9$^{lacZ/lacZ}$/BAC17 mice seemed to be characterized by the same lethal phenotype described for the Pax9$^{lacZ/lacZ}$ (Peters et al. 1998b). They showed a swollen belly and died soon

after birth. This phenotype had been described to be due to the presence of a cleft secondary palate with consequent impairment of respiration.

Surprisingly, anatomical analysis of these mice revealed a normally formed secondary palate, similar to the wild-type animals (Fig. 35). This result was in accordance with the expression of the transgene in the projecting tips of the secondary palate processes (Fig. 31) and it demonstrates that the expression level, although more restricted than the endogenous gene, is sufficient to carry out the function. Moreover, the rescue of the secondary palate defect but not of the lethality strongly indicates that this malformation is not the only cause of dearth of the Pax9 mutants, as previously described (Peters et al. 1998a).

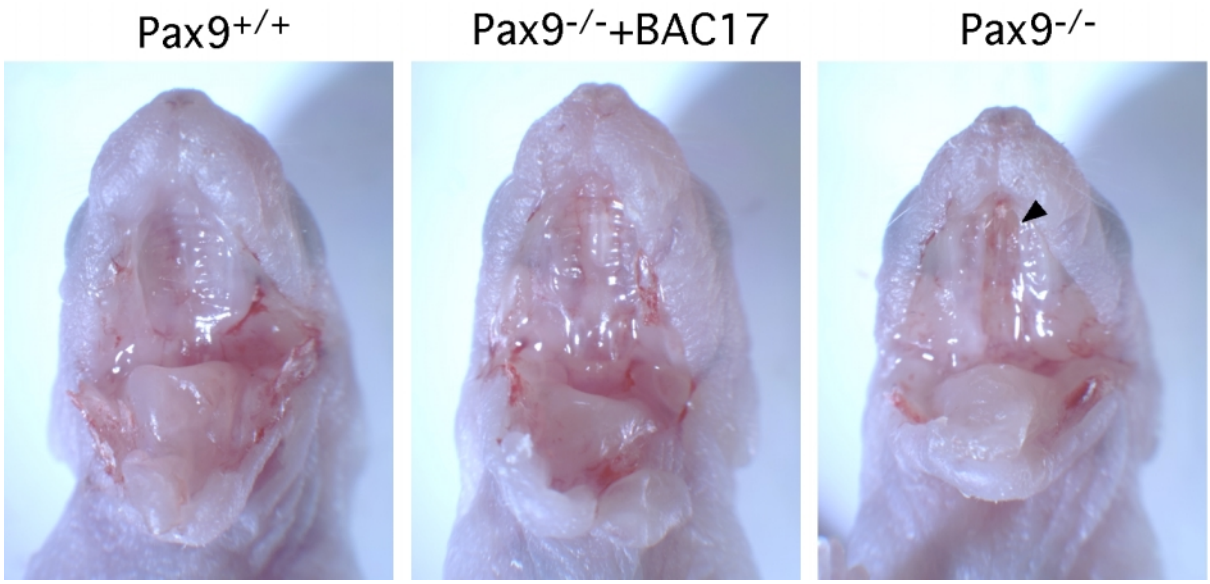


Fig. 35. Phenotypic rescue of cleft palate. Secondary palate of newborn mice with genotype indicated above each photography. The mouth opening was enlarged by cutting the cheeks along their antero-posterior length. In the $Pax9^{+/+}$ and $Pax9^{-/-}$+Bac17 mice the typical striped structure of the secondary palate is visible. Conversely, in the $Pax9^{-/-}$ mouse a central cleft leaves open view to the nasal coanes (arrowhead).

In accordance with the transgene expression in the developing limbs in a *Pax9*-like fashion, a rescue analysis of the limb defect was also conducted. The most evident limb malformation in the *Pax9* knockout mice is the preaxial duplication of the first digit in the hindlimbs (Peters et al. 1998a). As shown in Figure 36, the transgenic animals showed rather normal limbs with no sign of polydactily, suggesting again phenotypic rescue form the exogenous *Pax9* copy. No rescue analysis was possible in the tail, because the Pax9 mutant mice do not show any apparent anatomical defect (Peters et al. 1998a).

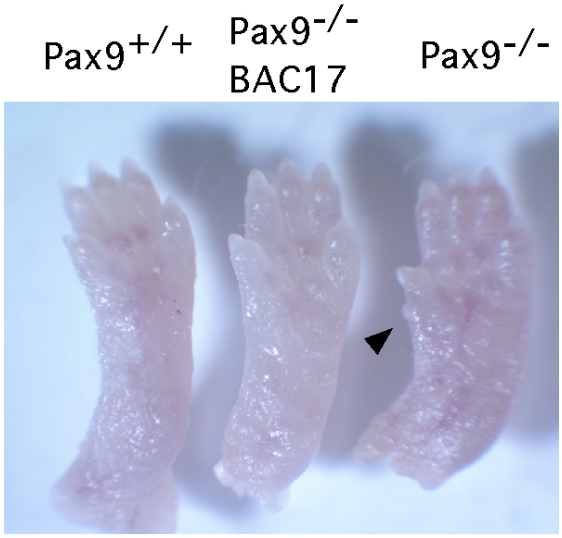


Fig. 36. Rescue of preaxial digit duplication. Left hindlimbs of newborn mice with genotype indicated above. In the $Pax9^{+/+}$ and $Pax9^{-/-}$+Bac17 mice, the general appearance of the limbs is normal with no sign of malformations. In the $Pax9^{-/-}$ mouse a smaller secondary first digit is formed (arrowhead).

4.6.5. Future BAC-transgenic experiments and construct preparation

One possible reasonable explanation for the incomplete expression pattern in the mBAC17-04 transgenic line could be that the BAC17 did not contain all the necessary elements for the full *Pax9* transcription regulation. Probably the two neighboring genes did not represent real boundaries to restrict the genomic sequence analysis. For that reason, it was decided to create a new transgenic construct, using a different BAC clone. The close vicinity of the *Odc* gene in a tail-to-tail orientation with respect to *Pax9* (the two 3'-UTRs are only 2 kb far apart) suggested a higher probability for more distant *Pax9* regulatory elements to reside on this side of the gene rather than over the *Nkx2-9* gene. The promoter of the *Odc* gene is 500 kb far away from *Pax9* and the presence of *Pax9* specific enhancers inside the *Odc* gene might not have any effect on the expression of the gene itself. Moreover, *Odc* is a constitutively expressed gene (Fiermonte et al. 2001) and extraneous regulatory elements of a different gene might be non-functional or irrelevant for its expression, while they could be deleterious for a tissue specific gene as *Nkx2-9*. Thus, the BAC15 represented the next best genomic fragment with the highest likeliness to include the specific *Pax9* control elements for the remaining expression domains (Fig. 26). BAC15 started around 10 kb upstream of *Pax9* 5'-end and stretched out to 190 kb downstream of the 3'-end with a total length of about 220 kb. It contained a 150 kb longer downstream genomic region than BAC17, including the last seven of the ten *Odc* gene exons. This BAC clone was successfully modified with an IRES-

βGeo cassette exactly as the BAC17 and the generation of a transgenic mouse line is at the moment in progress (data not shown).

Additionally, three more BAC constructs were generated. With the purpose to restrict the regions of the identified expression domains, three different large deletions were introduced in the modified BAC17. Two of these deletions covered the intergenic region between *Pax9* and *Nkx2-9*. The construct mBAC17-Δ1 carried a 47 kb deletion approximately 5 kb away from the 5'-end of *Nkx2-9*, while the construct mBAC17-Δ2 included a 41 kb deletion, 15 kb upstream of the *Pax9* gene and 26 kb overlapping with the mBAC17-Δ1 deletion.

The third deletion in the mBAC17-Δ3 included all the 3'-portion of the BAC17, including the terminal portion of the *Odc* gene (Fig. 37).

Each of these deletions was created with the same ET-cloning system that was initially used for the insertion of the IRES-βGeo cassette, confirming that this method is suitable for any type of modification of a large DNA construct. As only technical adaptation from the protocol, this second modification on the same BAC clone required the use of a kanamycin resistance marker cassette flanked by mutated Frt sites (Frt5), that would not recombine with the wild type Frt site, left over from the previous modification (see Materials and Methods and Schlake and Bode 1994).

The generation of transgenic mice with these new BAC constructs is in progress and will lead to a more precise localisation of the regulatory elements shown to be present on the BAC clone 17, allowing the creation of a large-range enhancer map as a basis for future closer investigations. The two upstream deletions were designed in order to further narrow the positioning of the enhancers. The loss of expression domains in both deletion constructs or only in one of them will restrict the localisation of the related regulatory sequences within the common overlapping region or in the outer non-overlapping regions respectively.

Furthermore, a deletion analysis of the genomic region between *Pax9* and *Nkx2-9* could turn out to be extremely interesting if it resulted in the ectopic expression of each of the two genes in the domains of the other one, unveiling the presence of a boundary element, e. g. an insulator sequence, that separates the domains of influence of the respective regulatory factors.

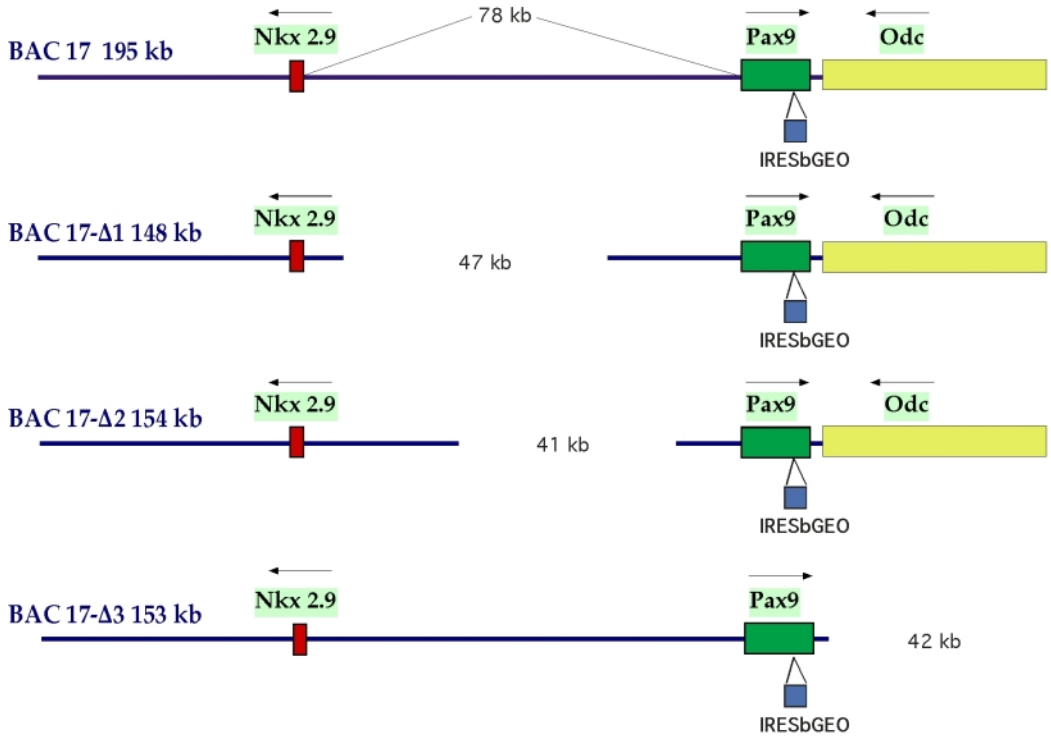


Fig. 37a. Deletion series of BAC17. Schematic representation of the modified BAC17 and the three derived deletion constructs generated by ET-cloning (BAC17-Δ1, BAC17-Δ2, and BAC17-Δ3). The three genes contained in the BAC17 are shown with the transcription orientations indicated by the arrows. An IRES-βgeo cassette is inserted in the *Pax9* gene. Sizes and distances given in kb are only indicative.

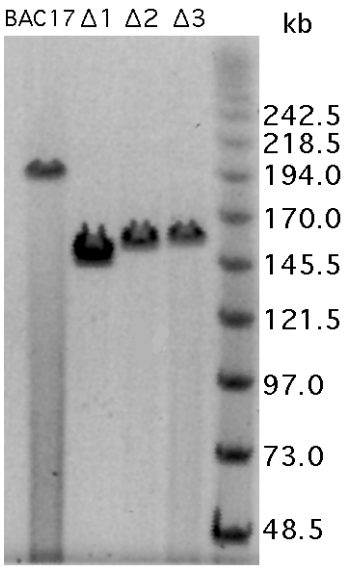


Fig. 37b. PFGE size analysis of BAC deletion constructs. The four constructs shown above in Fig. 35a were linearized with NotI and run on 1% PFGE to confirm the estimated sizes. The full-length BAC is in the first lane (BAC17). Δ1, Δ2, and Δ3 are the three deletion constructs. Notice that BAC17-Δ3 is perhaps slightly bigger than BAC17-Δ2, this inconsistency with the expected size is due to the incompleteness of the mouse sequence information.

5. DISCUSSION

5.1. Initial considerations about the project

The goal of this work was to identify the genomic sequences that regulate the *Pax9* gene expression. The isolation of such regulatory elements would be the first step for the identification of the binding transcription factors and consequently for the delineation of the molecular pathway(s), in which Pax9 is involved.

However, it is important to make some initial considerations about the feasibility of the experimental design and to explain what made this project not an easy issue to address.

The mechanisms that regulate *Pax9* expression during development are still basically unknown and appear to be rather complicated. It seems that the *Pax9* regulatory mechanisms can acquire competence to respond to particular molecular signaling in a tissue and time specific manner. The expression in the sclerotomal portion of the somites have been proven to be dependent on signaling from the adjacent notochord and this signal is mediated by the secreted factor Sonic Hedgehog (Shh) (Goulding et al. 1994; Neubüser et al. 1995; Müller et al. 1996). However, the same factor Shh coming from the epithelium of the presumptive tooth domain does not seem to be responsible for the *Pax9* expression in the tooth mesenchyme, while a different ectodermal signal, namely Fgf8, performs this function (Neubüser et al. 1997; Dassule et al. 2000; Mandler and Neubüser 2001). This suggests that, if there is a direct regulation by Shh signaling effectors (like the Gli transcription factors) in the sclerotome, the corresponding regulatory elements of the *Pax9* gene can modulate their availability to be bound by these effectors in different tissue types. One possible explanation could be that transcription factors from one single pathway are not enough to initiate *Pax9* transcription by themselves and that they require a synergistic co-operation of other specific factors, binding to other genomic elements. *Pax9* expression in the sclerotome depends for instance also on cell-autonomous mechanisms. Mutations in genes expressed in the somites and responsible for the normal patterning and antero-posterior polarization of the somites can result in *Pax9* downregulation, even without affecting the Shh signaling coming from the notochord (Mansouri et al. 1999; Leitges et al. 2000). If the two or more regulatory components

bind to sequences residing far apart from each other in the *Pax9* locus, their identification might result quite complicated.

Another consideration regards the location of the *cis*-regulatory elements with respect to the *Pax9* proximal promoter sequence. In many cases the analysis of up to 10 kb genomic sequences directly upstream of the transcription start site has led to the identification of part or all of the main regulatory elements of various genes, including several transcription factors with a complex expression pattern during development (some examples in Logan et al. 1993; Sasaki and Hogan 1996; MacKenzie et al. 1997; Kuschert et al. 2001). Conversely, a similar approach for the *Pax9* gene did not produce consistent results that would suit with the presence of real regulatory elements in the tested 15 kb upstream genomic sequence (H. Peters, unpublished).

It is known that promoters can be regulated by genomic sequences located at considerable distances from the transcribed regions. Molecular models have been proposed, in which facilitator factors between enhancers and promoters would then bring them into physical proximity to each other, overcoming the big genomic distance (Dorsett 1999). Examples of such long-range regulatory elements have been found for the human gene *IL5*, which has an enhancer 120 kb far off from the gene itself (Loots et al. 2000), and for the human *SOX9* gene, whose complete expression pattern could be only reproduced in transgenic mice carrying a 350 kb long YAC construct (Wunderle et al. 1998).

The large range genomic analysis established in this work was founded on the possibility that a similar situation could also apply to the *Pax9* gene.

This hypothesis was moreover sustained by observations conducted on *Pax1*, the paralogous gene highly related to *Pax9*. A transgenic approach with BAC clones (up to 130 kb long) encompassing the *Pax1* locus was not sufficient to rescue the *Pax1* knock-out phenotype, suggesting that the BAC sequences were not long enough to contain all the required elements for the normal expression of the gene (Kokubu et al. 2002).

A BAC based approach was nevertheless chosen also for this work on the *Pax9* gene. There was expectancy indeed that better results could be obtained for *Pax9*, first because of the use of a different BAC library (RPCI-23 mouse BAC), which consisted of significantly longer genomic inserts with an average of 200 kb in length, and second

since the finding that the intergenic regions between *Pax9* and its neighboring genes were much shorter than for *Pax1*, as it is discussed further on in this section.

5.2. Structural conservation of the *Pax9* gene

Eleven different *Pax9* BAC clones were isolated and a series of Southern blot hybridizations, using *Pax9* specific probes and BAC end probes, were performed in order to establish an ordered contig around the gene locus. The contig covers a genomic region of almost 400 kb in length and represents now a useful tool for further genomic analysis on *Pax9* and its neighboring genes.

