ELUCIDATION OF GENE REGULATORY NETWORK CONTROLLING EMBRYONIC SKELETAL DEVELOPMENT:

FROM THE PERSPECTIVE OF *Pax1* & *Pax9*

V SIVAKAMASUNDARI

(*B.Sc. (Hons.), NUS*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOLOGICAL SCIENCES

NATIONAL UNIVERSITY OF SINGAPORE

2012

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety.
I have duly acknowledged all the sources of information that have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

_____

V Sivakamasundari

Aug 2012

ACKNOWLEDGEMENTS

I am sincerely thankful and greatful to all the people who have helped me on this journey. I started my graduate studies with an important goal in mind, and it would not have been possible to achieve it without the assistance, guidance and support from many people.

I owe my earnest thanks to my supervisor Dr Thomas Lufkin for being a great mentor. He has been patient and given me the necessary independence to work on my project. I have learnt many valuable techniques in his lab, an opportunity I would have missed if I had had started my career elsewhere. His encouragements, guidance and confidence in my work have all been the motivating factors during the course of my study.

I would also like to thank Dr Christoph Winkler, my co-supervisor for his inputs on the thesis and sharing his insights on fish development.

Dr Tara Huber for stepping-in on my behalf to ensure that I received sufficient funding to complete my program. Her advice and guidance on the various aspects of my project and beyond were certainly invaluable.

My sincere thanks also goes to my friend and colleague Dr Chan Hsiao Yun, whom I've had great pleasure working with and shared many thoughts on science among other things. Her constant discussions and assistance on our lab projects were very refreshing and helpful.

Dr Petra Kraus has been an important pillar to all of our projects in the TL lab, with her skillful and tireless ability to generate and maintain the numerous mouse lines.

My gratitude goes to her for providing me the very much needed moral support during difficult times and sharing her insightful thoughts on the project.

Dr Shayam Prabhakar and his post-docs Dr Sun Wenjie, Hu Xiaoming, and Dr Vibhor Kumar, who were greatly helpful with the bioinformatics analysis and always went that extra mile.

A special thanks to my friend Dr Nirmala for sharing her experience, advice, encouragements and all the lunch hours filled with interesting chats on practically everything under the sun.

To all my colleagues who have helped me in different ways at some point of my project: Song Jie, Siew Lan, Sook Peng, Xing Xing, Serene Lee, Sumantra, Cecilia, Eileen Tan, Dr Sinnakaruppan Mathavan and the BSF FACS facility: Michelle Mok, Chee Zhe Jie Keefe, Leck Thye Seng and Toh Xue Yun. Especially Serene, Sumantra and Siew Lan for their encouragements.

Most importantly, words cannot express my appreciation and gratitude to my parents who were ever supportive of my pursuit of graduate studies, sisters Suchi and Indu, and friends Ashik, Nivetha, Ashita and Kaiwee for their immense support and for always going the extra mile to make my day. They had always believed in me and motivated me throughout this trying journey. Their moral support is what has helped me pull through and complete my dissertation.

TABLE OF CONTENTS

ABSTRACT

The osteogenic and chondrogenic lineages derived from mesenchymal stem cells (MSCs) are of immense biomedical importance especially in the area of regenerative therapy for numerous degenerative bone diseases and developmental defects. The coordinated expression of key transcription factors (eg. *Pax, Runx, Sox* etc.) orchestrate the commitment of the MSCs towards the chondro-osteogenic lineage. However, much remains to be learned about the regulatory relationships between these transcription factors (TFs) controlling embryonic skeletal development.

Immense research has been carried out to elucidate the roles of the *Sox* and the *Runx* family of TFs which are master regulators in the chondro-osteogenic pathway. Yet, less attention has been conferred upon other early acting TFs like *Pax1* and *Pax9* which are critical in patterning and differentiation of the sclerotomal cells that give rise to the vertebral bodies and intervertebral discs of the axial skeleton. Using mice as the experimental model, gene-targeting strategies and current genomic technologies were employed to identify, for the first time, the target genes of *Pax1* and *Pax9*, in a cell-type specific manner.

Pax1 and *Pax9* were knocked-out by the insertion of EGFP in their exons, in order to enrich for *Pax1* and *Pax9* cell lineages. For a WT comparison, EGFP was co-expressed with Pax1 using the F2A-peptide strategy. Besides, Pax1 and Pax9 proteins were successfully endogenously tagged with hemagglutinin (HA) epitope for use in TF mapping and other protein-related studies.

Using FACS, highly enriched populations of *Pax1*- and *Pax9*-specific cells were used on microarrays. Firstly, genes enriched in *Pax1*-specific cells at E12.5 and E13.5 stages were identified. Subsequently, the target genes of *Pax1* and *Pax9* were

discovered from the various knock-outs (*Pax1*$^{-/-}$, *Pax1*$^{-/-}$*Pax9*$^{+/-}$ & *Pax1*$^{-/-}$*Pax9*$^{-/-}$). The use of 3-allele and 4-allele knock-outs enabled the identification of *Pax1* and *Pax9* regulated genes that were masked in the *Pax1*$^{-/-}$ embryos by the functional redundancy between *Pax1* and *Pax9*.

In parallel, TF mapping performed on the wild-type embryos helped to distinguish the direct and indirect targets of Pax1 and Pax9. From this, the molecular functions of *Pax1* and *Pax9* could be delineated. *Pax1* and *Pax9* appear to have a role in regulating the early functions of intervertebral disc morphogenesis, i.e. cell proliferation, cell adhesion, cell motion, condensation, ECM organization and cartilage development. Also, a novel link between the *Pax* genes and *Sox5* has been identified. Moreover, the *Pax* genes regulate several of the genes that are known to be regulated by the *Sox* trio (*Sox5/Sox6/Sox9*). While the *Pax* genes are not master regulators of chondrogenesis, they probably play accessory roles by assisting the *Sox* genes in initiating the early expression of chondrogenic genes. Once the chondroblasts mature into chondrocytes, these *Pax* genes are down-regulated in the chondrocytes possibly by a negative feed-back mechanism.

In conclusion, this genome-wide, non-hypothesis driven study has provided a better understanding on the roles of *Pax1* and *Pax9* and helped to formulate more hypotheses regarding their molecular functions. The data and the numerous mouse lines generated in this study also serve as an invaluable resource to construct the gene regulatory network of embryonic skeletal development.

(505 words)

List of Tables

List of Figures

List of Abbreviations

TF – Transcription Factor

CRE – *Cis*-Regulatory Element

TRE – *Trans*-Regulatory Element

kb - kilobases

bp – base pairs

GRN – Gene Regulatory Network

GO – Gene Ontology

MSC – Mesenchymal Stem Cell

ESC – Embryonic Stem Cell

IVD – Intervertebral Disc

ECM – Extracellular Matrix

ChIP-Seq – Chromatin Immunoprecipitation Sequencing

HTH – Helix-turn-helix

UTR – Untranslated region

RIN – RNA Integrity Number

CHAPTER 1 – INTRODUCTION

1.1 Gene regulation – the central dogma, revised.

The sequencing of human genome to a "finished-grade" by 2004 has provoked an explosion of sequencing technologies over the past decade [1]. The burgeoning sequencing technologies have enabled us to probe the eukaryotic DNA and RNA sequences in greater depth and at a single base resolution [2]. This has revealed the unprecedented complexities of the genome architecture, whereby gene regulation is not really modular as once thought, but involves an intricate orchestration of protein molecules (transcription factor, TFs; co-factors; chromatin modifiers; transcription machinery complex) and RNAs (long non-coding RNAs, lncRNAs; lincRNAs; retrotransposon-derived RNAs; micro RNAs, miRNA etc) acting on segments of DNA (*cis*- & *trans*-regulatory elements, CREs/TREs and promoter) [3, 4].

The central dogma of genetics described a "gene" as a segment of DNA that could be transcribed into mRNA and then translated into a protein. Everything else was deemed to be "junk" DNA. However, in the past decade significant evidence has emerged to prove the importance of such "junk" DNA which do produce either non-coding RNAs or function as cis-regulatory elements, all of which are paramount to genetic regulation. Indeed, organismal complexity arises not just because of the increase in protein diversity (from alternative splicing of transcripts), but also because of the increased level of genomic regulation by the trans-acting factors (TFs and non-coding RNAs). For instance, while only ~3% of the protein-coding genes encompass TFs in a simple, unicellular eukaryote like *Saccharomyces cerevisiae*, in the more complex multicellular nematode *C.elegans* it is about ~5%, and in the much more complex mouse and humans, it is about ~10% [5]. Moreover, the percentage of non-protein-coding DNA in humans is ~98%; a drastic difference from that of a prokaryote

which is only ~12% [6]. The non-protein-coding DNA could encode non-coding RNAs (long RNAs, miRNAs etc) which function as trans-acting factors, or serve as cis-regulatory elements for TF binding. Indeed, the repetitive sequences in the human genome, mainly derived from transposable elements, have been shown to encompass TF binding sites (cis-elements) [7-9]. Transposable elements possessing TF binding motif precursor sequence, once integrated into the genome, could evolve into novel, species-specific TF binding sites [7-9]. Thus, the coding and non-coding components of the genome contribute to colossal numbers of permutations and combinations of trans-acting factors interacting with the cis-elements that presumably give rise to organismal complexity [6]. With that realization, it is evident that the genome is an efficiently organized information system and nothing is really "junk". This shift in the paradigm of gene regulation has completely transformed our interpretation of the genetic landscape and hence, our approach to unravelling its three-dimensional architecture.

1.2 The conceptual framework – the GRN

In a multicellular organism, the individual cell types are determined by differential gene expression. Such spatio-temporally regulated expression of a combination of genes, is called the "gene battery" [3, 10]. As a single gene can have multiple cis-regulatory elements (CREs) and a particular CRE can be bound by several TFs, to regulate the expression of that gene in a specific tissue and time. Thus, it is the trans-acting factors like the TFs, which bind to a subset of these CREs, and miRNAs[1] which regulate gene expression at post-transcriptional level, that will determine the composition of the gene battery. While non-coding RNAs are a recent

[1] miRNA mediate repression at post-transcriptional level by binding to the target transcripts at the 3'UTR and inhibiting its translation or reducing its stability. The mode of miRNA action is different from that of the TFs, which have the ability to activate a gene as well.

discovery whose regulatory functions are constantly being updated, the TFs have long taken the center-stage in our pursuit of understanding gene regulation. Genome-wide techniques such as microarray and ChIP-chip, and two-hybrid (yeast and mammalian) paved the way to examine gene expression patterns, protein-DNA and protein-protein interactions in a systematic fashion [5, 11]. Such complex interconnections of the TFs (with their interactors) with their CREs, and the causal links of the trans-acting factors with their target genes, can be mapped into a comprehensive conceptual framework – the gene regulatory network (GRN).

The transcriptional network regulates the expression of this "gene battery" and determines the differentiation program of stem cells into specific lineages. The composition of activated and repressed genes by a combination of TFs would in turn control the various signalling pathways to execute the specification, commitment and differentiation of the precursors to a particular lineage. Dysregulation of such transcriptional regulatory programs can give rise to diseases owing to aberrant behaviour of cells (eg. cancer, diabetes, congenital diseases and developmental defects) [12].

Modelling complex gene regulation as a network map presents numerous advantages. GRNs will enable us to interrogate the network motifs within, which may assist us in understanding the mechanisms of regulation of a specific biological process. For example, feed-forward loops result in a gene to be expressed quickly, while feed-back/ auto-regulatory loops either reinforce or further reduce the expression of a gene. Such observed patterns can then be coupled with the known functions of the process-in-question to comprehend that biological process. Moreover, such network maps allow formulation of hypotheses to be made which can be further tested experimentally. We can also predict the outcome of various

perturbations to the network, and thus design appropriate therapies (eg. regenerative therapies, tissue engineering, multi-target drugs) for numerous diseases [5, 12].

GRNs are composed of nodes and edges, whereby the nodes are the biological molecules (DNA, protein, miRNA etc) while the edges represent the functional association between them (eg. activation or repression) [5]. Thus, construction of GRNs for any process requires four key data sets: (1) the protein-DNA interaction, (2) protein-protein interaction, (3) the causal links between the TFs / miRNAs and their target genes and (4) spatio-temporal expression of genes [13]. Genome-wide *in vitro* or *in vivo* data for each of these components can be acquired via a myriad of techniques, which are summarized in Table 1.

Table1: Data sets required for the construction of a GRN and techniques that can be used to acquire those data.

	Information	*In vitro* technique(s)	*In vivo* technique(s)
1	Protein-DNA interaction	PBM, Y1H, B1H, SELEX, luciferase-based PDI mapping, microfluidics-based PBM, luciferase assay (small-scale), EMSA (small-scale)	ChIP-chip, ChIP-Seq.
2	Protein-protein interaction	Y2H, M2H, co-IP (small scale), affinity purification (small scale), mass spectrometry (small scale)	co-IP (small scale), affinity purification (small scale), mass spectrometry (small scale)
3	Causal links between TFs/ miRNAs and their target genes	RNAi	Microarray, RNA-Seq
4	Spatio-temporal expression	-	*In situ* hybridization, RT-qPCR

PBM- protein binding microarray; Y1H – yeast one hybrid; B1H – bacterial one hybrid; SELEX - systematic evolution of ligands by exponential enrichment; PDI – protein-DNA interaction; ChIP – chromatin immunoprecipitation; EMSA –

electrophoretic mobility shift assay; Y2H – yeast two hybrid; M2H – mammalian two hybrid; co-IP – co-immunoprecipitation.

Evidently, constructing a GRN for even a single process requires a vast amount of data, and acquiring that is a daunting task. Indeed such a task need not be handled independently since it is time-consuming, labour-intensive and simply very expensive. Researchers world-wide have been generating genome-wide data sets which can be integrated to eventually generate the GRN. For instance, GRNs have been constructed for (and are constantly being updated) endomesoderm specification of the sea urchin [14, 15], dorsal-ventral patterning of *Drosophila [16, 17]*, vulva development [18] and neuron cell type specification in *Caenorhabditis elegans* [19] and mesendoderm development in *Xenopus* [20, 21]. These networks were not constructed overnight but took decades of data collection and required the effort of numerous independent labs. Also, with the complexities of gene regulation, such networks have a long way to attain completion. Nonetheless, data collection is the first-most obligatory step for modelling such networks.

1.3 Bone development

Embryonic bone formation is a tightly regulated process that can occur through endochondral ossification or intramembranous ossification [22]. Most of the bones of the axial and appendicular skeleton and some craniofacial bones are formed by endochondral ossification. In this process, three major cell types are involved: the chondrocytes, osteoblasts and osteocytes, which are all derived from a common precursor, the msenchymal stem cell (MSC). The MSCs first form a condensation which is mostly complete by embryonic day (E) 10.5 [23]. These cells produce extracellular matrix (ECM) composed of collagen type I. The cells within the condensations then differentiate into chondrocytes and secrete ECM components rich in collagen type II and aggrecan. The peripheral cells of the condensation,

however, form the perichondrium and continue to secrete collagen type I instead. The cartilage condensation thus formed, acts as a template of the future bone. The chondrocytes of the cartilage subsequently become hypertrophic, secrete ECM composed of collagen type X, then undergo terminal differentiation and eventually die through apoptosis. In parallel, the perichondrial cells differentiate into osteoblasts upon Indian hedgehog (Ihh) signal induction from the pre-hypertrophic cells, thus forming the periosteum [23]. Meanwhile, the ECM in the immediate vicinity around the hypertrophic chondrocytes is degraded by matrix metalloproteinases (MMPs) and a disintegrin and metalloproteinase with thrombospondin motifs (ADAMTSs) family of enzymes [24]. This is followed by an invasion of blood vessels through a vascular endothelial growth factor (VEGF)-dependent pathway, which imports the osteoblast precursors, osteoclasts and bone marrow cells to the center of the cartilaginous template. While the osteoclasts play a critical role in bone resorption, the differentiating osteoblasts then replace the remnant cartilaginous template with bone. The mineralization of this cartilage matrix occurs through the deposition of hydroxyapatite [24].

Contrary to endochondral ossification, the intramembranous ossification process does not involve a cartilage intermediate. The MSCs differentiate directly into osteoblasts. These osteoblasts secrete a fibrillar, non-calcified ECM called osteoid, which in turn become mineralized to form the bone. This process forms parts of the skull bones (eg. the frontal and parietal bones of the neurocranium (skull roof)) and lateral parts of the clavicles [25].

1.3.1 Key players in skeletogenesis

Although both ossification mechanisms of embryonic bone development are distinct, cells in the majority of skeletal elements are derived from a common

precursor - the MSCs. As with other differentiation pathways, the restriction of MSCs towards the chondro-osteogenic lineage in skeletal development involves the coordinated and sequential expression of key TFs (e.g. *Bapx1 (Nkx3.2), Pax1, Pax9, Runx2, Runx3, Osterix, Sox9, Sox5, Sox6* etc.) and the involvement of various hormones (growth and thyroid hormone) and local secreted factors (Ihh, PTHrP, BMP, Wnt, FGFs) [24]. The various TFs involved in the chondro-osteogenic pathway are depicted in a schematic diagram in Figure 1.