Before starting the real search for the promoter and *cis*-regulatory sequences, the first part of this work was focused on the determination of the gene structure. The importance of knowing the structure of a gene in this type of studies relates with the necessity to eventually locate the identified regulatory sequences with respect to the gene itself. It has been observed that enhancers can be found not only in the vicinity of the promoter sequence at the 5'-end of the gene, but also within the gene itself, in the introns or in the 5'- and 3'-untranslated regions, or very often downstream of the polyadenylation signal (see for example Aparicio et al. 1995; Kwan et al. 2001; Morishita et al. 2001).

An accurate analysis of the human and mouse genes, supported by the availability of the human genomic and cDNA sequences and by the mouse cDNA sequence, allowed to define the structure of *Pax9* in both species. Some insight about the *Pax9* gene structure was already accessible thanks to previous data, reported in Peters et al. for mouse *Pax9* and in Stockston et al. for human *PAX9* (Peters et al. 1998b; Stockton et al. 2000). However, a detailed analysis of the complete exon-intron organization, including the exact exon-intron boundaries location, was still missing. The high degree of conservation between the two orthologous counterparts in mice and humans, which had been observed at the cDNA and gene product level (Peters et al. 1997), could be extended as well to the genomic level. Both genes feature a very similar physical arrangement consisting of 4 exons, which are distributed over 16 kb of genomic region, and showing identical size and localization of the interposing introns. A similar exon-intron organization had been shown for zebrafish *Pax9* (Nornes et al. 1996).

Beside the human and mouse *Pax9* genes, this work has reported the isolation and structural characterization of the orthologous counterpart from the Japanese pufferfish (*Takifugu rubripes*), commonly known as *Fugu*. The striking conservation of the paired domain within the first subgroup of the Pax family (Hetzer-Egger et al. 2000) has made it possible to design degenerated primer sequences for a first PCR screening of a *Fugu* cosmid library and subsequently to refine the screening among the isolated clones by cross-species hybridization with a mouse specific paired domain probe. The *Fugu Pax9* genomic sequence was consequently isolated, but since no *Fugu Pax9* cDNA sequence was available, the further structural characterization of the gene had to be conducted taking the zebrafish *Pax9a* cDNA as a reference for the determination of the exon-intron boundaries (Nornes et al. 1996).

The *Fugu Pax9* gene revealed a similar organization as its mouse and human counterparts (Fig. 9). It is as well composed of four exons and the intron positions coincide with the ones described for the other two species. The first intron is located after the first nucleotide of the second codon; the second intron occurs shortly after the octapeptide sequence; the third intron occurs after a short exon with little sequence conservation. The conservation of the gene structure does not surprise considering the latest data regarding the human-pufferfish genome comparison. It has been observed that almost all of the analyzed *Fugu* genes tend to maintain the same organization as in their human counterparts (Brunner et al. 1999; McLysaght et al. 2000). In particular in McLysaght et al., 199 pairs of orthologous introns from the corresponding 22 genes were found between human and *Fugu*. There were only six cases where an intron was present in one sequence but there was no equivalent intron nearby or out of phase in the other species. These observations suggest that the maintenance of the exon-intron structure of a gene is a common feature within the vertebrates and that a gene organization is likely to be strictly linked with its functionality.

In the case of the *Pax9* gene, the structure conservation is even more striking, considering that the amphioxus *Pax1/9* gene counterpart (*AmphiPax1*), which is considered related to the primitive ancestor gene of the vertebrate Pax1/9 gene subfamily, shows the same exon-intron organization as *Pax9* (Hetzer-Egger et al. 2000).

The only appreciable difference could be noticed in the gene size. *Fugu Pax9* is only about 6 kb long; that means that the mammal orthologues are about 2.7 fold longer. Even though the unavailability of the *Fugu* mRNA sequence including the 5' and 3'-UTRs did not allow a precise size estimation, this size difference agrees with the notion that the pufferfish genome is about 7.5 fold smaller than the human genome (400 Mb versus 3000 Mb), which is principally not due to a lower number of genes but to a reduced amount of repetitive and non-coding sequences (Elgar et al. 1996; Koop and Nadeau 1996).

The compaction of *Fugu* genes has been shown to be a general feature of almost all the analyzed *Fugu*-human gene pairs and it is mainly accounted for by a substantial difference in the intron size (McLysaght et al. 2000). A similar observation can be made for the *Fugu Pax9* gene, where each of the three introns appears to be smaller than the respective human or mouse orthologue, even though the compaction ratio does not seem to be homogeneous. Rather, the third intron is 4 fold shorter than the human/mouse counterpart, while the first and second introns show only a 2-fold shrinkage (Fig. 9). Assuming that the general function of orthologous genes in different species can be considered largely conserved, it is intriguing to think that the contraction of DNA sequences in the *Fugu* genome corresponds probably to a loss of rather non-functional DNA, where by functional it is not necessarily meant coding but also regulatory. The unique introns of the *Fugu* and mouse *Hoxb-4* genes have a similar size. Sequence and functional analyses have proven that this size conservation is due to the presence of a transcriptional regulatory element within the intron of the gene in both species, which has obviously determined a selective constraint during evolution against size reduction in the *Fugu* counterpart. If the non-homogeneous contraction of the *Fugu Pax9* introns is similarly due to a different content of functional information (i.e. less in the third intron respect to the first two), it cannot be discussed with the present data, but it remains anyway an appealing hypothesis.

Remarkably, a certain size reduction could be as well noticed in the coding sequence, being the deduced *Fugu* Pax9 protein about 10 aminoacids shorter than the Pax9 proteins from the other compared species, including the other teleost zebrafish (Fig. 13). It has to be taken into consideration that the *Fugu* Pax9 mRNA sequence is not known and that these data are only based on an extrapolation from the genomic sequence in correlation

with the zebrafish gene structure. However, the encountered alignment gaps with the other Pax9 sequences are very likely to correspond to the real situation, since they do not occur in proximity of the deduced exon-intron boundaries, where the sequence assembling is more error prone, but in the middle of exons flanked by homologous regions. In spite of the size discrepancy, the overall aminoacid identity between the mammal and the *Fugu* Pax9 proteins reaches up to 73%, which is considerably high when compared for instance to similar orthologue pairs, like *Fugu* and mouse Hoxb-4 proteins (56% identity) or *Fugu* and human Etv6 proteins (58% identity), both defined to be very conserved (Aparicio et al. 1995; Montpetit and Sinnett 2001). The sequence identity increases even to 98% inside the paired-domain and to 100% in the octapeptide domain, in agreement with the striking conservation of the two domains within the members of the Pax1/9 gene subfamily (Hetzer-Egger et al. 2000).

5.3. Conserved association to *Nkx2-9*

Another interesting finding that emerged from the analysis of the in human, mouse and pufferfish *Pax9* genomic regions was the striking conserved locus synteny. Starting from the information available with the human *PAX9* genome sequence regarding the presence of another gene 80 kb upstream, namely *NKX2-8*, it was consequently found that the same physical association was existing also in the mouse and *Fugu* genomes (Fig. 9). Actually, no direct *NKX2-8* orthologous gene had been yet definitely described in the mouse; however, sequence comparisons revealed that the mouse *Nkx2-9* was the closest related gene to human *NKX2-8* among the members of the NK-2 transcription factor family. The two genes were independently and almost contemporaneously described (Apergis et al. 1998; Pabst et al. 1998), but due to the different nomenclature, their direct orthology was never recognized, neither in later publications (Wang et al. 2000),

PCR and Southern blot analyses were performed on the mouse BAC clones and on the *Fugu* cosmid clone and allowed to successfully confirm the presence of the *Nkx2-9* gene in both organisms in association with the *Pax9* gene. In effect, no Fugu *Nkx2-9* gene was so far known; in fact since no counterpart had been isolated in non-mammalian vertebrates, it was even suggested that it could represent a new gene arisen late in some branches of the vertebrate lineage (Wang et al. 2000), However, the gene showed the

highest similarity to the human and mouse *Nkx2-9* gene and in addition to that it displayed the same physical association with the *Pax9* gene. These two features together strongly suggested that the identified gene was indeed the Fugu *Nkx2-9* gene.

The structural organization of the three *Nkx2-9* orthologues was determined by sequence comparison with the available cDNA sequences and identification of conserved splicing sites. A sequence alignment showed that the *Fugu* Nkx2-9 protein was homologous to the two mammal counterparts only in the three conserved domains, described for the members of the Nk-2 transcription factor family (Harvey 1996). Notably, Pabst et al. pointed out the absence of the TN domain near the amino terminus of the mouse Nkx2-9 protein, which is conversely present in all known Nk-2 genes, and therefore hypothesize a divergent origin and a different role for this gene (Pabst et al. 1998). In discord to that, a TN domain is here described in the Nkx2-9 protein of the three species (Fig. 13), even though it deviates significantly from the proposed consensus (Harvey 1996).

Regarding the *Nkx2-9* gene structure in the three organisms, it is implicit to make the same considerations brought up before about the *Pax9* gene. Again the *Fugu* gene shows the shortest gene length, which is mostly accounted for by a shorter intron sequence. The same principle of the general *Fugu* genome contraction can also explain the shorter intergenic distance between the two genes, which is only 10 kb in the *Fugu* sequence, in contrast to 80 kb in the human situation and 75 kb estimated for the mouse (Fig. 9).

5.4. Evolutionary considerations about the conserved syntenic region

It has recently been reported that *Nkx2-9* is closely linked to its related gene *Nkx2-1* on mouse chromosome 12 and human chromosome 14 (Wang et al. 2000). However, no information about the distance between the two genes is provided. In the present work, it was possible to identify the presence of the mouse *Nkx2-1* gene on the most 5' clones of the *Pax9* BAC contig, suggesting a physical distance of about 70 kb from *Nkx2-9*. Unfortunately, the *Fugu* cosmid clone did not extend enough from the *Nkx2-9* gene to verify a similar association with the *Nkx2-1* gene. Nevertheless, the *Fugu* genome annotation at Ensembl (http://www.ensembl.org/) includes this gene in the syntenic group.

Comparisons between the human and mouse genomes suggest that 1793 orthologous gene pairs fall into 201 synteny groups (DeBry and Seldin 1996, and its electronic update

available at http://www.ncbi.nlm.nih.gov/Homology/). It is predicted that the number of human-mouse syntenic groups will remain about 200 regardless of the further introduction of newly mapped genes into the comparative maps (Nadeau and Sankoff 1998). Obviously, the degree to which fragmentation has occurred between the genomes of lower vertebrates (like the fish) and mammals during 400 million years of evolution is expected to be much higher compared to species within the mammalian class, which diverged not earlier than 70 million years ago (Elgar et al. 1996; Koop and Nadeau 1996). However, even though no conclusion can yet be reached concerning large regions of DNA due to the unavailability of an arranged *Fugu* genome sequence, short-range conserved synteny has been demonstrated for a number of adjacent *Fugu* genes versus the equivalent human orthologues (Elgar et al. 1996; Brunner et al. 1999; Elgar et al. 1999; McLysaght et al. 2000). Despite the controversial results from different authors due to different stringency degrees on the definition of orthology, it was calculated that for at least 45% of linked *Fugu* genes the human orthologues were mapped on the same chromosome (McLysaght et al. 2000).

A fascinating aspect of these studies would be to understand whether the synteny conservation between distantly related species is only the result of incomplete genome shuffling, which has involved random blocks of genes instead of single gene units, or if the process was to a certain extent controlled by selective pressure, which operated in order to maintain compact clusters of linked genes. The Hox genes, for example, are organized in clusters in all the lower and higher metazoan species so far studied. The rigid preservation of this multigene organization is due to the sharing of common regulatory elements and to a global equilibrium of the transcriptional control (Duboule 1998).

Pax9 and *Nkx2-9* are both transcription factors with definite patterning roles during embryonic development but they do not show any type of expression overlap, being *Nkx2-9* exclusively expressed in the developing neural tube (Pabst et al. 1998). As already discussed above about *Pax9*, also *Nkx2-9* expression depends on Shh signaling, in particular emanating from the floor plate (Pabst et al. 2000). This initially raised the hypothesis that the two genes might reside in a common Shh control DNA region, even

though resulting in expression in different domains. Nevertheless, more recent results presented in this work tend to argue against this possibility (see 5.11.).

It is difficult to say if *Pax9* and *Nkx2-9* share some common regulatory mechanisms or if there is any other impelling force that binds them together or if their conserved linkage is just a random result of chromosomal rearrangements.

Recently, a similar situation has been described for the *Pax6* gene. A conserved syntenic locus comprising four genes, including *Wt1* and *Pax6*, has been identified in human and *Fugu* and, also in this case, a related possible functional significance associated to the *Pax6* regulation was proposed (Kleinjan et al. 2002).

Further considerations on this topic render the hypothesis of a functional linkage between *Pax9* and *Nkx2-9* still credible. Wang et al. have shown that *Nkx2-4* and *Nkx2-2*, the paralogous genes of *Nkx2-1* and *Nkx2-9*, respectively, are also linked on mouse chromosome 2 and human chromosome 20 (Wang et al. 2000). In addition to these data, the new human genome annotation has revealed the presence of *PAX1*, the paralogous gene of *PAX9*, in the vicinity of *NKX2-2*, confirming the original mapping data (Stapleton et al. 1993). The same linkage was observed for mouse *Pax1*, which as well maps next to *Nkx2-2* and *Nkx2-4* (Wang et al. 2000 and mouse genome annotation at Ensembl - http://www.ensembl.org/). Apparently, two equivalent blocks of paralogous genes have preserved their association during evolution at least in mammals and in the case of the *Pax9* syntenic region also in the fish and probably in all the vertebrates. Intriguingly, recent *Fugu* genome annotation data at Ensembl confirm the *Nkx2-4*/*Nkx2-2*/*Pax1* physical association also in this species.

It is also interesting to notice that lower chordates have only one copy for the paralogous pairs *Pax1/Pax9* (Holland et al. 1995; Ogasawara et al. 1999), *Nkx2-2/Nkx2-9* (Holland et al. 1998), and *Nkx2-1/Nkx2-4* (Venkatesh et al. 1999). This is not an exceptional case. Most of the vertebrate gene families that include two, three or four paralogous members for each gene type are restricted to only one member per paralogous group in the lower chordates (amphioxus and the tunicates); for example one Msx gene instead of the three found in the vertebrates or only one Hox gene cluster instead of four (Holland et al. 1994). It has been now firmly recognized that many paralogous gene pairs have arisen from common ancestor genes in the context of two waves of whole genome duplication events

that occurred through the establishment of the vertebrate lineage from the first primitive chordates. The duplicated genes have evolved independently acquiring new distinct functions accountable for the increasing complexity of the vertebrate body plan (Holland et al. 1994). The elaboration of the brain and its specialization in fore-, mid- and hindbrain regions and the onset of endoskeletal elements like cartilage and bone or other mineralized tissues (e. g. teeth) can be mentioned among the innovations of the vertebrate body plan (Shimeld and Holland 2000). The direct involvement of the Nk2 and Pax genes respectively in the development of these structures is a proof of the importance of gene duplication and diversification.

If the lower chordates represent a model for the ancestral genome (Corbo et al. 2001), then it can be logically deduced that a physical association between the ancestral *Pax1/9*, the ancestral *Nkx2-2/-9*, and the ancestral *Nkx2-4/-1* genes already existed in a primitive situation and that the entire locus duplication has generated the two syntenic groups presently known (Fig. 38). According to this hypothesis, the physical association of *Pax1* and *Pax9* with the Nk2 genes originated long before the vertebrate evolution and still has been preserved up to the present time. These observations do not add any direct functional evidence to the conserved synteny but diminish credits to a simple interpretation based on random genome shuffling. This important remark will be brought up again further on in the discussion of the final results (see 5.15.).

5.5. More insight in determining the *Pax9* mRNA structure

In the context of the determination of the *Pax9* gene structure, some work was done trying to determine the complete mRNA sequence. The original available information was limited to two cDNA clones, described in Neubüser et al., which comprehended a total length of 2.5 kb and therefore did not account for the over 4.5-5 kb band detected by Northern blot analyses (Neubüser et al. 1995; Peters et al. 1997).

The Northern blot data were reproduced in this work using two different *Pax9* probes on embryonic RNA extracts and on a commercial RNA blot from different adult mouse tissues (Fig. 14). The use of a paired-box probe on 11.5 dpc embryonic samples brought about the detection of two bands both in tail and in limb bud extracts. One band about 4.5 kb long corresponded to the transcript length described (Neubüser et al. 1995). The 2.5

kb band of weaker intensity did not match to any previous data. The possible existence of two different isoforms for *Pax9* was not confirmed by the Northern blot results obtained from the mouse adult tissue membrane, where only the 4.5 kb band could be detected. In this case a different probe was used from the exon 4 sequence. A possible interpretation for the additional 2.5 kb band could be a cross-hybridization of the paired-box sequence with a different transcript, maybe *Pax1*, whose paired domain is highly homologous to *Pax9* (Neubüser et al. 1995; Hetzer-Egger et al. 2000) and which is also expressed in the tail and limb buds of the same stage developing mouse embryos (Deutsch et al. 1988; Timmons et al. 1994). However, a similar probe was used in Neubüser et al. on embryonic extracts without clear detection of additional bands. Furthermore, the *Pax1* transcript size has been described to be around 3 kb (Deutsch et al. 1988), even though the lack of precise RNA molecular weight markers makes the size estimation of RNA molecules often arduous.