Of these, *Sox9* (*SRY-box containing gene 9*) is the master regulator of chondrogenesis while *Runx2* is the master gene for osteogenesis. *Sox9* is known to activate numerous chondrogenic markers like *Acan* (*aggrecan*), *Col2a1 (collagen, type II, alpha 1)*, as well as *Sox5* (*SRY-box containing gene 5*) and *Sox6* (*SRY-box containing gene 6*) TFs which are important for chondrocyte differentiation. It plays essential functions in promoting chondrocyte proliferation while inhibiting its hypertrophy. Moreover, loss-of-function mutations of *Sox9* gives rise to Campomelic dysplasia, which is a form of skeletal dysplasia, resulting in abnormalities of the head, neck and long bones and is often lethal [23, 26, 27].

Runx2 (runt related transcription factor 2), on the other hand, is essential for osteogenesis as *Runx2*$^{-/-}$ mouse mutants completely lack osteoblasts in all the skeletal elements. *Runx2* also regulates the expression of osteoblast-specific hormone, *Osteocalcin* and osteoblast-specific TF, *Osterix*. Besides, it also plays dual roles in chondrocytes; when expressed transiently in the pre-hypertrophic chondrocytes it promotes hypertrophy, whereas its constitutive expression in the perichondrium inhibits both chondrocyte proliferation and hypertrophy [28].

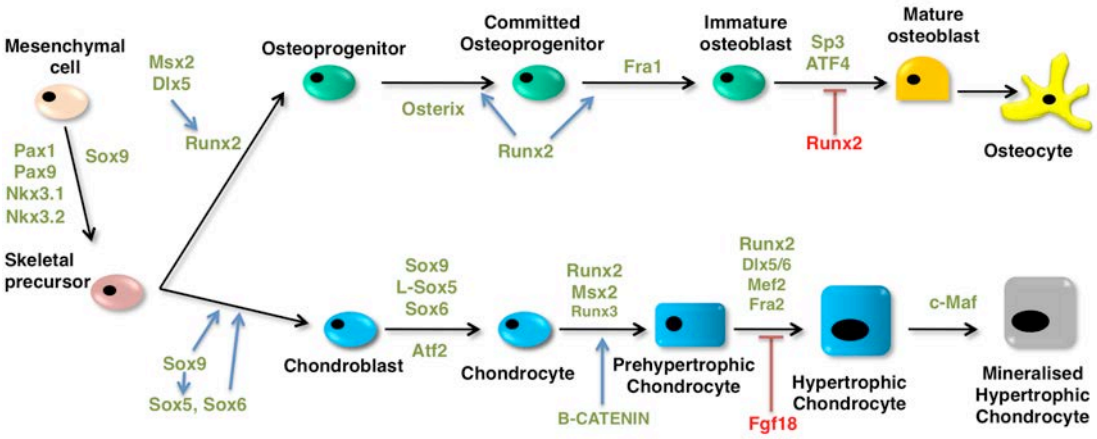


Figure 1: Transcription factors involved in the commitment of the mesenchymal stem cells in the chondro-osteogenic pathway. Positive regulation and positively acting factors are shown in green, negative regulation and negatively acting factors are shown in red. Reference for lineage commitment figure - Current Opinion in Genetics & Development 2009, 19:437–443 [25].

1.4 Vertebral Column Structure and Development

An important skeletal structure and the defining framework of all vertebrates, is the vertebral column, which confers support, flexibility and protects the spinal cord and crucial spinal nerves of the body. It constitutes metamerically arranged vertebral bodies linked together by intervertebral discs (IVDs) [29]. The IVD is an indispensable aspect of the vertebral design, withstanding biomechanical forces and conferring tensile strength. It also provides flexibility in motion to an otherwise rigid spine [30]. The mature IVD is a multi-component fibro-cartilaginous structure, with each component possessing distinct biochemical properties so as to execute their different biomechanical functions. The central nucleus pulposus of the IVD is semi-fluidic and acts as a shock absorber. The circular annulus fibrosus, which encases the nucleus pulposus, functions to withstand the compressive forces acting on the spine. It also endures tension and holds the nucleus pulposus in the center during compression. Rostrally and caudally positioned to the annulus fibrosus are the cartilaginous end plates which are composed of hyaline cartilage and connect the

adjacent bony vertebrae to the annulus fibrosus. These components thus function together to transmit and evenly distribute the body weight and load from physical activity [29].

1.4.1 Embryonic axial skeletogenesis and its genetic regulation

Axial skeletogenesis is a multi-step process beginning with somitogenesis and the paraxial mesoderm that is located adjacent to the neural tube [31]. Spherical balls of cells – the somites, "bud off" from the unsegmented mesoderm in a periodic manner. The Notch signalling pathway is instrumental in this segmentation process and *Paraxis* is required to form the metameric blocks of epithelial somites [32-34]. The cells within the somites become specified to different fates depending on the multiple extracellular signals they are exposed to from the surrounding tissues. Of these, Wnt signals from the dorsal neural tube and surface ectoderm, BMP signals from the dorsal neural tube or lateral plate mesoderm and *Sonic Hedgehog* (*Shh*) signals from the notochord and the floor plate (of neural tube) all play critical and antagonistic roles [35-37]. The Shh morphogen, by antagonizing Wnt signals, induces the ventral somites to a sclerotomal fate by activating the expression of *Paired-box 1* (*Pax1), Paired-box 9 (Pax9)* and *Mesenchyme forkhead-1* (*Mfh1)* [35, 38-41]. The proliferative function of *Shh* is thus believed to be mediated by the *Pax1/Pax9* and *Mfh1* [41]. Also, maintenance of a BMP-reduced zone in the ventral somite by BMP antagonists – *Noggin (Nog)* and *Gremlin (Grem1)*, is required for this sclerotomal specification [42-45]. Early exposure of the somites to BMP signals renders them to a lateral mesoderm fate, while a late exposure to specified sclerotomal cells (after exposure to Shh) promotes a chondrogenic fate [37, 46]. Thus, both *Shh* and *Nog* have the potential to induce *Pax1* expression in the ventral somites, through potentially parallel pathways. *Pax1* expression is detected, although

at diminished levels, in the *Shh*$^{-/-}$ embryos and *Nog* is able to induce *Pax1* expression even in the absence of *Hh* signalling [42, 47].

The ventro-medial cells of the somites proliferate and then de-epithelialize to become mesenchymal sclerotomal cells [31]. These sclerotomal cells then migrate to surround the notochord and form the mesenchymal prevertebrae. The sclerotomal cells in the immediate vicinity of the notochord give rise to the vertebral bodies and IVDs, while the lateral portions form the proximal parts of ribs, vertebral pedicles and laminae of the neural arch of the spine [48, 49]. The prevertebrae forms a segmented condensed and less condensed regions along the anterior-posterior (A/P) axis of the embryo at around E12.5. The former gives rise to the IVD anlagen while the latter develops into vertebral bodies [50].

Thus, with respect to vertebral column development, both IVD (annulus fibrosus) and vertebral bodies are derived from a specified pool of sclerotomal cells [48], while the nucleus pulposus is derived from the notochord [30]. Despite a shared cellular ancestry, the vertebral body and IVD fates are acquired through different underlying molecular mechanisms. Moreover, in the embryonic stages, the annulus fibrosus is composed of two portions - a cartilaginous inner annulus and a fibrous outer annulus. The genetic pathway leading to a vertebral fate has been well-studied, however there is a dearth of information regarding the molecular pathway leading to an annulus fibrosus fate.

While *Shh*, *Nog* and *Grem1* clearly play a role in sclerotome specification, *Pax1*, *Pax9* and *Mfh1* are essential in sclerotome maintenance by regulating their numbers. Indeed, *Pax1* and *Mfh1* are known to genetically interact as the *Pax1*$^{-/-}$ *Mfh1*$^{-/-}$ double-null mutants showed a reduced cell proliferation [41]. The fact that *Mfh1* expression was unaffected in the *Meox1*$^{-/-}$*Meox2*$^{-/-}$ (*Mesenchyme homeobox 1*

and 2) mutants, but *Pax1* expression (*Pax9* to a lesser extent) was reduced, indicates a potential parallel pathway in sclerotome maintenance by the *Pax1/Pax9* and *Mfh1* [51].

1.4.1.1 Vertebral body fate determination

The subsequent differentiation of the sclerotome into chondroblasts in the vertebral bodies requires *Meox1/Meox2*, *Nkx3.2/Nkx3.1* and *Sox9*.