Interestingly, in Peters et al. two *Pax9* transcripts were detected in mouse adult tissues, being the sizes somewhat different from what reported here, 5.3 kb and 2.2 kb, and in a human esophagus extract even three bands could be found (Peters et al. 1997). Another major difference between the Northern blot data shown in Peters et al. and in this work is the type of tissues where a *Pax9* transcript was detected. Apart from the thymus, which was confirmed to be *Pax9* positive in agreement with previous expression data (Neubüser et al. 1995; Peters et al. 1997) and with the assessed functional role of the gene in the development of this organ (Hetzer-Egger et al. 2002), two more organs appeared as well positive, the stomach and the lungs. These two organs had been formerly shown clearly not to express *Pax9* (Peters et al. 1997). Moreover, no functional data are available about a possible role of the gene in their formation or in their physiology. Nevertheless, *Pax9* expression could be detected by X-Gal staining in the bronchi and bronchioles of $Pax9^{lacZ}$ mice (I. Rodrigo, personal communication) and this relates with the documented expression of the gene in similar cartilaginous structures, such as the larynx and the thyroid cartilage (Peters et al. 1998b). As for the expression in the stomach, *Pax9* expression had been found particularly in the epithelium of the forestomach in continuation with the esophagus epithelium (Peters et al. 1998b). At the light of these considerations, it can be said that the controversial results of the Northern blot analysis

critically depended on the way the tissue sources were dissected and on the inclusion of neighboring tissues.

A 5' and 3' RACE PCR analysis did not provide sufficient information to explain the size discrepancy between Northern blot data and cDNA sequence. The identification of a new exon, 3.7 kb upstream of the exon1, and the extension of the total cDNA sequence of 431 bp cannot account for the missing mRNA sequence, because it only comes up to about 3 kb. The attempt to further elongate the 5'-end sequence directly adjacent to the newly identified exon 0 was unsuccessful. Similarly, the 3' sequence could not be extended. However, the absence of a canonical polyadenylation signal AATAAA, or of the only known relatively common variant ATTAAA (reviewed in Wahle and Ruegsegger 1999), directly upstream of the poly(A) tail suggests that the mRNA sequence might rather continue on this side and that the actual 3'-UTR of the gene is longer than in the cloned cDNA.

The closest consensus-like putative polyadenylation signal can be found 1575 bp in the genomic sequence further downstream of the published 3'-end of the gene. Remarkably, a GT-rich element, a downstream element usually located roughly within the first 30 nucleotides from the transcript cleavage site (Wahle and Ruegsegger 1999), is also present. If these sequences identify the real *Pax9* 3'-end, then the *Pax9* mRNA contains an extraordinary long 3'-UTR (2.9 kb) and its total length would finally match the size determined by Northern blot.

This point could be very important in the context of *Pax9* regulation. It is known that 3'-UTRs carry out fundamental regulatory roles and they are in most cases decisive for the fate of a particular mRNA. They are for instance responsible for the transcript stability, determining a long or short half-life of the mRNA in accordance with the long- or short-term activity of the gene respectively, and for its translatability through the interaction with regulatory binding factors (reviewed in Grzybowska et al. 2001). A very long 3'-UTR could be an indication of such a particular function.

Unfortunately, the 3'-RACE PCR approach was not helpful to prove the existence of such a long 3'-UTR, leaving the question unsolved. The difficulties in amplifying this sequence might have been due to inaccessible secondary structures formed by the mRNA.

Even though no canonical polyadenylation signal could be found close to the known 3'-end, it is difficult to give another explanation for the presence of a poly(A) stretch in the cDNA sequence. The shorter 1.6 kb cDNA described in Neubüser et al. is obviously the artificial product of a poly(T) primer alignment on a 10 A stretch present inside the 3'-UTR. The same cannot be argued for the longer cDNA clone. It is rather likely that this cDNA derives from a real mRNA product, generated by a non-canonical poly(A) signal, and represents the minor 2.5 kb mRNA band, only detectable with high signal intensity Northern blot hybridizations (e. g. with the embryo extracts in this work and in Peters et al. 1997). In this case, the possibility of an alternative polyadenylation signal usage should be taken into consideration, but the lack of solid evidence does not leave space for further discussion.

In conclusion, the divergence of these data does not help determine the real *Pax9* transcript size, neither it certainly proves the existence of different isoforms. With respect to this point, it has to be said that two distinct *Pax9* transcripts have been isolated in zebrafish, *Pax9a* and *Pax9b*, the former of which represents the homologous form to the mouse and human counterparts. The alternative isoform *Pax9b* originates from the splicing over of the third exon and codes for a 73 aa shorter protein to due a frame-shift in the C-terminal domain (Nornes et al. 1996). A series of RT-PCR (not shown here) excluded the possibility of a similar alternative splicing for the mouse *Pax9*. Nornes et al. identify a possible explanation for the alternative isoform in the splicing donor site of the third intron, where an A at position +5 of the consensus $GT^{A}_{G}AG$ might be responsible for the occasional skipping of exon 3. Conversely, a canonical donor site GTGAG is present in the intron 3 of the mouse gene.

The detection of a shorter RNA band (about 3 kb) in the testis sample of the adult tissue Northern blot initially brought up the idea that an isoform of the gene could be transcribed in this tissue (Fig. 14). However, a series of non-consistent RT-PCR data and the unsuccessful attempt to clone a *Pax9* cDNA from a mouse testis cDNA library did not allow to verify this hypothesis.

5.6. *Pax9* transcription is driven by two alternative TATA-less promoters

The presence of an additional exon at the 5' end of the mouse *Pax9* gene (Fig. 15) does not alter the basic conserved structure of the gene discussed above. The exon 0 does not seem to constitute a strong component of the *Pax9* transcript, since it is very low represented in the total mRNA population and it probably does not possess a particular functional significance.

The presence of two alternatively used promoters has sometimes been suggested to have a relevant role for the function of the gene itself. The quail *Pax6* gene can be transcribed from two different promoters about 3 kb far apart from each other. The situation resembles that of mouse *Pax9*. The two *Pax6* isoforms, synthesized from promoter P0 and P1 respectively, differ in the 5'-UTR. The P0 mRNA contains an additional exon0, as in *Pax9*, but the exon1 is shorter than in the P1 mRNA, while in *Pax9* isoform A exon1 is longer. It seems that the activities of the two *Pax6* promoters are temporally shifted in the development of the neuroretina. Promoter P1 is earlier activated in neuroretinal cells but later on the promoter P0 takes over for *Pax6* transcription and promoter P1 is slowly switched off (Plaza et al. 1995).

The authors interpret this promoter switch as a regulatory mechanism for the *Pax6* transcript level in the cells. They describe a change from the weak promoter P1 to the stronger promoter P0, registering an increase in the mRNA amount as the cells proceed in differentiation.

The lower activity of promoter A, verified both by in situ and RT-PCR analysis, might suggest that a similar transcription regulation could take place also for *Pax9*. If it is so, a closer investigation at a single tissue level should be performed, as in the case of *Pax6*. However, the present experimental progress does not provide any indications about which tissue(s) might be possibly involved in this type of analysis.

Another example of multiple promoter usage is in the human fibroblast growth factor 1 (*FGF1*) transcription, which is controlled by at least four distinct promoters in a tissue specific manner. Promoter 1.A is active in the kidney, 1.B in the brain, and 1.C and 1.D in a variety of cultured cells induced by different biological response effectors (reviewed in Chiu et al. 2001).

In the case of *Pax9*, no apparent differential usage of the two promoters could be demonstrated by *in situ* hybridization on developing embryos. Moreover, neither of the two distinct *Pax9* expressing cell lines showed appreciable activity of promoter A. Thus, the available data do not permit at the moment to conclude if there is any functional relationship between promoter A and B and tend to support the hypothesis that promoter B has to be considered the principle *Pax9* promoter, while promoter A rather shows a background activity.

An interesting point emerged in the analysis of the two *Pax9* promoters is that they are both lacking the common control element known as TATA box. The TATA box is usually located 25-30 bp upstream of the transcription start site and directs accurate transcription initiation. However, many promoters do not contain consensus TATA boxes, or even non-consensus TATA boxes and, although some TATA-less promoters retain the ability to direct transcription initiation from a specific nucleotide, others appear to have multiple start sites, ranging from few clustered to dozens spanning hundreds of nucleotides (Smale 1997). It appears to be the case for the *Pax9* promoter(s). Promoter A can direct transcription from at least 6 different start sites in an interval of 140 bp, while for promoter B 4 start sites were identified within 70 bp (Fig. 15). It seems that the strength of the promoter is correlated to its stability, showing the weaker promoter A a broader oscillation and uncertainty in the initiation.

Some alternative features can characterize TATA-less promoters. Usually, an initiator element (Inr) surrounding the transcription start site(s) assists the function of the TATA-box for the formation of the initiation complex and can take over the complete function in TATA-less promoters (Smale 1997). However, the consensus sequence for the Inr is so loose that it was extremely difficult to make a consistent prediction about its position in the *Pax9* promoter without a functional assay.

Ince and Scotto have identified a downstream element in almost all of the analyzed TATA-less promoters with multiple start sites. This element, called MED-1, has the quite conserved consensus sequence GCTCCC/G and it is able to define a distinctive window of multiple start sites (Ince and Scotto 1995). Strangely, no similar sequence was found around the *Pax9* transcribed sequence, which might suggest that the *Pax9* promoter does not fall in any of the typical promoter classes. However, the MED-1 element was isolated

from the alignment of a group of promoters sharing similar structural features, using the P-glycoprotein promoter as a model (Ince and Scotto 1995). This selection could have created a bias for only a subclass of promoters and the presence of a MED-1 element might not be so general as suggested by the authors.

In spite of all, the only formal proof that the sequences upstream of the identified 5'-ends were real promoters came only with the functional assay.

Unfortunately, only the promoter B showed a certain activity. Promoter A did not seem to function in this model system more significantly than a background level. Although one explanation could be found in the absence of the exon 0 isoform in the two *Pax9* expressing cell lines, on the other hand the basal activity registered for promoter B could be observed also in a non-specific cell line, namely NIH-3T3 cells. That means that a normal basal promoter activity is potentially measurable in any cell system and does not require a particular specificity. This phenomenon is not unusual in this kind of experimental assay. For example, the mouse *Sox9* promoter displays a similar level of activity in transfected cells from testis, ovary and liver, despite the fact that only the gonadal cells and not the liver cells were shown to express the gene (Kanai and Koopman 1999). In fact, DNase I hypersensitive site analysis clearly demonstrated that normally the endogenous *Sox9* promoter was in a close inactive conformation in the liver cells. The authors find an explanation in the fact that the extrapolation of the promoter sequence from its native genomic location into an episomal construct sets the sequence free from the chromatin conformation and enables it to recruit the transcription factors. A similar situation might be happening for the *Pax9* promoter B in the NIH-3T3 cells, justifying the absence of specificity in the other two cell lines. It is evident that a comparable level of activity in the three cell lines identifies only a basal promoter sequence with the absence of specific regulatory elements, same conclusion as for the *Sox9* promoter.

On the other hand, the absence of specific regulatory elements in the proximity of the promoter could be already presupposed from the failure of the first transgenic experiments mentioned before.

The impressive burst of activity exhibited upon deletion of more 5'-sequences of the promoter B construct is likely to be due to a greater accessibility of the basic transcription machinery on the DNA construct than on the sudden exposition of new regulatory

elements. Although the general promoter activity seemed to acquire higher specificity for the *Pax9* positive cells, the absence of more consistent data does not permit to draw any conclusion in this respect.

However, the deletion analysis has allowed the restriction of the basal promoter to not more than 400 bp upstream of the most 5' TSS and the loss of activity with the deletion of the TSS region has conferred authenticity to the experimental system, making possible to declare that the tested constructs proved to really enclose the *Pax9* promoter function.

The functionality of promoter A remains to be demonstrated, but the RT-PCR and in situ hybridization data on embryos are incontestable evidence for the presence of an additional upstream promoter.

5.7. Identification of candidate regulatory elements through comparative sequencing

The large-scale sequence alignment between the *Pax9* genomic regions from humans, mice and pufferfish was one of the two methods adopted for the identification of regulatory elements, based on the assumption that the patterns of gene regulation and the corresponding regulatory controls are often conserved across species (Duret and Bucher 1997; Hardison 2000; Wasserman et al. 2000). The application of the PIP algorithm (Percent Identity Plot) resulted in the detection of conserved fragments between the human, mouse, and *Fugu* sequences (Fig. 18 and 19).

The human-mouse alignment showed a very elevated sequence homology within the whole locus with peaks of identity fragments scattered all over the analyzed region, both inside and outside of the transcribed domains. Only some gaps were found in correspondence with repetitive DNA elements, previously properly masked, or with recognizable intervening insertions either in the mouse or in the human sequence that interrupted the homology continuity. Similar results have been obtained in other studies, always showing that the two genomes have maintained, in spite of 90 million years of evolutionary divergence, a high degree of sequence conservation (Göttgens et al. 2000 and reviewed in Hardison 2000).

This astounding conservation between the genomes of two species that apparently look so different is in reality the reason why the mouse is such a good model for human genetics. Nevertheless, in this experimental approach such a sequence similarity might rather be a

problem. Even though examples exist in which conserved regulatory elements have been found by human-mouse sequence comparison (Loots et al. 2000), the identification of so many conserved-non-coding sequences (CNSs) may lead into an intricate web of false positive results (Duret and Bucher 1997; Göttgens et al. 2000).

One suggestion to elude this problem was the development of new algorithms with the purpose to combine qualitative and quantitative comparisons, in order to specifically isolate only highly conserved, ungapped blocks in which regulatory elements are most likely to reside (Wasserman et al. 2000). However, looking at more distantly related species would be a more sensitive way to address this issue (Elgar et al. 1996; Duret and Bucher 1997; Göttgens et al. 2000).

For this reason, the comparative sequencing was extended to the *Fugu* genome. A similar comparative genomic approach has been carried out for the *Sox9* gene. In that case, a total of eight different conserved elements were identified around the gene locus between the human and *Fugu* sequences and between the mouse and *Fugu* sequences. These elements contained conserved consensus for known transcription factors and were presented as very strong candidates for regulatory sequences of the different expression domains of the gene (Bagheri-Fam et al. 2001).

The outcome of the human-*Fugu* alignment for the *Pax9* genomic sequence was conversely rather disappointing. Of the dense conserved element distribution observed between human and mouse, only two unique elements (CNS-6 and CNS+2) were found in the *Fugu* sequence, being the remaining homology strictly confined in the coding regions. This result could be explained with two different hypotheses. One is that fish may in some cases be phylogenetically too distant for this type of analysis. The situation of the *Sox9* gene indicates that *Fugu* is generally a useful model for comparative sequencing and other successful examples of conserved regulatory elements between mammals and fish have been reported (Aparicio et al. 1995; Rowitch et al. 1998; Zerucha et al. 2000). However, distantly related species are not always the best comparison model. The CNS-1 element located between the interleukins 4 and 13 in the human and mouse genomes and responsible for their specific expression in type 2 T-helper cells could not be clearly detected in chicken or *Fugu* (Loots et al. 2000). Likewise, not all of the human *SCL* gene enhancers could be detected by comparison with the chicken sequence, but

only with the mouse (Göttgens et al. 2000), and comparisons between mammalian and avian β-globin gene clusters failed to demonstrate any significant homology between regulatory elements, even when they were known to be functionally analogous (Hardison et al. 1997).

These observations clearly suggest that the choice of species to be compared is essential for the efficiency of the phylogenetic footprinting. If the species are too closely related, distinguishing highly constrained regulatory elements from non-functional regions will be impossible because there will not have been enough evolutionary time for the diversification of neutral sequences. But if the species are too distantly related, then detecting conserved regulatory elements may be impossible, either because they will have diverged too much to preserve any significant similarity or because the regulation processes are different in the two lineages (Duret and Bucher 1997).

5.8. The zebrafish *Pax9* expression pattern

One necessary control before starting a comparative sequencing analysis is the comparison of the gene expression pattern between the two species, because the similarity of expression domains may be an important indication (but not a definite proof) for the presence of common regulatory mechanisms. Vice versa, dissimilar expression patterns leave little expectation in finding conserved regulatory elements.

Because of the unavailability of pufferfish embryos for expression studies, zebrafish embryos were employed in this work, assuming that the expression pattern would be normally alike among teleost fish.

The zebrafish *Pax9* pattern was already partially known (Nornes et al. 1996). The herein presented results have added more information (Fig. 20). The expression in the sclerotome of the somites, detected at around 18 hrs of development, was confirmed. The disappearing of this expression in later stages (48 hrs) could not be explained. This difference might correspond to a divergence in the later developmental stages of these structures between mammals and fish and therefore to variations in the expression patterns of some genes. However, it is noteworthy that in the lamprey, which does not have a real vertebral column, no *Pax9* transcription in the somites could be seen, attesting that the expression of *Pax9* in the somites specifically appeared in association with the

emergence of the sclerotomal tissue and consequently with the formation of a vertebral column (Ogasawara et al. 2000). This expression domain has remained conserved up to the higher vertebrates (Neubüser et al. 1995; Müller et al. 1996).