Sox9 is a well-known master regulator of chondrogenesis [52]. *Nkx3.2 (a.k.a Bapx1, bagpipe homeobox gene 1 homolog)* also appears to be essential based on the observation that the pre-chondroblasts failed to differentiate into chondrocytes in the vertebral bodies of the *Bapx1*$^{-/-}$ mutants [53]. This was accompanied by a loss of *Sox9*, *Col2a1* and *fibroblast growth factor receptor 3* (*Fgfr3*) expression in the vertebral anlagen, all of which are chondrogenic differentiation markers [53]. At the same time, the ability of *Sox9* to drive *Bapx1* expression *in vitro*, by binding to its promoter, indicates a potential positive-regulatory loop between *Sox9* and *Bapx1[54]*. Indeed, *Sox9* and *Bapx1* were able to induce each others expression in chick explant cultures when over-expressed [54, 55].

Meox1 and *Meox2* are not needed for sclerotome specification but they act upstream of *Pax1*, *Pax9* and *Bapx1*. In the *Meox1*$^{-/-}$*Meox2*$^{-/-}$ mutants, *Pax1*, *Pax9* and *Bapx1* expression was lost in the sclerotome but not *Mfh1 [51]*. Furthermore, both *Meox1* and *Pax1* (*Pax9* to a smaller extent) have been shown to bind to the promoter of *Bapx1* and transactivate its expression *in vitro* [56, 57]. Also, *Meox1* transactivation of *Bapx1* expression was enhanced in the presence of *Pax1* and/or *Pax9* [56]. Indeed, *Pax1* and *Meox1* are known to interact at protein-level *in vitro* [58], and their binding sites in the promoter of *Bapx1* are adjacent to each other [56, 57]. This indicates a potential genetic hierarchy whereby *Meox1* is upstream of *Pax1*,

which is in turn upstream of *Bapx1*. However, in wild-types, *Bapx1* expression is maintained in the chondrogenic cells of the vertebral bodies, while *Pax1* and *Pax9* expression are down-regulated. This led to the hypothesis that *Pax1/Pax9* may be required for the initiation of *Bapx1* but not for its maintenance at later stages [57]. *Sox9*, which is expressed in the chondrogenic cells may be involved in its maintenance. It is postulated that the *Shh* induced positive regulatory loop between *Sox9* and *Bapx1* is subsequently maintained by BMP signals [55].

Thus, the current hypothesis is that *Pax1/Pax9* and *Mfh1* are needed to expand the sclerotome population to a sufficient density, upon which *Sox9* and *Bapx1* can then confer a chondrogenic potential, whereby the BMP signal is able to differentiate these competent cells to their chondrogenic fates.

1.4.1.2 Annulus fibrosus (IVD) fate determination

Mfh1 and TGF-beta signalling are known to be important players in annulus fibrosus fate determination,. The annulus fibrosus structure was reduced / abnormal in the *Mfh1*$^{-/-}$ and *Tgfbr2* conditional KO (in *Col2a1*-expressing cells) mutants [59, 60]. Moreover, gene expression profiling analysis of the IVD anlagen from *Tgfbr2* conditional KO mutants showed the mutant gene expression profile more closely resembled a wild-type vertebral body profile than the wild-type IVD. The authors had also observed a similar phenotype whereby the boundary between the vertebral body and IVD was not maintained. Moreover, *Fibromodulin* (*Fmod*, an IVD marker) and several other IVD markers were found to be regulated by TGF-beta signalling. Thus, TGF-beta signalling is required for annulus fibrosus differentiation and potentially prevents inappropriate chondrocyte differentiation in the IVD region [60, 61].

Owing to the tissue-specific expression of *Pax1* and *Pax9* in the IVD anlagen and the complete absence of the IVD structures in the *Pax1*$^{-/-}$*Pax9*$^{-/-}$ double-null

mutants indicates their potential role in IVD formation [62, 63]. It is not clear if it is simply a patterning and cell maintenance role, or if these genes are also essential in the differentiation of the sclerotome cells to an annulus fibrosus fate (fibrosus tissue). Whether *Pax1* or *Pax9* act as competence factors in sclerotome cells for their further differentiation into the annulus fibrosus by other signals (eg. TGF-beta signalling), is unknown. Also, *Sox5/Sox6* also appear to be important in annulus fibrosus differentiation since they are known to induce *Fmod* expression just like TGF-beta signalling. How exactly the *Pax* and *Sox* genes come into play in the IVD formation remains unknown.

Thus, a considerable amount of research has been carried out to reveal the numerous TFs involved in the general chondro-osteogenic pathway. Yet, little is known about the regulatory relationships between the TFs orchestrating the embryonic skeleton development. Moreover, the focus has largely been on the ossification processes in general, without much attention given to the annulus fibrosus fate determination, which forms an indispensable part of the vertebral column.

In this study, the main focus is on the *Pax1* and *Pax9* genes which are essential for axial skeletogenesis, especially for IVD formation. Some of the known characteristics and functions of these two *Pax* genes are discussed in the following section.

1.5 The *Pax* genes

The *Pax* gene family was initially identified through similarity to the “paired-box” in the *Drosophila* gene *gooseberry and paired* [64-66]. This *Pax* family constitutes a group of nine genes (*Pax1-Pax9*) encoding TFs with the highly conserved DNA-binding domain, the paired-box (of 128-amino acids). Genes within

the family are further divided into subfamilies based on their sequence similarity, the combination of domains they possessed: paired-domain containing two helix-turn-helix (HTH) motifs [67], paired-type homeodomain (full or truncated) and octapeptide motif (HSVSNILG) [68]; and their overlapping domains of expression [65].

Of these *Pax1* and *Pax9* belong to the same subfamily, essential for the early stages of axial skeleton formation. Moreover, *Pax1* and *Pax9* are the only *Pax* genes that are expressed in the sclerotomal cells [69]. They contain only the paired-domain and the octapeptide motif, and share a high protein sequence similarity of 79%, diverging mainly at their C-terminal ends. Within the paired-domain they differ in only 3 amino acids, at positions 82, 89 and 93, which belong to the C-terminal half of the paired box [50]. They also share similar expression domains, especially in the pharyngeal pouch endoderm, sclerotome and later in the intervertebral disc anlagen [50].

The Shh morphogen emanating from the notochord and the floor plate (of the neural tube) induces the expression of *Pax1* transcripts at E8.5 in the ventro-medial deepithelializing somites to specify their sclerotomal fates [35, 39, 40]. *Pax9* expression in the sclerotome is also mediated by notochordal signals [50]. It is believed *Pax1* and *Mfh1* mediate the proliferative functions of *Shh* [39, 41]. As mentioned earlier, besides *Shh*, *Nog* is also able to induce *Pax1* expression in the somites [42]. Other factors which are not known to independently induce the expression of *Pax1* or *Pax9*, but regulate their expression in the somite are *Pbx1*/*Pbx2* (TALE homeodomain TFs) and *Meox1/Meox2* TFs [51, 70]. In the *Pbx1*$^{-/-}$ *Pbx2*$^{-/-}$ mutants and *Meox1*$^{-/-}$*Meox2*$^{-/-}$ mutants *Pax1* and *Pax9* expression is diminished in the somites / sclerotome. Moreover, the effect of *Pbx1/Pbx2* and *Meox1/Meox2* is greater on the *Pax1* than *Pax9* [51, 70]. This indicates the presence of an alternative mechanism by which *Pax9* is induced in the sclerotome.

1.5.1 Spatio-temporal expression patterns of *Pax1* and *Pax9*

Pax1 expression begins in the deepithelializing somites at E8.5, while detectable levels of *Pax9* expression begin a little later at E9.0. Subsequently in the metameric condensing sclerotome, *Pax1* is initially expressed in both rostral and caudal segments while *Pax9* is restricted to the caudal half. Subsequently both *Pax1* and *Pax9* become restricted to the caudal half of the condensing cells of the intervertebral disc anlagen by E12.5. While *Pax1* expression is stronger in the medial portion of the IVD anlagen, *Pax9* expression is stronger in the lateral portion. *Pax1* remains strongly expressed in the intervertebral disc and the perichondrium of the vertebral bodies till E14.5, after which it declines with only the annulus fibrosus cells of the intervertebral disc expressing it in the post-natal mice. Similarly, *Pax9* is still expressed weakly in the E14.5 intervertebral discs, but by E16.5 its expression is no longer detectable in the vertebral column [50, 63, 71].

Beyond the axial skeleton, *Pax1* and *Pax9* are also expressed in the craniofacial mesenchyme, foregut, appendicular skeleton, pharyngeal arch and its derivatives (maxillary and mandibular arches, thymus, parathyroid glands, ultimobranchial bodies). Both *Pax* genes are expressed in the foregut at E8.5, the pharyngeal pouches at E9.5, anterior proximal limb buds at around E10.0 to E11.5 and thymus anlagen at E12.5 [50, 72, 73]. However the domains of expression in these sites are not always exactly overlapping. For instance, *Pax9* expression in the anterior proximal limb buds is adjacent to the *Pax1* domains of expression [50]. Moreover, *Pax9* also has its own unique sites of expression such as in the ultimobranchial bodies (pharyngeal arch derivative), hindgut, ventral tail mesenchyme, salivary glands, squamous epithelia of esophagus, tongue and tooth mesenchyme, while *Pax1* expression in the sternum at E13.0 has not been mentioned for *Pax9* [50, 73, 74]. Of these, the nasal, palatal and teeth mesenchyme

are neural-crest derived cells. Notably, the craniofacial bones, nasal & palatal bones, are mainly formed by intramembranous ossification without a cartilaginous intermediate [75]. The detailed list of all the expression sites, the stages during which the expression begins/ reported to be expressed at, and the corresponding references are shown in Table 2.

Table 2: *Pax1* and *Pax9* region of expression

Site of expression	*Pax1*	Ref	*Pax9*	Ref
Somite (ventro-medial)	E8.5	[71]	E9.0	[50]
Sclerotome	E8.5		E9.5	
Axial skeleton: IVD anlagen	E12.5		E12.5	
Perichondria lining VB, pedicles and proximal ribs	E12.5		E12.5	
Rib intercostal mesenchyme	E13.5		E12.5	[50, 74]
IVD: AF	~E15.5 onwards		~E15.5 to E16.5 (subsequently expression is lost)	
Ventral tail mesenchyme	Not expressed	NIL	E9.5	
Anterior proximal fore limb bud	E10.0	[76]	E11.5	
Anterior proximal hind limb bud	E10.5		E11.5	
Hindlimb knee joint	E12.5		E12.5	[50]
Hindlimb tarsal	E12.5		E12.5	[74]
Between digits	E12.5		Not mentioned	NIL
Sternum	E13.0	[73]	Not mentioned	
1st arch branchial pouch	E9.5	[50, 72]	E9.5	[50]
2nd arch branchial pouch	E9.5		E9.5	
3rd arch branchial pouch	E9.5 (weak)		E9.5	
4th arch branchial pouch	E10.5		E9.5	
Parathyroid glands	Not mentioned	NIL	E12.0	[74]
Thymus epithelium	E12.5	[72]	E12.0	
Ultimobranchial bodies	Not mentioned	NIL	E12.0	
Nasal process	E11.5	[50]	E10.5	[50]

Wolffian ridge	E11.5		Not expressed	NIL
Salivary glands	Not mentioned	NIL	E16.5	[74]
Esophagus (internal stratified squamous epithelia)	Not expressed		E13.5	[74, 75]
Tongue epithelium	Not expressed		E11.5	[50]
Foregut	E8.5	[72]	E8.5	
Middle ear (tympanic ring)	Not mentioned	NIL	E13.5	[74]
Teeth mesenchyme	Not expressed		E10.0	[74]
Hindgut	Not expressed		E9.5	[50]

1.5.2 Functions of *Pax1* and *Pax9*

The role of *Pax1* in axial and appendicular skeletogenesis was initially identified through spontaneous mouse mutants the *undulated* (*un*) [77], *Undulated short-tail* (Un^{s}) [78], *undulated-extensive* (un^{ex}) [79] and *undulated intermediate* (*un-i*) [80] which consisted of either point mutations or deletions in *Pax1* or deletion of its entire locus [68, 71, 73]. Targeted disruption of *Pax1* subsequently helped to clarify the structures *Pax1* was genuinely essential for - the vertebral bodies, intervertebral disc, scapula, sternum and pelvic girdle [73]. The $Pax1^{+/-}$ were externally similar to wild-type mice, viable and fertile, but with mild abnormalities of certain skeletal elements like the first two cervical vertebrae (atlas-axis), lumbar vertebrae and sternum, but with an overall penetrance of only 88%. The authors had attributed the lack of 100% penetrance of such skeletal defects to a genetic background effect [73]. The $Pax1^{-/-}$ mice, however, were smaller than wild-type and exhibited a shortened, kinked-tail phenotype but were still viable and fertile. They possessed more severe defects in the vertebral column, scapula, sternum and tail. The first two cervical vertebrae (atlas-axis) were inappropriately fused; pedicles were fused to the ventral ossification centres of their respective vertebrae from the thoracic segment onwards

till the caudal region; loss of the acromion process of the shoulder girdle; and inappropriate ossification of some of the intersternebra. The lumbar segments exhibited more severe defects with split vertebrae accompanied by the loss of IVD structures and formation of ventral rod-like cartilage structure [73].

Surprisingly, targeted inactivation of *Pax9* does not result in any vertebral column abnormalities, but shows limb, craniofacial, teeth and thymus defects. Although the *Pax9* heterozygotes were perfectly normal, resembling a wild-type, the *Pax9*$^{-/-}$ mutants possessed numerous defects and die post-natally. They showed lack of pharyngeal arch derivatives - thymus, parathyroid glands and ultimobranchial bodies; absence of all teeth (molars & incisors), palatal processes of mandible and maxilla, cleft secondary palate (neural-crest derived structures), displayed preaxial polydactyly of both fore- and hind-limbs and tympanic ring was hypoplastic (inner ear). However, there were no vertebral column defects observed. The *Pax9*$^{-/-}$ mice die shortly after birth exhibiting difficulties in respiration and a bloated abdomen [74, 75].

Notably, the facial and dentition defects are unique to *Pax9* and were not found in *Pax1*$^{-/-}$, which corroborates with the distinct expression sites of *Pax9* in these neural-crest derived structures. Although defects in the thymus were not reported by the authors for the *Pax1*-targeted null mice, thymus size reduction has been reported for the *Pax1* spontaneous mouse mutants [72].

Since both *Pax1* and *Pax9* belonged to the same sub-family, and there was an absence of vertebral column defects in the *Pax9*$^{-/-}$ mice, it was hypothesized that both *Pax* genes may have redundant roles in axial skeletogenesis. This prompted the generation of multiple allele knock-outs through the inter-mating of *Pax1*$^{+/-}$ and *Pax9*$^{+/-}$ mutants.

An analysis of the various *Pax1/Pax9* mutants revealed the synergistic roles played by *Pax1* and *Pax9* in the vertebral column. Conforming to their redundant roles, a clear gene-dosage effect was observed; with the loss of more alleles of *Pax1/Pax9*, the vertebral column malformations were more severe. Furthermore, in the vertebral column, while *Pax9* could only partially compensate for the loss of *Pax1*, *Pax9* deficiency was fully rescued by *Pax1*. The most severe axial skeleton defect was seen in the $Pax1^{-/-}Pax9^{-/-}$ (double-null) mutants, whereby there was a complete loss of vertebral bodies and IVDs, deformed proximal parts of the ribs and the mice died post-natally. These axial skeleton defects were more drastic than those seen in the single-null mutants, thus demonstrating the synergistic functions of *Pax1* and *Pax9* [62, 73, 74].

In the double-null mutants, besides the vertebral column, malformations in the other organs / structures were exclusive to the loss of either *Pax1* or *Pax9*. The preaxial polydactyly, cleft secondary palate and lack of ultimobranchial bodies (4^{th} pharyngeal arch derivative) seen in the double-null mutant was also seen in $Pax9^{-/-}$ mice. Likewise, defects in the scapula seen in the $Pax1^{-/-}$ mice were also observed in the double-null mutants. Abnormalities of the thymus glands, however, were not reported in the study owing to their primary focus being the axial skeleton [62, 73, 74]. All the phenotypes of the *Pax1* and *Pax9* targeted mutants are summarized in Table 3.

Thus, it was hypothesized that *Pax1* and *Pax9* are not needed for sclerotome formation per se, but are imperative to maintain the proliferative capacity of the sclerotomal cells, enough for a critical size of mesenchymal condensation to form, upon which endochondral ossification can successfully take place. Indeed, the vital role of *Pax1* in controlling cell proliferation is apparent through its genetic interaction with another TF - *Mfh1*. Also expressed in the sclerotome, *Mfh1* has been shown to

synergize with *Pax1* to regulate the mitotic activity of sclerotomal cells [41]. Similarly, several other *Pax* genes like *Pax5* and *Pax6* are well-known to regulate proliferation of B cells and diencephalic precursor cells respectively [81, 82]. Therefore, the *Pax* genes in general may possess a conserved role of regulating cell proliferation.