In addition to that, a very strong expression in the pharyngeal region could be observed at 48 hrs, even though initial traces of the *Pax9* transcript could already be seen at 18 hrs in a corresponding area, erroneously confused with facial mesenchyme in Nornes et al. The pharyngeal endoderm is the most primitive *Pax9* expression domain. It has not only been documented in all the vertebrate species so far analyzed, but studies in lower chordates, like amphioxus and ascidians, which have as already mentioned only one member of the Pax1/9 paralogous pair, have shown that this domain predates the duplication and differentiation of the *Pax1* and *Pax9* genes (Holland et al. 1995; Ogasawara et al. 1999).

Finally, a certain positive signal was to be seen in the facial region, may be in correspondence with the position of the olfactory organs. This domain might be related to the mouse *Pax9* expression in the nasal region (nasal processes and nasal capsule) (Neubüser et al. 1995; Peters et al. 1998b). The zebrafish *Pax9* staining in the anterior edge of the olfactory organs could be recognized by comparison with the expression pattern of other genes observed in the same region, like *MsxB* (Ekker et al. 1997). In conclusion, it can be said that the *Pax9* expression pattern in the fish nicely overlaps the one observed in the mammals. From this point of view, a fish is potentially a good model for the search of conserved regulatory elements, at least for the common domains. Conversely, the fine regulation of the expression in more precise structures of the face mesenchyme or in the limbs, where no corresponding expression at all has been so far detected in the fish (Fig. 20 e), cannot be investigated with this system.

Of course, the possibility remains that the presence of an equivalent expression domain does not correspond to the presence of homologous regulatory elements. For example, the mouse *Dll1* gene and its orthologue in zebrafish *DeltaD* are both expressed in the somitic and presomitic mesoderm and in the developing neural tube (Bettenhausen et al. 1995; Haddon et al. 1998). However, while two neural tube specific enhancers were found to be conserved between the two species, no homology could be found between the identified mouse somite enhancer and the zebrafish somite enhancer (Beckers et al. 2000)

suggesting that only some regulatory elements might have been preserved both at the sequence and at the functional level in the evolution of the two lineages.

The second hypothesis to explain the scarcity of conserved non-coding sequences around the human and the *Fugu Pax9* genes is that the analyzed genomic region does not contain the control elements for the expression in the described domains. This hypothesis was sustained by the BAC transgenic experiment as discussed further on (see 5.13.).

5.9. Comparative sequencing reveals an extended conserved syntenic region

One of the most interesting things that emerged from the human/*Fugu* comparative sequencing was that most of the identified conserved elements downstream of *Pax9* turned out to coincide with the exons of the gene coding for the mitochondrial oxodicarboxylate carrier (Odc). This finding added credit to the interest for this highly conserved syntenic region, which proved to extend further at least in one direction. In fact, analysis of the mouse and human genome annotations, available from the internet (Ensembl Genome Browser http://www.ensembl.org/ and Human Genome Sequencing at NCBI http://www.ncbi.nlm.nih.gov/genome/seq/), showed that even one more gene was included. The gene *Foxa1*, coding for Hnf-3α, one of the members of the winged-helix transcription factor family, maps directly upstream of the *Odc* gene both in the mouse and in the human genomes. Up to now, the whole mouse and human conserved syntenic region includes at least five genes, *Nkx2-1*, *Nkx2-9*, *Pax9*, *Odc* and *Foxa1*, three of which (*Nkx2-9*, *Pax9* and *Odc*) were shown here to be associated also in the *Fugu* genome.

For the considerations made before about the locus duplication event that originated the *Pax9* and the *Pax1* syntenic regions, it is reasonable to think that also *Nkx2-1* is present in the same association in *Fugu*. This hypothesis was confirmed by analysis of the *Fugu* genome sequence data in the Ensembl database (see above).

This gene association was presumably at least present in the common ancestral genome and that gives reason to think that the situation has not changed up to now, as shown for the mammals and the fish. The same considerations can be probably extended to the rest of the group of genes. The next known gene downstream of *Pax1* is *Foxa2*, the paralogous counterpart of *Foxa1*, and the same gene can be again found in the human and in the mouse genomes. It appeared however that a gap was present between the two

paralogous syntenic regions, since no known gene seemed to interpose between *Pax1* and *Foxa2*, as does *Odc* between *Pax9* and *Foxa1*. A gene loss had occurred in the diversification of the two regions, an internal deletion that did not affect the flanking genes and that left a gap subsequently filled up with hundreds of kb of neutral sequence. However, this hypothesis was still not fully convincing. Looking again more carefully at the human genome annotation, it seemed that a putative gene was indeed present between *Pax1* and *Foxa2*. The gene was only suggested by exon prediction programs, which identified some putative coding sequences that did not correspond to any known gene, neither to ESTs. Surprisingly, the sequence prediction classified the gene as a putative mitochondrial carrier, exactly as *Odc*. The fact that this gene does not seem to correspond to a transcribed sequence suggests that it represents a pseudogene or more simply what is left of a gene that has lost its functional significance.

Finally, it can be concluded that in the delineation of the *Pax1*/*Pax9* genomic regions five neighboring genes have maintained their physical association conserved from a primitive situation predating the locus duplication up to the present time. Following this duplication event, four of the five paralogous pairs have diversified their roles, contributing to the variety of the gene functionality in the vertebrates. Only from the duplication of the *Odc* gene, one of the two copies has not evolved to a different active form and it has been lost through millions of years of genetic drift.

It is interesting to point out, that *Odc* is the only one of these genes that does not code for a transcription factor or for a protein with a role in body development. The oxodicarboxylate carrier performs a central role in the mitochondrial metabolism and it is a component of a biochemical process that has maintained conserved from the yeast up to the most evolved pluricellular organisms (Fiermonte et al. 2001). Thus, there was probably no need for another similar gene in such a well-established metabolic pathway and this could explain why no paralogue has evolved after the duplication event. However, even though the function of the Odc-like putative gene has not evolved, the gene has maintained its structure and location at the genomic level. The hypothesis of some biological significance connected to the intimate physical association of this group of genes builds up conspicuously.

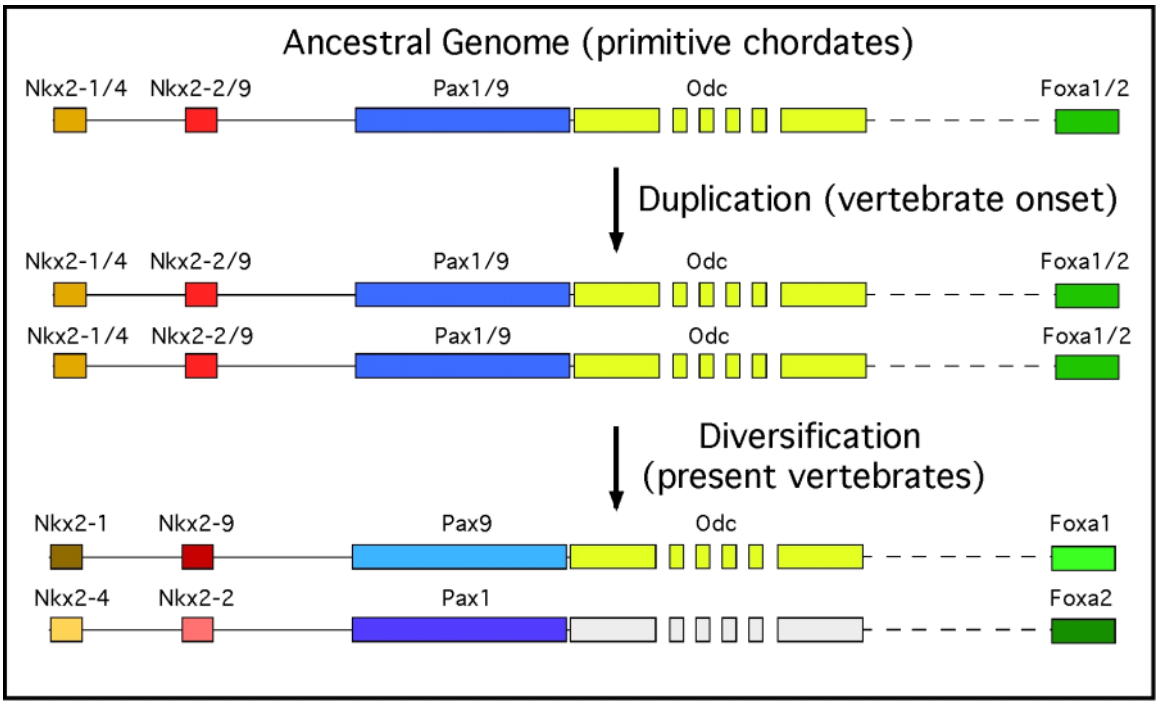


Fig. 38. Hypothesis about the origin of the Pax1/Pax9 syntenic regions. From the conserved synteny of the Pax1/Pax9 genomic regions and the genetic information about lower chordates like amphioxus, it can be assumed that in the ancestral genome of a chordate progenitor a set of at least five genes, including Pax1/9 ancestor, were tightly associated. A series of genome duplication events brought about the evolutionary burst that caused the vertebrate origin. The Pax1/9 syntenic region underwent as well duplication originating two sets of paralogous genes that afterwards independently diversified acquiring distinct functions up to the present situation. Only the Odc gene did not evolve in two different forms; one copy (associated to Pax1) became inactive. In this work it is suggested that the presence of interspersed regulatory cis-elements throughout this genomic region has represented the driving force that has kept these genes tightly associated through evolution.
The length of the Odc gene (500 kb) is not proportionally shown (broken bar), so is not the physical distance from Foxa1/2 (dashed line).

5.10. Experimental approaches for the identification of regulatory elements: cell culture versus transgenesis

Two methods were adopted in this work in order to test the functional activity of the selected conserved non-coding sequences as regulatory elements.

With the first method, it was tried to take advantage of a cell culture system, using the two *Pax9* expressing cell lines, AT478 and MLB13myc, already employed for the promoter assay. The experimental system was based on the concept that a regulatory element active in either cell line would have positively or negatively changed the *Pax9* promoter activity.

As a comparable example, it can be reported that a similar system was used to identify and dissect two enhancers for the neuroretinal specific *Pax6* P0 promoter. The two enhancers showed as well a high degree of conservation between the human and the quail sequences and when cloned in front of the basal promoter they were independently able to specifically amplify the expression of a reporter gene only in quail neuroretinal cells and not in other cell types (Plaza et al. 1999).

The same kind of approach with the *Pax9* CNSs cloned in front of the promoter B did not produce any appreciable and reproducible result. However, it has to be said that the identification of the *Pax6* enhancers resulted from an initial promoter assay with a longer genomic region, subsequently dissected into shorter functional components, and the sequence conservation was ascertained only after the experimental proof.

Conversely, the *Pax9* test fragments were simply characterized by sequence comparison and no discriminating evidence was at the basis of their selection for the functional assay. Hence, the lack of indicative experimental data could not assure whether among the tested elements there was some potential candidate, nor if the two cell lines were an appropriate system for a reliable enhancer assay.

These remarks do not intend to discredit the experimental procedure, which could have been indeed a potentially good system for the rapid identification of specific regulatory elements, rather they justify the failure of the approach in this particular case. Moreover, no similar examples could be found for comparison in the literature.

The production of transient transgenic mice was the alternative method engaged to assay for enhancer function. The advantage of the method is that the constructs can be tested under more physiological conditions than what can be reproduced in a cell culture system and that there are no cell type-specific restrictions, at least within the chosen developmental stages when the analysis is performed. On the other hand, the disadvantage of this method compared to a cell culture assay is that it is based on an extremely more complicated and expensive technique (pronuclear injection and embryo transfer) and that it is quite a long-term experiment, both due to the variable efficiency of the technique, which allows to test only one construct at the time, and to the incubation times required for the embryo production. For this reason, an exhaustive analysis of all the available CNS constructs was not possible within the terms of this work and the

experiments were focused on the two most conserved elements, CNS-6 and CNS+2, which could be found in the human, mouse and *Fugu* sequences.
Indeed, both elements turned out to have a real regulatory activity specifically driven in some target tissues.

5.11. A transient transgenic assay identifies an *Nkx2-9* neural tube enhancer

The CNS-6 element embodied the ability to direct transcription of the reporter gene mainly in the ventral half of the neural tube (Fig. 22). This pattern of expression was immediately associated to the *Nkx2-9* gene, which had been described to be expressed in the ventral domains of the neural tube and brain (Pabst et al. 1998). That the CNS-6 represented most likely an *Nkx2-9* rather than a *Pax9* regulatory sequence was somehow expected. The element was located about 1.5 kb away from the 5'-end of the gene and it enclosed a CpG island (Fig. 21). CpG islands are often involved in the regulation of a gene transcription and they are located in proximity of the transcription start site (Ioshikhes and Zhang 2000).
Interestingly, a CpG island was also detected by PIP analysis in the vicinity of the *Pax9* promoter (2.5 kb upstream of the transcription start site B, GenBank Z63201), but oppositely to *Nkx2-9* no homology was found with the mouse sequence. Moreover, no available experimental evidence proved that the putative *Pax9* CpG island performs a real functional control on the transcription of the gene. In fact, there was indication that this element does not have tissue-specific enhancer activity, unlike in the case of *Nkx2-9*.
Although the finding of the *Nkx2-9* neural enhancer did not directly relate to the *Pax9* regulation analysis, it was important for the general investigation of the intergenic region and as a demonstration of the technical validity of the experimental approach, that is identification of evolutionarily conserved elements and functional testing through transgenesis.
Nkx2-9 spatial and temporal expression pattern through development is more complicated than what was reproduced with the transgenic construct. The first expression domain of the gene at around E7.0 can be observed in the endoderm underlying the anterior neural plate and only later between E7.5 and E8.0 it is shifted to the to the floor plate region in the neuroectoderm. The expression extends along the entire neuraxis until E10.5, when

Nkx2-9 transcripts are still detected in the brain and the caudal part of the neural tube (shown also in this work by in situ hybridization – Fig. 23) and contemporaneously it moves from the floor plate to more lateral positions within the neuroectoderm (Pabst et al. 1998).

All the analyzed transgenic embryos were about 10.5/11 dpc old, hence too advanced in development to attest whether the same element CNS-6 would as well be competent to drive the expression of the gene in the early endoderm. This possibility is rather unlikely. It was generally shown that the endodermal expression of the NK2 genes depends on different Shh-independent regulatory mechanisms than in the neural tissue (Pabst et al. 2000).

The persistent detection of β-galactosidase activity even when the expression of the endogenous gene starts to fade off could be partly due to an incomplete spatio-temporal information of the transgenic construct but mostly to the relatively high stability of the *lacZ* mRNA and of the β-galactosidase protein itself, so that transgenic expression can still be detected also when transcription has ceased. In some older specimens X-Gal staining of the neurons emerging from the ventral neural tube could be observed. These neurons conceivably derived from formerly *Nkx2-9* expressing cells that still preserved residual β-galactosidase activity.

The analysis of the CNS-6 sequence could not only show the conservation degree among the three species, but it revealed a putative binding site for the Gli proteins with a 100% matching to the sequence identified in Sasaki et al. 1997. The authors originally identified an enhancer downstream of the *Hnf3-β* gene (*Foxa2*) that specifically drove the expression of a *lacZ* reporter gene in the floor plate along the whole neural tube (Sasaki and Hogan 1996). This enhancer appeared to recruit the Gli transcription factors, particularly upon Shh induction, promoting activation of a downstream reporter gene both in vivo and in a cell culture assay. This finding was supported by the fact that the Gli factors overlapped their expression with *Hnf3-β* in the ventral neural tube and floor plate, region of Shh production (Sasaki et al. 1997).

Interestingly, this expression pattern overlaps as well with that of *Nkx2-9*. Moreover, it is known that *Nkx2-9* expression in the neural tube is dependent on the Shh signaling

coming from the floor plate, since in Shh knockout mice none of the Nkx2 genes could be detected in the neural domain (Pabst et al. 2000).

These observations taken together strongly suggest that the identified sequence in the CNS-6 is the Gli-binding site through which Shh transcription regulation of *Nkx2-9* takes place. Notably, the putative Gli-binding site is conserved in the three species (Fig. 21). The sequence conservation extends far outside the consensus suggesting that most probably more factors synergistically co-operate to the gene regulation. Accordingly, also in the case of *Hnf3-β* a construct containing only multiple copies of the Gli-binding site was not sufficient to reconstitute the function of the entire enhancer in a transgenic experiment (Sasaki et al. 1997).

In addition to the expression in the ventral part of the neural tube, four out of five CNS-6 transgenic embryos also showed X-Gal staining in the dorsal area, at the level of the roof plate. This result is of difficult interpretation because it is known that Shh represses the transcription of genes normally expressed in the dorsal neural tube (Goulding et al. 1994 and reviewed in McMahon 2000). Presumably, a cryptic site for floor plate specific expression is present within the CNS-6 sequence, which is normally inactive in the endogenous gene and gets activated after extrapolation from the native genomic context.