While proliferation might be an early role of the *Pax1/Pax9*, they evidently have a late function in IVD development. Therefore, other hypotheses are that these two *Pax* genes may regulate other processes or factors essential for the condensation process, such as cell shape, cell size, cell adhesion and junction, and ECM production and degradation [62, 71]. Also, as mentioned earlier, *Pax1* and *Pax9* have been shown, *in vitro*, to directly bind to the promoter and trans-activate *Bapx1*, another TF known to be critical for the proper differentiation of prechondroblast into chondrocytes in axial skeletogenesis [53, 57]. This led to the hypothesis that both *Pax* genes may have roles in early chondrogenesis and that they are crucial for the early stages of axial skeleton formation.

Currently, besides *Bapx1* as one of the potential early targets, other down-stream targets of *Pax1* and *Pax9* are not known. Therefore, identification of the target genes of *Pax1* and *Pax9* during early stages of sclerotome differentiation and their late functions in IVD development will help illuminate their functions in axial skeletogenesis.

Table 3: Summary of *Pax1* and *Pax9* targeted mouse mutant phenotypes

Pax Mutants	Phenotype	Survival & references
Pax1+/-	- Mild defects in axial skeleton: ▪ Fused atlas-axis (C1 & C2 cervical vertebrae) ▪ Lumbar vertebrae ▪ Sternum - Semi-dominant haploinsufficient	Viable & fertile [73]
Pax1-/-	More severe defects in axial skeleton: - Fused atlas-axis (C1 & C2 cervical vertebrae) - Lumbar vertebrae – VBs & IVDs - Sternum - Scapula (Pectoral girdle - acromion) - Pelvic girdle - Tail → short & strongly kinked - Lack derivatives of 3rd & 4th pharyngeal pouches (thymus, parathyroid glands) - Kinked tails (more severe)	Viable & fertile [73]
Pax9+/-	No axial skeleton defects; Mice are phenotypically normal	Viable & fertile [74]
Pax9-/-	- Cleft secondary palate - Absence of teeth (incisors & molars) - Absent palatal processes of premaxilla & coronoid process of the mandible - Hypoplastic tympanic ring (ear) - Preaxial polydactyly (ectopic cartilage formation) of fore- & hind-limbs - Lack derivatives of 3rd & 4th pharyngeal pouches (thymus, parathyroid glands) - Distended abdomen at birth; no feeding - No vertebral column abnormalities	Die shortly after birth; gasping respiration & bloated abdomen [74, 75]

Pax1-/- Pax9-/-	- No vertebral bodies or IVDs - Missing: ▪ Proximal parts of most ribs, ▪ all skeletal elements of tail, ▪ connection between sacrum & pelvic girdle - ectopic dorsal cartilage formation	Post-natal lethality (no embryonic lethality) [62]

1.5.2.1 Pleiotropic roles of *Pax1* and *Pax9*.

Similar to several other TFs, *Pax1* and *Pax9* have pleiotrophic roles, in that they are important for the development of other structures besides the vertebral column. *Pax1* is also necessary for scapula, pelvic girdle and thymus development. Although the thymus was not investigated by the authors in the *Pax1*-targeted null mutants, analysis of *Pax1 undulated* mutants (*un, un*^{ex} and *un*^{s}) had revealed a reduction in the size of thymus and also altered maturation of thymocytes.

Similarly, *Pax9* is critical for digit, teeth, craniofacial bones (mandible and maxilla), thymus and parathyroid gland development, as seen from the defects in all of these structures in the *Pax9*$^{-/-}$ mutants (Table 3).

Thus, these TFs expression sites correlate with their roles in numerous organs / structures. Evidently, the developmental pathways must be different for the various structures, and it is highly likely that they regulate such distinct organ development through cooperation with varied partners in the different tissues.

1.5.3 *Pax1/ Pax9* related defects in humans

The importance of *Pax1* and *Pax9* can also be perceived through the emergence of vertebral column anomalies in humans, associated with the counterpart human *PAX1* and *PAX9* genes. Skeletal defects such as Jarcho-Levine syndrome [83], Klippel-Feil syndrome [84] and kyphoscoliosis or variants of such

vertebral segmentation [85, 86] defects have been associated with the *PAX1* and *PAX9*.

In brief, Jarcho-Levine syndrome is a lethal, autosomal recessive developmental disorder, comprising a variety of vertebral and rib deformities. The associated phenotypes include a “crab-like” thoracic spine and ribs, fused vertebrae or ribs, kyphoscoliosis and short stature, which may occur in conjunction with cardiovascular disorder, renal defects or neural tube anomalies. The axial skeleton phenotypes similar to that of *Pax1/Pax9* compound mice mutants have been identified in human foetuses suffering from the Jarcho-Levine syndrome [62, 83]. A significant decrease in the PAX1 and PAX9 proteins were detected in the vertebral column of the autopsied foetuses. Morphologically, the thoracic and lumbar vertebrae were defective with fusions and irregular shape [83].

Thus, *Pax1* and *Pax9* have conserved roles in axial skeletogenesis in mouse and humans. Moreover, the high similarity of the paired-domain sequence between murine and human *Pax* genes indicates the suitability of the mouse as a model system to study such developmental disorders.

1.6 Research Aims, Strategy and Significance

1.6.1 Objective

With the ultimate objective of putting together a comprehensive and accurate gene regulatory network (GRN) of the embryonic chondro-osteogenic pathway, this study is specifically centered on unravelling a portion of this GRN by experimenting on two of the genes, *Pax1* and *Pax9*. The objective of this study is to identify the target genes of *Pax1* and *Pax9* in the IVD, using traditional gene-targeting strategies (loss-of-function study) and current genomic technologies (microarray and chromatin

immunoprecipitation-sequencing (ChIP-SEQ)). The information derived from these genome-wide techniques can then be linked to the existing chondro-osteogenic pathway, and thus help in the construction of the GRN. From that, we also aim to understand the roles of *Pax1* and *Pax9* in IVD development and decipher how they are linked to the other TFs involved in the MSC commiment towards the osteo-chondrogenic lineages.

Specific Aims

Identify:

1. The target genes of *Pax1* in the early stages of IVD development, in a cell-type specific manner.
2. The common set of *Pax1* and *Pax9* targets that cannot be picked up in the single-null mutants due to the redundant roles of these paralogous genes.
3. The corresponding direct binding targets of Pax1 and Pax9 in the IVD development.
4. Potential link between *Pax1* and *Pax9* and the other TFs in the chondro-osteogenic pathway.

1.6.2 Strategy

(I) Generate *Pax1* and *Pax9* WT and knock-out (KO) mouse-lines for use in gene expression profiling and TF mapping studies to identifying the target genes of *Pax1* and *Pax9*

A traditional transgenic approach was adopted to KO these two genes in the mouse using BAC templates and gene targeting in mouse embryonic stem cells (mESCs) cells by homologous recombination. *Pax1* and *Pax9* were knocked-out

(KO) by insertion of EGFP in order to enrich for *Pax1* and *Pax9* cell lineages (using Fluorescence-Activated Cell Sorting, FACS) specifically from the vertebral column, which were subsequently used for gene expression profiling profiling. Also, for comparison with WT cells, EGFP was co-expressed with Pax1 via the IRES or F2A peptide concatenating strategies. This way, cells producing WT Pax1 and EGFP from the *Pax1* locus were enriched from the vertebral column using FACS, and subsequently used on microarrays for gene expression analyses.

For ChIP-Seq, there are only two ChIP-grade antibodies available (for Pax1 and Pax9 each) commercially at the moment. To overcome potential cross-reactivity issues with these commercial antibodies, the Pax1 and Pax9 proteins were endogenously tagged with hemagglutinin (HA) epitope in mice. These mouse lines were generated as an alternative in case the gene-specific antibodies were not sensitive and specific enough. Moreover, the HA epitope was chosen as several commercial ChIP-grade HA-antibodies are available.

(II) Specific Aim 1: Identify targets regulated by *Pax1* in the early stages of IVD development

Gene expression profiling profiling of *Pax1* (*Pax1*$^{E/E}$ vs *Pax1*$^{-/-}$) was performed on cells sorted by FACS from E12.5 and E13.5 vertebral column, stages when the IVD anlagen is being formed.

(III) Specific Aim 2: Find the common set of *Pax1* and *Pax9* targets that cannot be picked up in the single-null mutant owing to the redundant roles of these paralogous genes

In the presence of *Pax1*, *Pax9*$^{+/-}$ and *Pax9*$^{-/-}$ show no vertebral defects, probably because of compensation by *Pax1* in the vertebral column. Therefore, it is

unlikely that any significant number of *Pax9* regulated genes in the axial skeleton will be picked up in the array comparison of WT vs *Pax9*$^{-/-}$. However, the *Pax1*$^{-/-}$*Pax9*$^{+/-}$ (three allele knock-out) and *Pax1-/-Pax9-/-* (double-null) mice exhibit a progressively more severe phenotype of the vertebral column than the single-null mutants, indicating an obvious gene-dosage and redundant effects of these two genes in axial skeleton formation.

Hence, to identify the target genes of *Pax9* that were obscured by *Pax1* owing to the redundant roles, the *Pax1*$^{-/-}$*Pax9*$^{+/+}$ vs *Pax1*$^{-/-}$*Pax9*$^{-/-}$ and *Pax1*$^{-/-}$*Pax9*$^{+/-}$ vs *Pax1*$^{-/-}$*Pax9*$^{-/-}$ mutants were analyzed. This helped to identify the genes regulated by 2 copies and 1 copy of *Pax9* in the absence of *Pax1*. Moreover, the double-null (*Pax1*$^{-/-}$ *Pax9*$^{-/-}$) vs WT enabled the identification of the whole array of target genes of *Pax1* and *Pax9* that were obscured by their redundancy.

In the double-null embryos, the first signs of abnormalities in cell proliferation and apoptosis of sclerotomal cells and an arrest of chondrogenesis was observed at E12.5, so the gene expression profiling was performed at E12.5.

A potential rate-limiting step in this specific aim was to obtain sufficient cells for gene expression profiling, as the chances of getting a *Pax1/Pax9* double-null embryo from mating the *Pax1*$^{+/-}$*Pax9*$^{+/-}$ (double heterozygotes) is 1 in 16 (based on the Mendellian ratio of inheritance). As the triple allele mutants do not survive (*Pax1*$^{-/-}$ *Pax9*$^{+/-}$ undergoes early postnatal lethality; *Pax1*$^{+/-}$*Pax9*$^{-/-}$ also undergo postnatal lethality), they could not be used for mating.

<u>(IV) Specific Aim 3: Identify the direct binding targets of Pax1 and Pax9 in the early stages of IVD development</u>

To distinguish the direct from the indirect targets of *Pax1* and *Pax9*, ChIP-Seq was performed on E12.5 and E13.5 CD1 WT mouse embryonic tissues (enriched only for the vertebral column) using commercial anti-Pax1 and anti-Pax9 antibodies, with input chromatin as background control. By overlapping the TF mapping data and gene expression profiling data (WT vs *Pax1*$^{-/-}$*Pax9*$^{-/-}$), the direct and indirect targets of *Pax1* and *Pax9* were distinguished.

To overcome potential cross-reactivity issues with commercial antibodies against Pax1 and Pax9, and also because there are only two ChIP-grade antibodies available (for each gene-product) commercially at the moment, these two genes were also endogenously tagged with hemagglutinin (HA) epitope sequence in mice. This was meant to serve as an alternative if those Pax-specific antibodies were not sensitive and specific enough. Moreover, HA epitope was chosen as there are several commercial ChIP-grade HA-antibodies available.

(V) Specific Aim 4: Potential link between *Pax1* and *Pax9* and the other TFs in the chondro-osteogenic pathway

The list of direct and indirect targets of *Pax1* and *Pax9* identified through the gene expression profiling and TF mapping studies were further scrutinized manually to identify any of the known TFs involved in the chondro-osteogenic pathway. This served as the link between the *Pax1*/*Pax9* TFs and the other TFs in the chondro-osteogenic pathway. Furthermore, literature search and existing gene expression data for other TFs (eg. *Sox5*, *Sox6*, *Sox9*) in the chondro-osteogenic pathway (generated by other researchers in the lab via the same strategies for the same embryonic stages and tissue) were mined for elucidating the relationship between the *Pax1/Pax9* and the other TFs.

1.6.3 Significance

Mesenchymal stem cells (MSCs) are multi-potent progenitor cells that possess the ability to differentiate into osteoblasts, chondrocytes, myocytes, adipocytes, tendon cells, fibroblasts or neuronal cells [87]. Understanding the mechanisms by which these multi-potent stem cells differentiate into specific lineages is essential for therapeutic applications in the area of regenerative medicine, such as reprogramming of patient-specific somatic cells or tissue engineering. The chondro-osteogenic lineage is of immense biomedical importance as it gives rise to most of the skeletal components of the body such as the axial, appendicular and some parts of craniofacial bones. Comprehending how the MSCs are patterned and become committed into chondrocytes or osteoblasts to form the skeleton will also shed light on the basis for congenital diseases/ syndromes and degenerative disorders of the bone like Klippel-Feil syndrome, Jarcho-Levine syndrome, spondylocostal dysostosis, Campomelic dysplasia, osteopenia, osteoporosis, arthritis, intervertebral disc degeneration and osteoporosis pseudoglioma (OPPG).

More importantly, identification of the target genes of *Pax1* and *Pax9* will help to illuminate the early events of regulation involved in the commitment of MSCs towards the annulus fibrosus fate in the IVD. Furthermore, by elucidating the network of chondro-osteogenic cell, we can also predict the outcome of various perturbations to the system. This would enable us to design better and more appropriate therapies for various bone diseases.

CHAPTER 2 – MATERIALS AND METHODS

2.1 BAC Modification and Subcloning

C57BL/6J mouse strain-derived bacterial artificial chromosomes (BAC) clones RP24-88N2 and RP24-211J10 containing the *Pax1* (mouse chromosome 2) and *Pax9* (mouse chromosome 12) gene loci respectively, were obtained from the BACPAC Resources Centre at Children's Hospital Oakland Research Institute (CHORI). BAC clones were confirmed by polymerase chain reaction (PCR) and sequencing before proceeding with gene manipulation. All the primers for BAC screening, modification and subcloning were designed using the Primer 3 (v. 0.4.0) web-based software (http://frodo.wi.mit.edu/primer3/). Genetic modifications of the clones were performed using the Gene Bridges Quick and Easy BAC Modification kit (Cat #K001) through the Red/ET recombineering technology, according to the manufacturer's protocol (Figures 2- 5). Briefly, appropriate 50 bp homology arms flanking the cassettes (*F2A-EGFP-FRT-PGKgb2-Neo-FRT*; *HA_3-TGA-loxP-PGKgb2-Neo-loxP*) were added via PCR. *pRed/ET* plasmids were first transformed into the *E.coli* strain DH10B containing the required BAC clone with selection by chloramphenicol (12.5 µg/mL) and tetracycline (3 µg/mL) antibiotics. The cassettes were subsequently transformed into the *E.coli* containing the *pRed/ET* plasmids. Colonies with successful homologous recombination were selected by their resistance to chloramphenicol (12.5 µg/mL) and kanamycin (20 µg/mL) antibiotics and screened by colony PCR. Positive clones were further screened for errors such as point mutations or deletions by sequencing the PCR products of the inserted cassette with multiple overlapping primer sets. Mutation-free, successfully modified BAC clones were subcloned into a minimal vector using the Gene Bridges Quick and Easy BAC Subcloning kit (Cat # K003) using the same Red/ET recombineering technology as per the manufacturer's protocol (Figure 5). Grabbing arms were

chosen in a repeat-free region, with a short arm of at least 1.5kb and a long arm of at least 7kb. A *PmeI* restriction site was added at one end of the grabbing arms to facilitate plasmid linearization for electroporation later. The selection of the positive recombinants (subclones) was performed with kanamycin (20 µg/mL) and ampicillin (100 µg/mL) antibiotics and screened by colony PCR. Clones were further confirmed for absence of errors by sequencing the PCR products of the modified region with multiple overlapping primer sets. Mutation-free subclones were linearized with PmeI restriction enzyme and electroporated into mouse ES cells.