The isolation of the putative Shh-dependent enhancer of *Nkx2-9* rules out the possibility presented before that the gene might share Shh regulatory elements with *Pax9*, since this sequence specifically drives the transgene expression only in the neural tube. Thus, the functional reason for the conserved association between the two genes has to reside in a higher scale of regulatory processes that probably operate over large genomic regions. This hypothesis is more precisely explained farther in the discussion (see 5.15.).

5.12. Identification of a *Pax9* medial nasal process enhancer

The result obtained with the CNS+2 transgene was as interesting as unexpected. This element represents the first *Pax9* regulatory sequence so far identified and this regulation seems to be particularly restricted to a very specific area.

Although it has never been described in details, *Pax9* is intimately coupled to the development of the nose, in particular expressed in the mesenchyme of neural crest cell derivation that originates from the midbrain and condenses in the facial region to form

the lateral and medial nasal processes and the maxillary processes. These structures arise originally separately and then they converge anteriorly and fuse in the midline giving rise to the nasal cavities (Kaufman 1992). *Pax9* is expressed very early at the onset of nasal placode formation in the underlying mesenchyme and as the nasal pits grow inside forming the olfactory epithelium, *Pax9* expression more and more extensively marks the surrounding nasal capsule (A. Neubüser, unpublished observations).

The finding that the element CNS+2 was able to direct the transgene expression selectively in the medial nasal processes and not in the remaining mesenchymal tissues suggests that each of the *Pax9* expressing structures is independently regulated. This result is in agreement with the outcome from the BAC transgenesis, where *lacZ* expression could be observed only in single restricted facial elements, especially in the nose mesenchyme. The regulatory mechanisms that control the whole cranio-facial development appears to be very complex and diverse, as testified by the composite distribution of signaling molecules and transcription factors that pattern the entire area. Components of all the known signaling factor types, such as Fibroblast Growth Factors (FGFs), Bone Morphogenetic Proteins (BMPs), Hedgehog and Wnt family members, as well as various transcription factors like homeobox-containing proteins (of Msx-, Dlx-, Otx-type) and paired-box gene products other than Pax9 (like Pax3, Pax6, Pax7) can be enumerated (reviewed in Francis-West et al. 1998).

It is however arduous to predict as for the CNS-6 which transcription factors possibly bind to CNS+2 or which signaling molecules trigger off the molecular cascade upstream of the pathway. The candidates are numerous and there is no significant evidence up to now that can restrict the circle.

The conservation of the CNS+2 element with the pufferfish sequence strongly suggests that the same element might be responsible for the *Pax9* expression in the ventral nasal region in the fish, as described in zebrafish for the first time in this work, and that consequently the same factors bind to this sequence and promote the transcription of the gene. Of course this possibility has to be proved, for example by likewise using the pufferfish CNS+2 sequence in a zebrafish transgenic model. The feasibility of using pufferfish sequences in zebrafish transgenesis has been recently suggested as an efficient way of mapping complex regulatory elements in the fish (Rothenberg 2001).

As during the tooth development, *Pax9* expression in the nasal processes could be similarly regulated by a competing mechanism involving Fgf and Bmp signaling. In particular Fgf8 is secreted by the ectoderm of the nasal pits and overlying the nasal mesenchyme of the medial processes that express *Pax9*. Similarly, other members of the Fgf family, Fgf9, Fgf17 and Fgf18, could be as well detected along the oral edge of the medial nasal process (Bachler and Neubüser 2001). The overlapping expression of these genes in this domain indicates some important role for this structure in future developmental steps, as for example during the fusion with the undergrowing maxillary processes. Interestingly, the involvement of the Fgf signaling from the overlaying ectoderm in the regulation of several transcription factors in the chick nasal mesenchyme (such as Tbx2, Erm, Pea3, and Pax3) has been very recently described (Firnberg and Neubüser 2002). If these molecules mediate the positioning of *Pax9*, as they do in the tooth mesenchyme (Neubüser et al. 1997), other factors must play a role in determining which cells should be competent to specifically respond to the ectodermal signaling, because *Pax9* is much more restricted than the diffuse Fgf expression patterns. This conception is supported by the identification of the CNS+2 enhancer, which obviously contains a very specific positional information. Otx2 is definitely one of the main candidates to be considered, since it is expressed in the migrating neural crest and the derived mesenchyme of the first branchial arch and frontonasal mass. In agreement with this, *Otx2* heterozygous mutants have defects in the anterior skull and distal jaws (reviewed in Kuratani et al. 1997). Various members of the aristaless-like homeobox gene family are expressed in similar patterns in neural crest-derived mesenchyme of developing craniofacial regions, in particular *Alx3* and *Alx4* expression overlaps with some *Pax9* positive domains including the medial nasal processes (Beverdam and Meijlink 2001). However, the involvement of these two genes in *Pax9* regulation seems unlikely, since recently the *Alx3*/*Alx4* double mutant mice have been shown to still express *Pax9* despite the severe malformations of the nasal structures (Beverdam et al. 2001).

Another gene known to be widely expressed in the facial mesenchyme including the nasal region is *Msx1* and its importance in the patterning of the nasal bones is confirmed by the phenotype of the corresponding knock-out mouse (Satokata and Maas 1994).

Interestingly, the zebrafish *MsxB* and *MsxC* genes display a certain overlap with the *Pax9* expression described in this work, that is in the ventral area of the olfactory placodes (Ekker et al. 1997). Taking again the tooth bud as a model, *Pax9* and *Msx1* appear rather to synergistically co-operate in the dental mesenchyme instead of being in hierarchical relationships with respect to each another (Peters and Balling 1999). A similar synergistic action might take place also in the development of the nasal capsule and nasal bone. However, despite the widespread *Pax9* expression in the nasal region, no evident defect has been described in the external nasal structures of the knock-out mice (Peters et al. 1998b) suggesting that the role of *Pax9* might be compensated by some other gene(s).

5.13. A 195-kb genomic region is not enough to fully reproduce the *Pax9* expression

The result obtained with the CNS+2 transgenic construct, showing a restricted *Pax9* expression in the ventral domain of the medial nasal processes, is in agreement with the outcome of the BAC transgenesis. The CNS+2 element was included in the sequence of the BAC17 clone used for the transgenic experiment. BAC transgenic embryos at developmental stage E11.5 showed X-Gal staining in the same region as in the CNS+2 transgenic embryos; however, the staining was very faint, indicating a very low *lacZ* expression level. The reason for this weak expressivity could not be investigated. The unavailability of a second transgenic line leaves open the possibility of a positional effect that drastically reduced the efficiency of the CNS+2 enhancer. The location of the enhancer relatively close to the 3' end of the construct (with respect to the *Pax9* orientation) made it more predisposed to interfering effects of neighboring sequences at the transgenic insertion site. The same consideration could be brought up for the remaining *Pax9* domains that could not be reproduced by the BAC17 construct, especially the rest of the nasal mesenchyme, supposing that more similar regulatory elements reside in relative vicinity. However, not even a faint *lacZ* expression could be detected in any of the missing domains.

The limited expression in the ventral nasal region of 13.5 dpc embryos could be still connected to the activity of the CNS+2 enhancer. It is unlikely that it merely reflects the presence of residual β-galactosidase in cells with earlier CNS+2 activity, because the expression level is definitely higher. The regulatory element could undergo a second

round of activation at a later stage and produce the pattern observed in the BAC transgenic embryos. Alternatively, an additional element might supervene with a different function. The *lacZ*-positive structures did not seem however to delineate a definite region, rather they appeared as part of an incomplete expression domain. A possible interpretation for this restrained expression could be that the regulatory sequences responsible for the spatial information, determining expression in the correct domains, need additional temporal information for the maintenance of the expression. If the maintenance is carried out by elements located outside the BAC clone, the result will be a drastic limitation to few positive cells that have newly started to express the gene. The same applies to the expression in the primary and secondary palate anlagen. Again the X-Gal staining in the BAC transgenic mice was much restricted compared to the normal situation, represented by the Pax9^{lacZ} mice. In particular in the secondary palate processes that protrude from the maxilla, expression could be detected only at the extreme tips, which are likely to consist of the most recently formed cells. If this restricted expression is enough to fulfill the *Pax9* function in the formation of a secondary palate, it can only be verified with the attempt to rescue the cleft palate phenotype by the generation of BAC17 transgenic mice in the *Pax9*$^{-/-}$ background (see 5.14.).

The correct expression in a particular domain can be determined by the concomitant function of separate control elements, even at long genomic distance from each other. For example the mouse *Myf5* gene expression pattern during development is very complex and it is regulated by elements scattered in a very large genomic region (up to 140 kb far away on the 5' side of the gene). Some closer elements failed to faithfully reproduce the complete expression pattern when tested in a conventional transgenic assay with 8.8 kb of upstream sequence, but a BAC transgenic analysis fulfilled all the required regulatory information, in particular adjusting the distribution and the maintenance of expression in the somites and in the branchial arches (Carvajal et al. 2001) The situation in the expression of the *Pax9* gene, in the nasal and maxillary domains, might be similar. The closer elements are not enough to generate a complete expression pattern and require the contribution of distal elements to complement their function. The specificity of the CNS+2 enhancer in tissue targeting is a clear example of how an apparently unique

expression domain can be functionally subdivided in single smaller subdomains with individual regulatory elements.

In addition, the full expression of a gene in a particular domain does not always rely only on the presence of a specific enhancer. Other elements different in structure and mode of activity, like the locus control regions (LCRs), are sometimes required to complement a regulatory function. An LCR operates in order to activate gene expression by affecting the opening state of the chromatin over a relatively large genomic region. LCRs are usually tissue-specific and in this sense they are similar to enhancers, but they are not enough to drive a gene expression in a particular domain. They are only responsible to render a region of DNA transcriptionally active enabling the therein-included promoters and enhancers. For this reason they can only work if integrated in the genome and not in cell culture assays based on transient expression of episomal constructs. The removal of these elements (for example in a transgenic experiment) can drastically reduce or totally abolish the gene expression (reviewed in Li et al. 1999).

Other than in the facial area, a regulatory subdivision of the expression domains seems to be a general rule for *Pax9*. Again the BAC construct was not able to reproduce a complete expression pattern in the limbs. At every stage of development the dorsal expression of the gene was completely missing, clearly indicating a separation of the dorsal and ventral control domains at the genomic level. This is not surprising, considering that during the establishment of the dorso-ventral axis of the developing limb buds different sets of genes are expressed. The secreted factor encoded by the gene *Wnt7a*, which is expressed in the dorsal ectoderm, is a good candidate to convey a dorsal signal, while the ventral ectoderm is characterized by the expression of the gene *En1* (reviewed in Capdevila and Izpisua Belmonte 2001). Thus, a model of two distinct molecular mechanisms for the onset of *Pax9* expression in the dorsal and ventral domains appears consequently reasonable. In accordance to that, it has been observed that mouse mutants for *Wnt7a* show ventralization of the dorsal side of the limbs associated to ectopic dorsal expression of *Pax9* (Parr and McMahon 1995).

In the tail region, the expression in the mesenchyme that will form the connective tissue around the muscles was limited to the lateral zones, leaving out the ventral domain. This indicates that two distinct elements control *Pax9* expression in the lateral and in the

ventral structures of the tail or that the complete expression can only be achieved by concerted function of separate elements without a clear territorial assignment. Also in this case, different sets of lateral and ventral signals can induce the expression of the gene. In accordance to this, it has been observed that the *Pax9* transcript can be detected already at about E10.5 in the ventral mesenchyme underneath the hindgut, while the expression in the lateral mesenchyme appeared only later starting from E12.5 (Neubüser et al. 1995 and unpublished observations). It is so far not even clear if there is a relationship between the ventral and the lateral expression and if the cells from the two domains concur to the formation of similar or distinct structures.

5.14. Transgenic rescue of palatoschisis does not rescue the *Pax9* mutant lethality

The rescue analysis conducted on the Pax9$^{-/-}$ /BAC17 transgenic mice has unveiled some interesting aspects of the *Pax9* function. First of all, it was a direct proof of the *Pax9* gene expression from the BAC transgene copy, which could be before only assessed by reporter gene expression analysis. Of course, it has to be assumed that the exogenous *Pax9* expression faithfully reproduced the *lacZ* expression. This can only be definitely ascertained by *Pax9* in situ hybridization on Pax9$^{-/-}$/BAC17 embryos, but the in situ data obtained with the *lacZ* RNA probe on the transgenic embryos, which showed a nice overlap with the X-Gal staining data, are already a good indication, because the *Pax9* and *lacZ* genes are transcribed as one common bicistronic mRNA.

Based on the expression pattern and on the rescue data, it can be deduced that the restricted *Pax9* expression in the very terminal end of the secondary palate processes is enough to fulfill the gene function. Palate development is a multistep process. In all vertebrates, the secondary palate arises as bilateral outgrowths from the maxillary processes. In birds and most reptiles, these palatal shelves grow initially horizontally, but do not fuse with each other resulting in physiological cleft palate. Mammalian palatal shelves initially grow vertically down the side of the tongue, but elevate at a precise time to a horizontal position above the dorsum of the tongue and fuse with each other to form an intact palate (Ferguson 1988).

It is not known what the exact role of *Pax9* is in the process, but it was observed that in *Pax9* mutant mice the palatal shelves are abnormally broadened and that they fail to

elevate over the tongue (Peters et al. 1998a). Palatal shelf-elevation is the result of an intrinsic shelf elevating force, chiefly generated by the progressive accumulation and hydration of hyaluronic acid in the extracellular space of the palatal process mesenchyme (Ferguson 1988). *Pax9* could be responsible for the production and/or distribution of extracellular matrix molecules. Alternatively, the abnormal morphology of the palatal shelves in the Pax9 knock-out mice suggests that the gene could regulate their growth. The failure of the elevation could be a secondary effect due to the thickening of these structures, which constraints them laterally between the tongue and the cheeks. This shaping role would fit with the strong *Pax9* expression in the growing tips of the shelves, which have to be maintained in the right size during their extension, and it would better explain the ability to rescue the mutant phenotype despite the spatially limited expression of the transgene. Of course, this hypothesis can only be proven with a direct analysis of *Pax9* function.

The most interesting aspect that emerged from the experiment was the inability to rescue or at least to reduce the lethality of the *Pax9* mutation. The transgenic newborn mice died just as soon as the *Pax9* mutants and in a similar fashion. This finding was particularly surprising, because it had been suggested that the presence of a cleft secondary palate was the main cause of death for these mice (Peters et al. 1998a). The new results partly contradict this interpretation and suggest that the cleft palate cannot account on its own for the rapid postnatal death. Other defects might contribute or even play a more significant role in the respiratory failure that characterizes the Pax9 knock-out mice.

Unfortunately, no detailed analysis of the reasons for this lethal phenotype could be carried out. Moreover, it is difficult to predict which of the other described malformations could account for it. It seems that a general problem in the respiration mechanism affects these mice. To this point, it is interesting to remind of the Pax9 expression in the bronchioli observed by Northern blot analysis (this work) and by X-Gal staining in the $Pax9^{lacZ}$ line (I. Rodrigo, personal communication). Since this expression domain was previously not known, no anatomical investigation has been conducted on the lungs of the Pax9 mutants.

Alternatively, the absence of parathyroids and ultimobranchial bodies, which regulate the calcium homeostasis through the release of parathormone and calcitonin, respectively, might impair the muscular contractions required for the active respiration.

5.15. Open questions and conclusive remarks (an evolutionary interpretation)

A BAC transgenic approach was chosen in this work as a long-range genomic system for the identification of the *Pax9* regulatory elements, after that preliminary observations indicated that the sequences responsible for the entire control of the gene transcription might span over very long genomic distances. In the light of the presented data, it must be recognized that the situation is probably more complicated than what it was believed and that not even a BAC system is sufficient to fulfill this type of investigation. Several expression domains remained excluded from the analysis, some of which are of great interest in order to elucidate the molecular mechanisms in which the gene is involved.

Unveiling the regulation of the expression in the sclerotome of the somites, for instance, would be particularly interesting in the frame of the studies focused on the molecular pathways of chondrogenesis and endochondral ossification. In particular, it would be interesting to know if the same regulatory mechanisms that drive *Pax9* expression in these structures are as well responsible for the expression of *Pax1*. It is known that both genes are dependent on Shh signaling emanating from the notochord (Koseki et al. 1993; Goulding et al. 1994; Neubüser et al. 1995; Müller et al. 1996). However, it is also known that the temporal and spatial expression of the two genes is not exactly coincident. *Pax1* is expressed earlier on shortly before de-epithelialization of the somites, and it extends along the whole rostro-caudal axis of each single somite; subsequently it concentrates in the posterior half and in the ventral domain of the sclerotome, right around the notochord. *Pax9* expression starts only later and it concerns the dorso-lateral portion of the sclerotome (Deutsch et al. 1988; Neubüser et al. 1995). However, the two genes perform concomitant functions in the development of the vertebral column in a synergistic action and the disruption of *Pax1* leads to upregulation and expansion of the *Pax9* expression (Peters et al. 1999). Is there a reciprocal regulatory mechanism at the basis of this model? It would be conceivable that an equilibrated reciprocal transcription

inhibition is the essence of this mechanism, which keeps the genes active in distinct sclerotomal subdomains and triggers the expansion of one in the absence of the other.

A comparison of the regulatory sequences of *Pax1* and *Pax9* could also help answer some evolutionary questions. For example, the expression in the pharyngeal endoderm of the *Pax1/9* gene in lower chordates (Holland et al. 1995; Ogasawara et al. 1999) is a strong hint of a common original regulatory mechanism for both genes in this tissue and therefore similarities in the respective control elements are expected. On the other hand, recent studies on lamprey (*Lampetra japonica*), an agnathan regarded as at the lowest evolutionary level of the vertebrate lineage, proposed a different model for the origin of expression in the somites. Both genes are apparently present in the genome of this organism, although only the *Pax9* cDNA could be successfully cloned. However, no *Pax9* somite expression could be detected, in accordance to the very primitively developed sclerotome of these animals (Ogasawara et al. 1999). But if the duplication of the Pax1/9 ancestor and the origin of the two paralogous genes predated the acquisition of somite specific elements, how can the similar expression pattern be explained? To this point, it has to be mentioned that a sequence alignment of the *Pax9* and *Pax1* genomic regions carried out during this work did not reveal any homology other than the paired domain, initially suggesting no similarity in the regulatory sequences. Now we can explain this result with the absence of most of the common regulatory elements within the analyzed genomic regions.

Interesting insight would come also from the identification of the elements that regulate *Pax9* expression in the tooth mesenchyme. The upstream factors of this pathway are already known. Fgf and Bmp signals determine the position of tooth bud formation and the expression of *Pax9* in the mesenchyme underneath (Neubüser et al. 1997). Finding the sequences that control this mechanism would eventually allow to identify the downstream factors of the molecular cascade and help establish new regulatory relationships among the genes known to participate in the process. Moreover, the involvement of *Pax9* in human patients with oligodontia could be ascertained by the identification of mutations in the regulatory elements instead of in the coding sequence.

More and more mutations responsible for human diseases have been mapped far beyond the transcription unit of the affected genes, suggesting the involvement of distal

regulatory elements. This is the case of the preaxial polydactily locus associated with the function of the gene *SHH*, but mapping at a physical distance of about 1 Mb (Lettice et al. 2002), or of aniridia-associated translocations, whose breakpoints map more than 150 kb distal to the affected gene *PAX6* (Kleinjan et al. 2001).

For this and other studies the possibility to create a different experimental approach remains open. If a BAC system is not enough to address the topic of an exhaustive analysis of *Pax9* regulation, may be a YAC approach should be envisaged. Of course any conclusion drawn from this analysis is inevitably shaded by the risk that the BAC transgenesis outcome was affected by some technical artifacts, above all the possibility of negative position effects in the site of integration. It is generally accepted that host sequences surrounding the place of transgene integration can modify the expected expression pattern, potentially causing it to be ectopic, weak or even undetectable and this is the main cause of variability among different specimens in conventional transgenic experiments. In this work, the lack of at least a second transgenic line did not allow to confirm, or in case to contradict the presented results, and the problem of the position effect remains unsolved. However, some arguments can be brought up in favor of a more confident interpretation of the results. The use of YACs, BACs and PACs in transgenesis is generally recognized as one of the best strategy to overcome the problem of position effect. The size of the transgenic construct can be regarded as a sort of protection against the negative influence of neighboring sequences in the site of integration and the benefits and applications of their use, in terms of optimal and reproducible gene expression level, have been reported in a constantly increasing number of cases (reviewed in Giraldo and Montoliu 2001). Hence, the utilization of a BAC construct guarantees per se a minimalization of undesired position effects, which cannot be therefore considered as a general cause for the defective expression pattern of the transgene. Moreover, considering the elevated number and diversity of missing expression domains, it is very unlikely that the presence of suppressor sequences or the general refractory structure of the integration locus affected the function of so many different regulatory elements within a genomic region of 195 kb, in particular because that would apply only to a selected subset of regulatory elements and not to all of them. The affected elements are furthermore not necessarily functionally and structurally related, thus also the vicinity of

tissue specific silencers is not a solid contrasting argument. The expression of the reporter gene in a few of the *Pax9* positive tissues can at least rule out the possibility of integration in a generally silent chromatin domain. In conclusion, it is not possible to say if some enhancers and which ones were suppressed by a position effect, but nevertheless this possibility cannot be accepted as a general interpretation for the whole outcome of the work.

A more consistent explanation implies the absence of many *Pax9* regulatory elements in the BAC clone used for the transgenic construct. In awareness of all the considerations made above, this interpretation sustains here the most accredited hypothesis about *Pax9* transcription regulation.

The discussion about the *Pax9* genomic region has brought out the strong suggestion of a biological significance for the evolutionary synteny conservation in the vertebrates and perhaps even in lower chordates. According to this hypothesis, the *Pax9* gene function appears to be strictly connected to its genomic environment. This physical association might reflect a fixed multigenic transcriptional domain whose members (at least up to five different genes) cannot be taken apart without compromising their normal function. The nature of this functional bond has still to be demonstrated, but interspersed and interdigitated regulatory elements over a widespread multigenic region can already represent a decisive factor.

A simplified example is the physical linkage between the *Myf5* and the *Mrf4* genes likewise conserved in all the vertebrate species so far analyzed. Recent studies about the distribution of the respective regulatory elements have revealed an intricate net of intermingling sequences respectively responsible for the tissue specific expressions of either gene (Carvajal et al. 2001). It is logical to deduce that this genomic organization represents an irresolvable constraint for the two genes. In addition, the regulatory elements of both *Myf5* and *Mrf4* trespass the limits of gene boundary and localize within the introns of a neighboring gene, the protein tyrosine phosphatase-RQ gene, forcing it probably into the same physical linkage.

The finding of the CNS+2 enhancer in one of the introns of the adjacent *Odc* gene is strong evidence that a similar complex situation applies also for the *Pax9* gene. The *Odc* gene might likewise host several more *Pax9* regulatory elements. The generation of a

new transgenic model using the modified BAC clone 15 as a construct, which extends 150 kb further into the *Odc* gene, might then unveil new *Pax9* elements. But, considering that *Odc* spans for about 500 kb of genomic sequence, the search for these elements overpowers a simple BAC transgenic approach and it has to be pursued with different strategies. As already mentioned, YAC transgenesis would allow much higher sequence coverage and it could be more informative. However, the subsequent restriction down to the single functional sequences would become very complicated.

At that point, comparative sequencing could turn extremely helpful for the fine mapping. Also in this work, it proved to be a very powerful method for the identification of conserved non-coding sequences with a real functional significance. It is interesting to notice that the almost total absence of conserved non-coding sequences between human/mouse and Fugu in the tested region coincided with the actual absence of the regulatory elements for the homologous expression domains, such as the somites and the pharyngeal endoderm. Accordingly, the third described common expression domain in the nasal region is probably associated to the CNS+2 element. Conversely, for the other domains reproduced by the BAC17 no homology has been observed in the *Pax9* expression between zebrafish and mouse. For example, no *Pax9* expression could be detected in the developing zebrafish fin buds. In any case, the limit of a comparative sequencing approach is that differences in gene expression patterns might lead to loss of precious information. Significantly, the regulatory elements for all the expression domains observed in the BAC transgenic embryos will not be identified by comparison with the *Fugu* sequence. Thus, a comparative sequencing analysis should not be restricted to two or three organisms. The availability of the complete genome sequence of an increasing number of species will allow to conduct an extended interspecies comparison and it will raise the chance to identify functional sequences in this type of research.

It has to be considered that searching inside the *Odc* gene introns might still not be enough for a complete analysis of the *Pax9* regulation at the genome level. The conserved syntenic locus embraces more genes that locate as well on the other side respect to *Pax9* and that similarly could retain regulatory elements in their structures. The long-range power and specificity of some cis-acting elements should not be underestimated. Recently, the genomic region containing the limb specific control

elements of the *Shh* gene has been identified about 800 kb away from the gene itself inside the introns of the *Lmbr1* gene in humans and mice. Moreover, at least one more gene has been located between *Shh* and *Lmbr1*, suggesting that these elements can exert their specific function on the target gene despite large genomic distances and the presence of other intervening genes (Lettice et al. 2002).

In conclusion, this work has represented only a first step to elucidate the terms of a research line that has revealed to be much more complex and sophisticated than what was expected. However, contemplating these preliminary results in the light of more and more similar examples in the literature (e. g. Kleinjan et al. 2002) and with a broader point of view leaves open space to some new general considerations. The presence of intersecting regulatory sequences through multigenic genomic regions has conceivably represented a key point in the genome evolution. The emergence and fixation of regulatory elements inside the territory of neighboring genes have constituted a functional bond resulting in the undisruptable physical association between the genes. These associations might have with time extended involving entire blocks of genes. Thus, it logically followed that the plasticity of the genome in the events of shuffling and reorganization occurred in the course of millions of years of evolution has been inevitably limited. The rearrangement units have not been the single genes but the blocks of physically linked genes. Hence in the future, the search for *cis*-regulatory elements will not only lead to a better understanding of the molecular mechanisms that control gene expression but it will also provide more insight into genome evolution.

The availability of the complete human and mouse genome sequences and soon of many other species, including, lower chordates, lamprey, zebrafish, chick, *Xenopus*, and primates, will allow in the next years a much broader comparative analysis of the relationship between genomic organization and gene regulation mechanisms. New computational approaches will be then indispensable for the establishment of comprehensive information out of a large amount of high-throughput measurements and data coming from this kind of analysis. Systems biology is a modern discipline that will enable us to create models of complex networks, including gene regulatory systems, out of experimental data of single components (Kitano 2002). This will also include modeling of genome evolution based not only on the phylogenetical relations between

species but also on the localization and distribution of the genes and their *cis*-regulatory elements and on the functional networks, in which they interplay.

6. BIBLIOGRAPHY

Aparicio, S., A. Morrison, A. Gould, J. Gilthorpe, C. Chaudhuri, P. Rigby, R. Krumlauf and S. Brenner (1995). "Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes." *Proc Natl Acad Sci U S A* **92**(5): 1684-8.

Apergis, G. A., N. Crawford, D. Ghosh, C. M. Steppan, W. R. Vorachek, P. Wen and J. Locker (1998). "A novel nk-2-related transcription factor associated with human fetal liver and hepatocellular carcinoma." *J Biol Chem* **273**(5): 2917-25.

Ashery-Padan, R. and P. Gruss (2001). "Pax6 lights-up the way for eye development." *Curr Opin Cell Biol* **13**(6): 706-14.

Bachler, M. and A. Neubüser (2001). "Expression of members of the Fgf family and their receptors during midfacial development." *Mech Dev* **100**(2): 313-6.

Bagheri-Fam, S., C. Ferraz, J. Demaille, G. Scherer and D. Pfeifer (2001). "Comparative genomics of the SOX9 region in human and Fugu rubripes: conservation of short regulatory sequence elements within large intergenic regions." *Genomics* **78**(1-2): 73-82.

Balczarek, K. A., Z. C. Lai and S. Kumar (1997). "Evolution of functional diversification of the paired box (Pax) DNA- binding domains." *Mol Biol Evol* **14**(8): 829-42.

Balling, R., U. Deutsch and P. Gruss (1988). "undulated, a mutation affecting the development of the mouse skeleton, has a point mutation in the paired box of Pax 1." *Cell* **55**(3): 531-5.

Beckers, J., A. Caron, M. Hrabe de Angelis, S. Hans, J. A. Campos-Ortega and A. Gossler (2000). "Distinct regulatory elements direct delta1 expression in the nervous system and paraxial mesoderm of transgenic mice." *Mech Dev* **95**(1-2): 23-34.

Bettenhausen, B., M. Hrabe de Angelis, D. Simon, J. L. Guenet and A. Gossler (1995). "Transient and restricted expression during mouse embryogenesis of Dll1, a murine gene closely related to Drosophila Delta." *Development* **121**(8): 2407-18.

Beverdam, A., A. Brouwer, M. Reijnen, J. Korving and F. Meijlink (2001). "Severe nasal clefting and abnormal embryonic apoptosis in Alx3/Alx4 double mutant mice." *Development* **128**(20): 3975-86.

Beverdam, A. and F. Meijlink (2001). "Expression patterns of group-I aristaless-related genes during craniofacial and limb development." *Mech Dev* **107**(1-2): 163-7.

Birnboim, H. C. and J. Doly (1979). "A rapid alkaline extraction procedure for screening recombinant plasmid DNA." *Nucleic Acids Res* **7**(6): 1513-23.

Bono, H., T. Kasukawa, M. Furuno, Y. Hayashizaki and Y. Okazaki (2002). "FANTOM DB: database of Functional Annotation of RIKEN Mouse cDNA Clones." *Nucleic Acids Res* **30**(1): 116-8.

Bopp, D., M. Burri, S. Baumgartner, G. Frigerio and M. Noll (1986). "Conservation of a large protein domain in the segmentation gene paired and in functionally related genes of Drosophila." *Cell* **47**(6): 1033-40.

Bopp, D., E. Jamet, S. Baumgartner, M. Burri and M. Noll (1989). "Isolation of two tissue-specific Drosophila paired box genes, Pox meso and Pox neuro." *Embo J* **8**(11): 3447-57.

Breitling, R. and J. K. Gerber (2000). "Origin of the paired domain." *Dev Genes Evol* **210**(12): 644-50.

Brunner, B., T. Todt, S. Lenzner, K. Stout, U. Schulz, H. H. Ropers and V. M. Kalscheuer (1999). "Genomic structure and comparative analysis of nine Fugu genes: conservation of synteny with human chromosome Xp22.2-p22.1." *Genome Res* **9**(5): 437-48.

Burri, M., Y. Tromvoukis, D. Bopp, G. Frigerio and M. Noll (1989). "Conservation of the paired domain in metazoans and its structure in three isolated human genes." *Embo J* **8**(4): 1183-90.

Capdevila, J. and J. C. Izpisua Belmonte (2001). "Patterning mechanisms controlling vertebrate limb development." *Annu Rev Cell Dev Biol* **17**: 87-132.

Carvajal, J. J., D. Cox, D. Summerbell and P. W. Rigby (2001). "A BAC transgenic analysis of the Mrf4/Myf5 locus reveals interdigitated elements that control activation and maintenance of gene expression during muscle development." *Development* **128**(10): 1857-68.

Chen, Y., Y. Zhang, T. X. Jiang, A. J. Barlow, T. R. St Amand, Y. Hu, S. Heaney, P. Francis-West, C. M. Chuong and R. Maas (2000). "Conservation of early odontogenic signaling pathways in Aves." *Proc Natl Acad Sci U S A* **97**(18): 10044-9.

Chi, N. and J. A. Epstein (2002). "Getting your Pax straight: Pax proteins in development and disease." *Trends Genet* **18**(1): 41-7.

Chiu, I. M., K. Touhalisky and C. Baran (2001). "Multiple controlling mechanisms of FGF1 gene expression through multiple tissue-specific promoters." *Prog Nucleic Acid Res Mol Biol* **70**: 155-74.

Corbo, J. C., A. Di Gregorio and M. Levine (2001). "The ascidian as a model organism in developmental and evolutionary biology." *Cell* **106**(5): 535-8.

Czerny, T., G. Schaffner and M. Busslinger (1993). "DNA sequence recognition by Pax proteins: bipartite structure of the paired domain and its binding site." *Genes Dev* **7**(10): 2048-61.

Dahl, E., H. Koseki and R. Balling (1997). "Pax genes and organogenesis." *Bioessays* **19**(9): 755-65.

Damante, G., G. Tell and R. Di Lauro (2001). "A unique combination of transcription factors controls differentiation of thyroid cells." *Prog Nucleic Acid Res Mol Biol* **66**: 307-56.

Das, P., D. W. Stockton, C. Bauer, L. G. Shaffer, R. N. D'Souza, T. Wright and P. I. Patel (2002). "Haploinsufficiency of PAX9 is associated with autosomal dominant hypodontia." *Hum Genet* **110**(4): 371-6.

Dassule, H. R., P. Lewis, M. Bei, R. Maas and A. P. McMahon (2000). "Sonic hedgehog regulates growth and morphogenesis of the tooth." *Development* **127**(22): 4775-85.

DeBry, R. W. and M. F. Seldin (1996). "Human/mouse homology relationships." *Genomics* **33**(3): 337-51.

Deutsch, U., G. R. Dressler and P. Gruss (1988). "Pax 1, a member of a paired box homologous murine gene family, is expressed in segmented structures during development." *Cell* **53**(4): 617-25.

Dohrmann, C., P. Gruss and L. Lemaire (2000). "Pax genes and the differentiation of hormone-producing endocrine cells in the pancreas." *Mech Dev* **92**(1): 47-54

Dorsett, D. (1999). "Distant liaisons: long-range enhancer-promoter interactions in Drosophila." *Curr Opin Genet Dev* **9**(5): 505-14.

Dressler, G. R., U. Deutsch, R. Balling, D. Simon, J. L. Guenet and P. Gruss (1988). "Murine genes with homology to *Drosophila* segmentation genes." *Development* **Supplement 104**: 181-186.

Dressler, G. R. and E. C. Douglass (1992). "Pax-2 is a DNA-binding protein expressed in embryonic kidney and Wilms tumor." *Proc Natl Acad Sci U S A* **89**(4): 1179-83.

Dressler, G. R. and P. Gruss (1988). "Do multigene families regulate vertebrate development?" *Trends Genet* **4**(8): 214-9.

Dressler, G. R. and A. S. Woolf (1999). "Pax2 in development and renal disease." *Int J Dev Biol* **43**(5): 463-8.

Duboule, D. (1998). "Vertebrate hox gene regulation: clustering and/or colinearity?" *Curr Opin Genet Dev* **8**(5): 514-8.

Dunn, L. C., S. Glücksohn-Schönheimer and V. Bryson (1940). "A new mutation in the mouse affecting spinal column and urogenital system." *Journal of Hereditary* **31**: 343-348.

Duret, L. and P. Bucher (1997). "Searching for regulatory elements in human noncoding sequences." *Curr Opin Struct Biol* **7**(3): 399-406.

Ekker, M., M. A. Akimenko, M. L. Allende, R. Smith, G. Drouin, R. M. Langille, E. S. Weinberg and M. Westerfield (1997). "Relationships among msx gene structure and function in zebrafish and other vertebrates." *Mol Biol Evol* **14**(10): 1008-22.

Elgar, G., M. S. Clark, S. Meek, S. Smith, S. Warner, Y. J. Edwards, N. Bouchireb, A. Cottage, G. S. Yeo, Y. Umrania, G. Williams and S. Brenner (1999). "Generation and analysis of 25 Mb of genomic DNA from the pufferfish Fugu rubripes by sequence scanning." *Genome Res* **9**(10): 960-71.

Elgar, G., R. Sandford, S. Aparicio, A. Macrae, B. Venkatesh and S. Brenner (1996). "Small is beautiful: comparative genomics with the pufferfish (Fugu rubripes)." *Trends Genet* **12**(4): 145-50.

Ferguson, M. W. (1988). "Palate development." *Development* **103**(Suppl): 41-60.

Fiermonte, G., V. Dolce, L. Palmieri, M. Ventura, M. J. Runswick, F. Palmieri and J. E. Walker (2001). "Identification of the human mitochondrial oxodicarboxylate carrier. Bacterial expression, reconstitution, functional characterization, tissue distribution, and chromosomal location." *J Biol Chem* **276**(11): 8225-30.

Firnberg, N. and A. Neubüser (2002). "FGF signaling regulates expression of *Tbx2*, *Erm*, *Pea3*, and *Pax3* in the early nasal region." *Dev Biol* **247**: 237-250.

Francis-West, P., R. Ladher, A. Barlow and A. Graveson (1998). "Signalling interactions during facial development." *Mech Dev* **75**(1-2): 3-28.

Frazier-Bowers, S. A., D. C. Guo, A. Cavender, L. Xue, B. Evans, T. King, D. Milewicz and R. N. D'Souza (2002). "A novel mutation in human PAX9 causes molar oligodontia." *J Dent Res* **81**(2): 129-33.

Frigerio, G., M. Burri, D. Bopp, S. Baumgartner and M. Noll (1986). "Structure of the segmentation gene paired and the Drosophila PRD gene set as part of a gene network." *Cell* **47**(5): 735-46.

Galili, N., R. J. Davis, W. J. Fredericks, S. Mukhopadhyay, F. J. Rauscher, 3rd, B. S. Emanuel, G. Rovera and F. G. Barr (1993). "Fusion of a fork head domain gene to PAX3 in the solid tumour alveolar rhabdomyosarcoma." *Nat Genet* **5**(3): 230-5.

Galliot, B. and D. Miller (2000). "Origin of anterior patterning. How old is our head?" *Trends Genet* **16**(1): 1-5.

Giraldo, P. and L. Montoliu (2001). "Size matters: use of YACs, BACs and PACs in transgenic animals." *Transgenic Res* **10**(2): 83-103.

Glardon, S., P. Callaerts, G. Halder and W. J. Gehring (1997). "Conservation of Pax-6 in a lower chordate, the ascidian Phallusia mammillata." *Development* **124**(4): 817-25.

Glardon, S., L. Z. Holland, W. J. Gehring and N. D. Holland (1998). "Isolation and developmental expression of the amphioxus Pax-6 gene (AmphiPax-6): insights into eye and photoreceptor evolution." *Development* **125**(14): 2701-10.

Gossler, A. and M. Hrabe de Angelis (1998). "Somitogenesis." *Curr Top Dev Biol* **38**: 225-87.

Göttgens, B., L. M. Barton, J. G. Gilbert, A. J. Bench, M. J. Sanchez, S. Bahn, S. Mistry, D. Grafham, A. McMurray, M. Vaudin, E. Amaya, D. R. Bentley, A. R. Green and A. M. Sinclair (2000). "Analysis of vertebrate SCL loci identifies conserved enhancers." *Nat Biotechnol* **18**(2): 181-6.

Goulding, M., A. Lumsden and A. J. Paquette (1994). "Regulation of Pax-3 expression in the dermomyotome and its role in muscle development." *Development* **120**(4): 957-71.

Grzybowska, E. A., A. Wilczynska and J. A. Siedlecki (2001). "Regulatory functions of 3'UTRs." *Biochem Biophys Res Commun* **288**(2): 291-5.

Guttenberger, R., J. Kummermehr and D. Chmelevsky (1990). "Kinetics of recovery from sublethal radiation damage in four murine tumors." *Radiother Oncol* **18**(1): 79-88.

Haddon, C., L. Smithers, S. Schneider-Maunoury, T. Coche, D. Henrique and J. Lewis (1998). "Multiple delta genes and lateral inhibition in zebrafish primary neurogenesis." *Development* **125**(3): 359-70.

Harafuji, N., D. N. Keys and M. Levine (2002). "Genome-wide identification of tissue-specific enhancers in the Ciona tadpole." *Proc Natl Acad Sci U S A* **99**(10): 6802-5.

Hardison, R., J. L. Slightom, D. L. Gumucio, M. Goodman, N. Stojanovic and W. Miller (1997). "Locus control regions of mammalian beta-globin gene clusters: combining phylogenetic analyses and experimental results to gain functional insights." *Gene* **205**(1-2): 73-94.

Hardison, R. C. (2000). "Conserved noncoding sequences are reliable guides to regulatory elements." *Trends Genet* **16**(9): 369-72.

Harvey, R. P. (1996). "NK-2 homeobox genes and heart development." *Dev Biol* **178**(2): 203-16.

Helwig, U., K. Imai, W. Schmahl, B. E. Thomas, D. S. Varnum, J. H. Nadeau and R. Balling (1995). "Interaction between undulated and Patch leads to an extreme form of spina bifida in double-mutant mice." *Nat Genet* **11**(1): 60-3.

Hetzer-Egger, C., M. Schorpp and T. Boehm (2000). "Evolutionary conservation of gene structures of the Pax1/9 gene family." *Biochim Biophys Acta* **1492**(2-3): 517-21.

Hetzer-Egger, C., M. Schorpp, A. Haas-Assenbaum, R. Balling, H. Peters and T. Boehm (2002). "Thymopoiesis requires Pax9 function in thymic epithelial cells." *Eur J Immunol* **32**(4): 1175-81.

Hobert, O. and G. Ruvkun (1999). "Pax genes in Caenorhabditis elegans: a new twist." *Trends Genet* **15**(6): 214-6.

Hoch, M. and H. Jackle (1993). "Transcriptional regulation and spatial patterning in Drosophila." *Curr Opin Genet Dev* **3**(4): 566-73.

Hol, F. A., M. P. Geurds, S. Chatkupt, Y. Y. Shugart, R. Balling, C. T. Schrander-Stumpel, W. G. Johnson, B. C. Hamel and E. C. Mariman (1996). "PAX genes and human neural tube defects: an amino acid substitution in PAX1 in a patient with spina bifida." *J Med Genet* **33**(8): 655-60.

Holland, L. Z., M. Schubert, Z. Kozmik and N. D. Holland (1999). "AmphiPax3/7, an amphioxus paired box gene: insights into chordate myogenesis, neurogenesis, and the possible evolutionary precursor of definitive vertebrate neural crest." *Evol Dev* **1**(3): 153-65.

Holland, L. Z., T. V. Venkatesh, A. Gorlin, R. Bodmer and N. D. Holland (1998). "Characterization and developmental expression of AmphiNk2-2, an NK2 class homeobox gene from Amphioxus. (Phylum Chordata; Subphylum Cephalochordata)." *Dev Genes Evol* **208**(2): 100-5.

Holland, N. D., L. Z. Holland and Z. Kozmik (1995). "An amphioxus Pax gene, AmphiPax-1, expressed in embryonic endoderm, but not in mesoderm: implications for the evolution of class I paired box genes." *Mol Mar Biol Biotechnol* **4**(3): 206-14.

Holland, P. W., J. Garcia-Fernandez, N. A. Williams and A. Sidow (1994). "Gene duplications and the origins of vertebrate development." *Dev Suppl*: 125-33.

Hoshiyama, D., H. Suga, N. Iwabe, M. Koyanagi, N. Nikoh, K. Kuma, F. Matsuda, T. Honjo and T. Miyata (1998). "Sponge Pax cDNA related to Pax-2/5/8 and ancient gene duplications in the Pax family." *J Mol Evol* **47**(6): 640-8.

Ince, T. A. and K. W. Scotto (1995). "A conserved downstream element defines a new class of RNA polymerase II promoters." *J Biol Chem* **270**(51): 30249-52.

Ioshikhes, I. P. and M. Q. Zhang (2000). "Large-scale human promoter mapping using CpG islands." *Nat Genet* **26**(1): 61-3.

Ishihara, K., N. Hatano, H. Furuumi, R. Kato, T. Iwaki, K. Miura, Y. Jinno and H. Sasaki (2000). "Comparative genomic sequencing identifies novel tissue-specific enhancers and sequence elements for methylation-sensitive factors implicated in Igf2/H19 imprinting." *Genome Res* **10**(5): 664-71.

Jessen, J. R., A. Meng, R. J. McFarlane, B. H. Paw, L. I. Zon, G. R. Smith and S. Lin (1998). "Modification of bacterial artificial chromosomes through chi-stimulated homologous recombination and its application in zebrafish transgenesis." *Proc Natl Acad Sci U S A* **95**(9): 5121-6.

Kanai, Y. and P. Koopman (1999). "Structural and functional characterization of the mouse Sox9 promoter: implications for campomelic dysplasia." *Hum Mol Genet* **8**(4): 691-6.

Kaufman, M. H. (1992). The atlas of mouse development. London.

Kawai, J., A. Shinagawa, K. Shibata, M. Yoshino, M. Itoh, Y. Ishii, T. Arakawa, A. Hara, Y. Fukunishi, H. Konno, J. Adachi, S. Fukuda, K. Aizawa, M. Izawa, K. Nishi, H. Kiyosawa, S. Kondo, I. Yamanaka, T. Saito, Y. Okazaki, T. Gojobori, H. Bono, T. Kasukawa, R. Saito, K. Kadota, H. Matsuda, M. Ashburner, S. Batalov, T. Casavant, W. Fleischmann, T. Gaasterland, C. Gissi, B. King, H. Kochiwa, P. Kuehl, S. Lewis, Y. Matsuo, I. Nikaido, G. Pesole, J. Quackenbush, L. M. Schriml, F. Staubli, R. Suzuki, M. Tomita, L. Wagner, T. Washio, K. Sakai, T. Okido, M. Furuno, H. Aono, R. Baldarelli,

G. Barsh, J. Blake, D. Boffelli, N. Bojunga, P. Carninci, M. F. de Bonaldo, M. J. Brownstein, C. Bult, C. Fletcher, M. Fujita, M. Gariboldi, S. Gustincich, D. Hill, M. Hofmann, D. A. Hume, M. Kamiya, N. H. Lee, P. Lyons, L. Marchionni, J. Mashima, J. Mazzarelli, P. Mombaerts, P. Nordone, B. Ring, M. Ringwald, I. Rodriguez, N. Sakamoto, H. Sasaki, K. Sato, C. Schonbach, T. Seya, Y. Shibata, K. F. Storch, H. Suzuki, K. Toyo-oka, K. H. Wang, C. Weitz, C. Whittaker, L. Wilming, A. Wynshaw-Boris, K. Yoshida, Y. Hasegawa, H. Kawaji, S. Kohtsuki and Y. Hayashizaki (2001). "Functional annotation of a full-length mouse cDNA collection." *Nature* **409**(6821): 685-90.

Kilchherr, F., S. Baumgartner, D. Bopp, E. Frei and M. Noll (1986). "Isolation of the *paired* gene of *Drosophila* and its spatial expression during early embriogenesis." *Nature* **321**: 493-499.

Kitano, H. (2002). "Systems biology: a brief overview." *Science* **295**(5560): 1662-4.

Kleinjan, D. A., A. Seawright, G. Elgar and V. van Heyningen (2002). "Characterization of a novel gene adjacent to PAX6, revealing synteny conservation with functional significance." *Mamm Genome* **13**(2): 102-7.

Kleinjan, D. A., A. Seawright, A. Schedl, R. A. Quinlan, S. Danes and V. van Heyningen (2001). "Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6." *Hum Mol Genet* **10**(19): 2049-59.

Kokubu, C., B. Wilm, T. Kokubu, I. Rodrigo, N. Sakai, F. Santagati, M. Wahl, M. Suzuki, K. Yamamura, K. Abe and K. Imai (2002). "Undulated short-tail deletion mutation ablates *Pax1* and leads to ectopic expression of its neighboring gene *Nkx2-2* in domains that normally express *Pax1*." *Submitted*.

Koop, B. F. (1995). "Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution." *Trends Genet* **11**(9): 367-71.

Koop, B. F. and J. H. Nadeau (1996). "Pufferfish and new paradigm for comparative genome analysis." *Proc Natl Acad Sci U S A* **93**(4): 1363-5.

Kornberg, T. B. and T. Tabata (1993). "Segmentation of the Drosophila embryo." *Curr Opin Genet Dev* **3**(4): 585-94.

Koseki, H., J. Wallin, J. Wilting, Y. Mizutani, A. Kispert, C. Ebensperger, B. G. Herrmann, B. Christ and R. Balling (1993). "A role for Pax-1 as a mediator of notochordal signals during the dorsoventral specification of vertebrae." *Development* **119**(3): 649-60.

Kothary, R., S. Clapoff, S. Darling, M. D. Perry, L. A. Moran and J. Rossant (1989). "Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice." *Development* **105**(4): 707-14.

Kozmik, Z., U. Sure, D. Ruedi, M. Busslinger and A. Aguzzi (1995). "Deregulated expression of PAX5 in medulloblastoma." *Proc Natl Acad Sci U S A* **92**(12): 5709-13.

Krelova, J., L. Z. Holland, M. Schubert, C. Burgtorf, V. Benes and Z. Kozmik (2002). "Functional equivalency of amphioxus and vertebrate Pax258 transcription factors suggests that the activation of mid-hindbrain specific genes in vertebrates occurs via the recruitment of Pax regulatory elements." *Gene* **282**(1-2): 143-50.

Kroll, T. G., P. Sarraf, L. Pecciarini, C. J. Chen, E. Mueller, B. M. Spiegelman and J. A. Fletcher (2000). "PAX8-PPARgamma1 fusion oncogene in human thyroid carcinoma [corrected]." *Science* **289**(5483): 1357-60.

Krumlauf, R. (1994). "Hox genes in vertebrate development." *Cell* **78**(2): 191-201.

Kuratani, S., I. Matsuo and S. Aizawa (1997). "Developmental patterning and evolution of the mammalian viscerocranium: genetic insights into comparative morphology." *Dev Dyn* **209**(2): 139-55.

Kuschert, S., D. H. Rowitch, B. Haenig, A. P. McMahon and A. Kispert (2001). "Characterization of Pax-2 regulatory sequences that direct transgene expression in the Wolffian duct and its derivatives." *Dev Biol* **229**(1): 128-40.

Kwan, C. T., S. L. Tsang, R. Krumlauf and M. H. Sham (2001). "Regulatory analysis of the mouse Hoxb3 gene: multiple elements work in concert to direct temporal and spatial patterns of expression." *Dev Biol* **232**(1): 176-90.

Lalioti, M. and J. Heath (2001). "A new method for generating point mutations in bacterial artificial chromosomes by homologous recombination in Escherichia coli." *Nucleic Acids Res* **29**(3): E14.

Leitges, M., L. Neidhardt, B. Haenig, B. G. Herrmann and A. Kispert (2000). "The paired homeobox gene Uncx4.1 specifies pedicles, transverse processes and proximal ribs of the vertebral column." *Development* **127**(11): 2259-67.

Lettice, L. A., T. Horikoshi, S. J. Heaney, M. J. Van Baren, H. C. Van Der Linde, G. J. Breedveld, M. Joosse, N. Akarsu, B. A. Oostra, N. Endo, M. Shibata, M. Suzuki, E. Takahashi, T. Shinka, Y. Nakahori, D. Ayusawa, K. Nakabayashi, S. W. Scherer, P. Heutink, R. E. Hill and S. Noji (2002). "Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly." *Proc Natl Acad Sci U S A* **99**(11): 7548-53.

Li, Q., S. Harju and K. R. Peterson (1999). "Locus control regions: coming of age at a decade plus." *Trends Genet* **15**(10): 403-8.

Logan, C., W. K. Khoo, D. Cado and A. L. Joyner (1993). "Two enhancer regions in the mouse En-2 locus direct expression to the mid/hindbrain region and mandibular myoblasts." *Development* **117**(3): 905-16.

Loots, G. G., R. M. Locksley, C. M. Blankespoor, Z. E. Wang, W. Miller, E. M. Rubin and K. A. Frazer (2000). "Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons." *Science* **288**(5463): 136-40.

MacKenzie, A., L. Purdie, D. Davidson, M. Collinson and R. E. Hill (1997). "Two enhancer domains control early aspects of the complex expression pattern of Msx1." *Mech Dev* **62**(1): 29-40.

Mandler, M. and A. Neubüser (2001). "FGF signaling is necessary for the specification of the odontogenic mesenchyme." *Dev Biol* **240**(2): 548-59.

Mansouri, A. (1998). "The role of Pax3 and Pax7 in development and cancer." *Crit Rev Oncog* **9**(2): 141-9.

Mansouri, A., G. Goudreau and P. Gruss (1999). "Pax genes and their role in organogenesis." *Cancer Res* **59**(7 Suppl): 1707s-1709s; discussion 1709s-1710s.

Mansouri, A., A. K. Voss, T. Thomas, Y. Yokota and P. Gruss (2000). "Uncx4.1 is required for the formation of the pedicles and proximal ribs and acts upstream of Pax9." *Development* **127**(11): 2251-8.

McLysaght, A., A. J. Enright, L. Skrabanek and K. H. Wolfe (2000). "Estimation of synteny conservation and genome compaction between pufferfish (Fugu) and human." *Yeast* **17**(1): 22-36.

McMahon, A. P. (2000). "Neural patterning: the role of Nkx genes in the ventral spinal cord." *Genes Dev* **14**(18): 2261-4.

Miller, D. J., D. C. Hayward, J. S. Reece-Hoyes, I. Scholten, J. Catmull, W. J. Gehring, P. Callaerts, J. E. Larsen and E. E. Ball (2000). "Pax gene diversity in the basal cnidarian Acropora millepora (Cnidaria, Anthozoa): implications for the evolution of the Pax gene family." *Proc Natl Acad Sci U S A* **97**(9): 4475-80.

Montero, J. A., Y. Ganan, D. Macias, J. Rodriguez-Leon, J. J. Sanz-Ezquerro, R. Merino, J. Chimal-Monroy, M. A. Nieto and J. M. Hurle (2001). "Role of FGFs in the control of programmed cell death during limb development." *Development* **128**(11): 2075-84.

Montpetit, A. and D. Sinnett (2001). "Comparative analysis of the ETV6 gene in vertebrate genomes from pufferfish to human." *Oncogene* **20**(26): 3437-42.

Morishita, M., T. Kishino, K. Furukawa, A. Yonekura, Y. Miyazaki, T. Kanematsu, S. Yamashita and T. Tsukazaki (2001). "A 30-base-pair element in the first intron of SOX9 acts as an enhancer in ATDC5." *Biochem Biophys Res Commun* **288**(2): 347-55.

Müller, T. S., C. Ebensperger, A. Neubüser, H. Koseki, R. Balling, B. Christ and J. Wilting (1996). "Expression of avian Pax1 and Pax9 is intrinsically regulated in the pharyngeal endoderm, but depends on environmental influences in the paraxial mesoderm." *Dev Biol* **178**(2): 403-17.

Muyrers, J. P., Y. Zhang, G. Testa and A. F. Stewart (1999). "Rapid modification of bacterial artificial chromosomes by ET- recombination." *Nucleic Acids Res* **27**(6): 1555-7.

Nadeau, J. H. and D. Sankoff (1998). "The lengths of undiscovered conserved segments in comparative maps." *Mamm Genome* **9**(6): 491-5.

Narayanan, K., R. Williamson, Y. Zhang, A. F. Stewart and P. A. Ioannou (1999). "Efficient and precise engineering of a 200 kb beta-globin human/bacterial artificial chromosome in E. coli DH10B using an inducible homologous recombination system." *Gene Ther* **6**(3): 442-7.

Neubüser, A., H. Koseki and R. Balling (1995). "Characterization and developmental expression of Pax9, a paired-box- containing gene related to Pax1." *Dev Biol* **170**(2): 701-16.

Neubüser, A., H. Peters, R. Balling and G. R. Martin (1997). "Antagonistic interactions between FGF and BMP signaling pathways: a mechanism for positioning the sites of tooth formation." *Cell* **90**(2): 247-55.

Nieminen, P., S. Arte, D. Tanner, L. Paulin, S. Alaluusua, I. Thesleff and S. Pirinen (2001). "Identification of a nonsense mutation in the PAX9 gene in molar oligodontia." *Eur J Hum Genet* **9**(10): 743-6.

Noll, M. (1993). "Evolution and role of Pax genes." *Curr Opin Genet Dev* **3**(4): 595-605.

Nornes, S., I. Mikkola, S. Krauss, M. Delghandi, M. Perander and T. Johansen (1996). "Zebrafish Pax9 encodes two proteins with distinct C-terminal transactivating domains of different potency negatively regulated by adjacent N-terminal sequences." *J Biol Chem* **271**(43): 26914-23.

Nüsslein-Volhard, C. (1991). "Determination of the embryonic axes of Drosophila." *Dev Suppl* **1**: 1-10.

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

Nutt, S. L., D. Eberhard, M. Horcher, A. G. Rolink and M. Busslinger (2001). "Pax5 determines the identity of B cells from the beginning to the end of B-lymphopoiesis." *Int Rev Immunol* **20**(1): 65-82

Ogasawara, M., Y. Shigetani, S. Hirano, N. Satoh and S. Kuratani (2000). "Pax1/Pax9-Related genes in an agnathan vertebrate, Lampetra japonica: expression pattern of LjPax9 implies sequential evolutionary events toward the gnathostome body plan." *Dev Biol* **223**(2): 399-410.

Ogasawara, M., H. Wada, H. Peters and N. Satoh (1999). "Developmental expression of Pax1/9 genes in urochordate and hemichordate gills: insight into function and evolution of the pharyngeal epithelium." *Development* **126**(11): 2539-50.

Pabst, O., H. Herbrand and H. H. Arnold (1998). "Nkx2-9 is a novel homeobox transcription factor which demarcates ventral domains in the developing mouse CNS." *Mech Dev* **73**(1): 85-93.

Pabst, O., H. Herbrand, N. Takuma and H. H. Arnold (2000). "NKX2 gene expression in neuroectoderm but not in mesendodermally derived structures depends on sonic hedgehog in mouse embryos." *Dev Genes Evol* **210**(1): 47-50.

Parr, B. A. and A. P. McMahon (1995). "Dorsalizing signal Wnt-7a required for normal polarity of D-V and A-P axes of mouse limb." *Nature* **374**(6520): 350-3.

Peters, H. and R. Balling (1999). "Teeth. Where and how to make them." *Trends Genet* **15**(2): 59-65.

Peters, H., U. Doll and J. Niessing (1995). "Differential expression of the chicken Pax-1 and Pax-9 gene: in situ hybridization and immunohistochemical analysis." *Dev Dyn* **203**(1): 1-16.

Peters, H., A. Neubüser and R. Balling (1998a). "Pax genes and organogenesis: Pax9 meets tooth development." *Eur J Oral Sci* **106 Suppl 1**: 38-43.

Peters, H., A. Neubüser, K. Kratochwil and R. Balling (1998b). "Pax9-deficient mice lack pharyngeal pouch derivatives and teeth and exhibit craniofacial and limb abnormalities." *Genes Dev* **12**(17): 2735-47.

Peters, H., G. Schuster, A. Neubüser, T. Richter, H. Hofler and R. Balling (1997). "Isolation of the Pax9 cDNA from adult human esophagus." *Mamm Genome* **8**(1): 62-4.

Peters, H., B. Wilm, N. Sakai, K. Imai, R. Maas and R. Balling (1999). "Pax1 and Pax9 synergistically regulate vertebral column development." *Development* **126**(23): 5399-408.

Plaza, S., C. Dozier, N. Turque and S. Saule (1995). "Quail Pax-6 (Pax-QNR) mRNAs are expressed from two promoters used differentially during retina development and neuronal differentiation." *Mol Cell Biol* **15**(6): 3344-53.

Plaza, S., S. Saule and C. Dozier (1999). "High conservation of cis-regulatory elements between quail and human for the Pax-6 gene." *Dev Genes Evol* **209**(3): 165-73.

Quandt, K., K. Frech, H. Karas, E. Wingender and T. Werner (1995). "MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data." *Nucleic Acids Res* **23**(23): 4878-84.

Rosen, V., J. Nove, J. J. Song, R. S. Thies, K. Cox and J. M. Wozney (1994). "Responsiveness of clonal limb bud cell lines to bone morphogenetic protein 2 reveals a sequential relationship between cartilage and bone cell phenotypes." *J Bone Miner Res* **9**(11): 1759-68.

Rothenberg, E. V. (2001). "Mapping of complex regulatory elements by pufferfish/zebrafish transgenesis." *Proc Natl Acad Sci U S A* **98**(12): 6540-2.

Rowitch, D. H., Y. Echelard, P. S. Danielian, K. Gellner, S. Brenner and A. P. McMahon (1998). "Identification of an evolutionarily conserved 110 base-pair cis-acting regulatory sequence that governs Wnt-1 expression in the murine neural plate." *Development* **125**(14): 2735-46.

Santagati, F., J. K. Gerber, J. H. Blusch, C. Kokubu, H. Peters, J. Adamski, T. Werner, R. Balling and K. Imai (2001). "Comparative analysis of the genomic organization of Pax9 and its conserved physical association with Nkx2-9 in the human, mouse, and pufferfish genomes." *Mamm Genome* **12**(3): 232-7.

Sasaki, H. and B. L. Hogan (1996). "Enhancer analysis of the mouse HNF-3 beta gene: regulatory elements for node/notochord and floor plate are independent and consist of multiple sub-elements." *Genes Cells* **1**(1): 59-72.

Sasaki, H., C. Hui, M. Nakafuku and H. Kondoh (1997). "A binding site for Gli proteins is essential for HNF-3beta floor plate enhancer activity in transgenics and can respond to Shh in vitro." *Development* **124**(7): 1313-22.

Satokata, I. and R. Maas (1994). "Msx1 deficient mice exhibit cleft palate and abnormalities of craniofacial and tooth development." *Nat Genet* **6**(4): 348-56.

Schlake, T. and J. Bode (1994). "Use of mutated FLP recognition target (FRT) sites for the exchange of expression cassettes at defined chromosomal loci." *Biochemistry* **33**(43): 12746-51.

Schwartz, S., Z. Zhang, K. A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison and W. Miller (2000). "PipMaker--a web server for aligning two genomic DNA sequences." *Genome Res* **10**(4): 577-86.

Shimeld, S. M. and P. W. Holland (2000). "Vertebrate innovations." *Proc Natl Acad Sci U S A* **97**(9): 4449-52.

Smale, S. T. (1997). "Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes." *Biochim Biophys Acta* **1351**(1-2): 73-88.

Stapleton, P., A. Weith, P. Urbanek, Z. Kozmik and M. Busslinger (1993). "Chromosomal localization of seven PAX genes and cloning of a novel family member, PAX-9." *Nat Genet* **3**(4): 292-8.

Stockton, D. W., P. Das, M. Goldenberg, R. N. D'Souza and P. I. Patel (2000). "Mutation of PAX9 is associated with oligodontia." *Nat Genet* **24**(1): 18-9.

Sun, H., A. Rodin, Y. Zhou, D. P. Dickinson, D. E. Harper, D. Hewett-Emmett and W. H. Li (1997). "Evolution of paired domains: isolation and sequencing of jellyfish and hydra Pax genes related to Pax-5 and Pax-6." *Proc Natl Acad Sci U S A* **94**(10): 5156-61.

Swaminathan, S., H. M. Ellis, L. S. Waters, D. Yu, E. C. Lee, D. L. Court and S. K. Sharan (2001). "Rapid engineering of bacterial artificial chromosomes using oligonucleotides." *Genesis* **29**(1): 14-21.

Timmons, P. M., J. Wallin, P. W. Rigby and R. Balling (1994). "Expression and function of Pax 1 during development of the pectoral girdle." *Development* **120**(10): 2773-85.

Treisman, J., E. Harris and C. Desplan (1991). "The paired box encodes a second DNA-binding domain in the paired homeo domain protein." *Genes Dev* **5**(4): 594-604.

Venkatesh, T. V., N. D. Holland, L. Z. Holland, M. T. Su and R. Bodmer (1999). "Sequence and developmental expression of amphioxus AmphiNk2-1: insights into the evolutionary origin of the vertebrate thyroid gland and forebrain." *Dev Genes Evol* **209**(4): 254-9.

Wada, H., P. W. Holland, S. Sato, H. Yamamoto and N. Satoh (1997). "Neural tube is partially dorsalized by overexpression of HrPax-37: the ascidian homologue of Pax-3 and Pax-7." *Dev Biol* **187**(2): 240-52.

Wada, H., H. Saiga, N. Satoh and P. W. Holland (1998). "Tripartite organization of the ancestral chordate brain and the antiquity of placodes: insights from ascidian Pax-2/5/8, Hox and Otx genes." *Development* **125**(6): 1113-22.

Wahle, E. and U. Ruegsegger (1999). "3'-End processing of pre-mRNA in eukaryotes." *FEMS Microbiol Rev* **23**(3): 277-95.

Wallin, J., H. Eibel, A. Neubüser, J. Wilting, H. Koseki and R. Balling (1996). "Pax1 is expressed during development of the thymus epithelium and is required for normal T-cell maturation." *Development* **122**(1): 23-30.

Wallin, J., Y. Mizutani, K. Imai, N. Miyashita, K. Moriwaki, M. Taniguchi, H. Koseki and R. Balling (1993). "A new Pax gene, Pax-9, maps to mouse chromosome 12." *Mamm Genome* **4**(7): 354-8.

Wallin, J., J. Wilting, H. Koseki, R. Fritsch, B. Christ and R. Balling (1994). "The role of Pax-1 in axial skeleton development." *Development* **120**(5): 1109-21.

Walther, C., J. L. Guenet, D. Simon, U. Deutsch, B. Jostes, M. D. Goulding, D. Plachov, R. Balling and P. Gruss (1991). "Pax: a murine multigene family of paired box-containing genes." *Genomics* **11**(2): 424-34.

Wang, C. C., T. Brodnicki, N. G. Copeland, N. A. Jenkins and R. P. Harvey (2000). "Conserved linkage of NK-2 homeobox gene pairs Nkx2-2/2-4 and Nkx2-1/2-9 in mammals." *Mamm Genome* **11**(6): 466-8.

Wasserman, W. W., M. Palumbo, W. Thompson, J. W. Fickett and C. E. Lawrence (2000). "Human-mouse genome comparisons to locate regulatory sites." *Nat Genet* **26**(2): 225-8.

Watanabe, Y., D. Duprez, A. H. Monsoro-Burq, C. Vincent and N. M. Le Douarin (1998). "Two domains in vertebral development: antagonistic regulation by SHH and BMP4 proteins." *Development* **125**(14): 2631-9.

Wilkins, S. W. (1993). Genetic Analysis of Animal Development. New York, Wiley-Liss

Wilm, B., E. Dahl, H. Peters, R. Balling and K. Imai (1998). "Targeted disruption of Pax1 defines its null phenotype and proves haploinsufficiency." *Proc Natl Acad Sci U S A* **95**(15): 8692-7.

Wunderle, V. M., R. Critcher, N. Hastie, P. N. Goodfellow and A. Schedl (1998). "Deletion of long-range regulatory elements upstream of SOX9 causes campomelic dysplasia." *Proc Natl Acad Sci U S A* **95**(18): 10649-54.

Yang, X. W., P. Model and N. Heintz (1997). "Homologous recombination based modification in Escherichia coli and germline transmission in transgenic mice of a bacterial artificial chromosome." *Nat Biotechnol* **15**(9): 859-65.

Yu, D., H. M. Ellis, E. C. Lee, N. A. Jenkins, N. G. Copeland and D. L. Court (2000). "An efficient recombination system for chromosome engineering in Escherichia coli." *Proc Natl Acad Sci U S A* **97**(11): 5978-83.

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**(2): 709-21.

Zhang, Y., J. P. Muyrers, G. Testa and A. F. Stewart (2000). "DNA cloning by homologous recombination in Escherichia coli." *Nat Biotechnol* **18**(12): 1314-7.

Postscriptum

Since the first submission of this PhD thesis, additional experiments have failed to reproduce the rescue of the Pax9 knockout phenotype with the BAC transgene, as described in the Result section 4.6.4.
At this moment, further analysis is required to verify the initial observations.

I apologize for the inconvenience, believing that this does not affect the general significance of the work.

Fabio Santagati