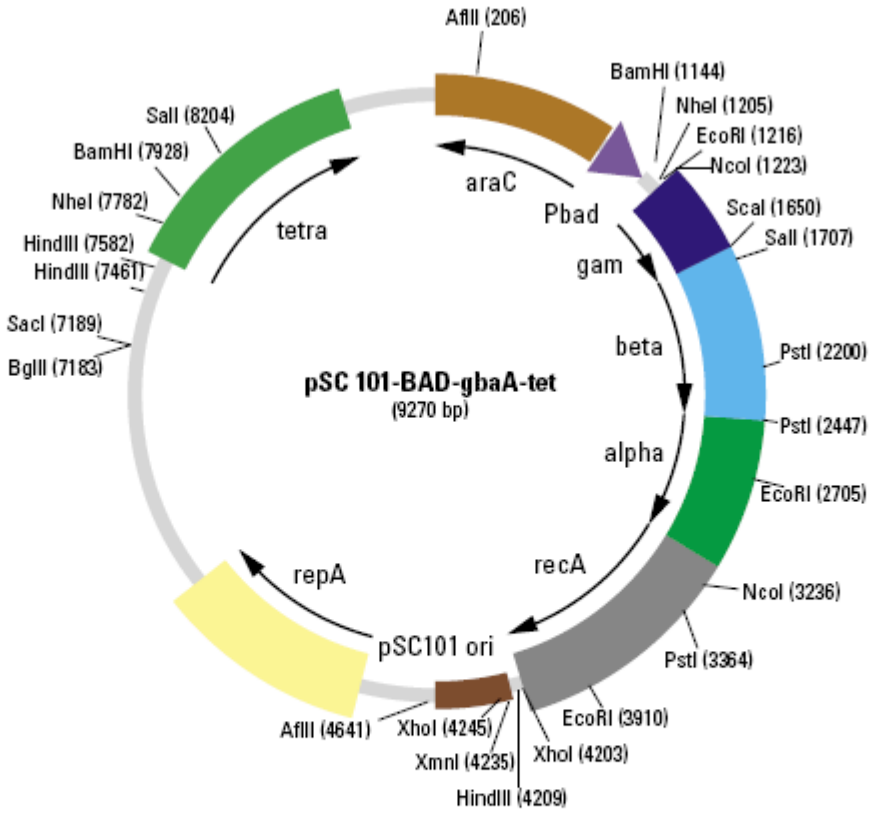


Figure 2: Map of the Red/ET expression plasmid pSC101-BAD-gbaA^{tet}. Transformation of *E.coli* hosts with this plasmid is selected for by acquisition of tetracycline resistance at 30°C. Expression of the Red/ET recombination proteins is induced by L-arabinose activation of the BAD promoter at 37°C. Picture adapted from Gene Bridges Quick and Easy BAC Modifi cation Kit By Red®/ET® Recombination Version 2.4 (February 2005) technical protocol.

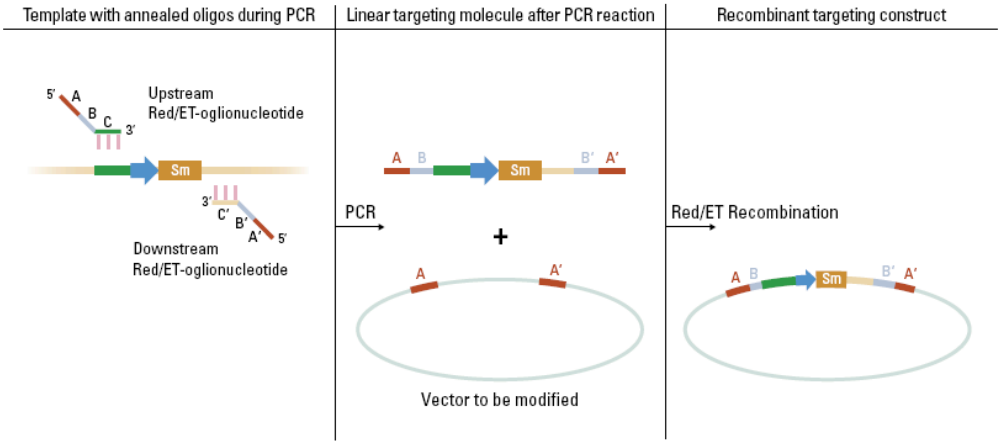


Figure 3: Illustration of the principle for modifying bacterial artificial chromosomes (BACs). *Left:* Cassettes with selectable marker gene are first synthesized using appropriate primers with homology arms by polymerase chain reaction (PCR). *Middle:* PCR product (cassette) is electroporated into host *E.coli* cells containing the BAC (vector) to be modified. *Right:* The BAC is thus modified by Red/ET mediated recombination. *Sm* - selectable marker; the *small blue arrow* - prokaryotic promoter. Picture adapted from Gene Bridges Quick and Easy BAC Modifi cation Kit By Red®/ET® Recombination Version 2.4 (February 2005) technical protocol.

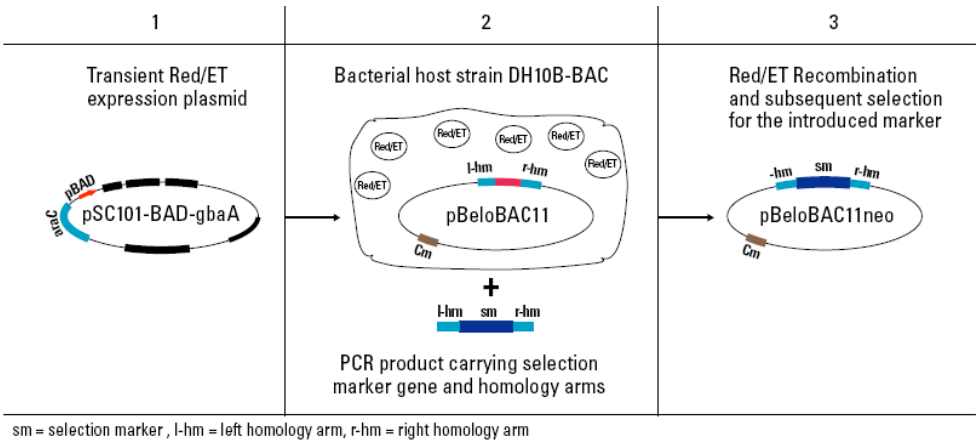


Figure 4: Mutagenesis strategy for inserting the cassette-of-interest into a bacterial artificial chromosome (BAC).1: Red/ET expression plasmid is first electroporated into the DH10B *E.coli* strain containing the BAC-of-interest. 2: Cassette with a selectable marker gene is then electroporated into cells containing the Red/ET plasmid. 3: Presence of selectable markers allows the identification of colonies with modified BAC. Picture adapted from Gene Bridges Quick and Easy BAC Modifi cation Kit By Red®/ET® Recombination Version 2.4 (February 2005) technical protocol.

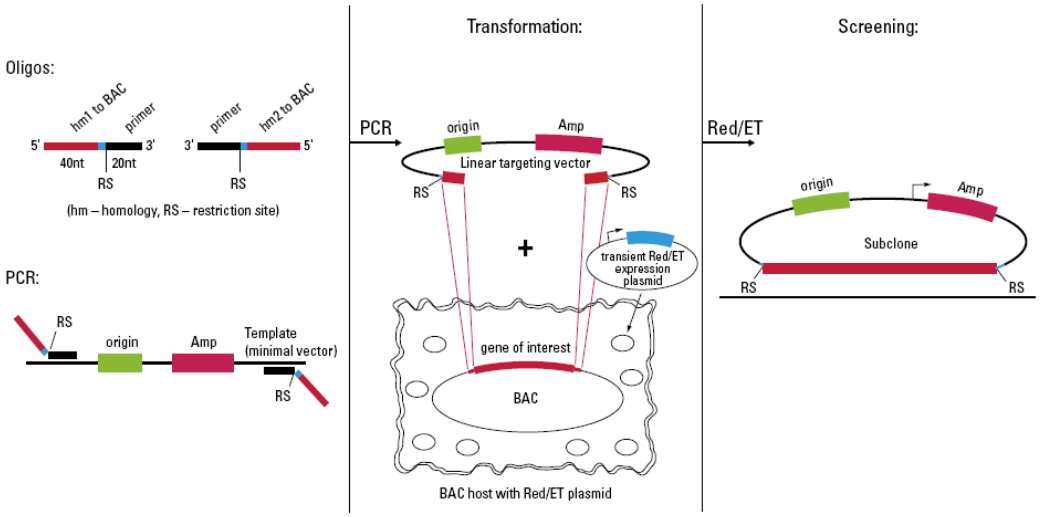


Figure 5: BAC subcloning by recombineering technology. Left: Primers with homology to region-of-interest for subcloning is synthesized first. Using the plasmid template provided in the kit for the minimal vector containing the ampicillin selection marker, PCR products are generated. Middle: Red/ET expression plasmid is first electroporated into the clones with modified BAC. The PCR products for subcloning are then electroporated into the cells. Right: Through homologous recombination, the region-of-interest is subcloned into the minimal vector which contains the ampicillin selectable marker. Presence of selectable markers (ampicillin + kanamycin) allows the identification of colonies with modified BAC that have been successfully subcloned. Picture adapted from Gene Bridges Quick and Easy BAC subcloning Kit By Red®/ET® Recombination Version 2.4 (February 2005) technical protocol. Hm – homology arm; RS – restriction site; Amp – ampicilin.

2.2 Homologous Recombination in Mouse ES Cells

2.2.1 ES Cell Culture

Either the R1 (derived from 129X1 x 129S1 strain) or the V6.4 (hybrid C57BL/6J x 129) mouse ES cells (p17 - p20) were used for gene targeting of the different constructs. All the mouse ES cells were grown on gelatinized plates (0.1%) containing irradiated mouse primary embryonic fibroblast (PEF). Cells were grown in ES media (ESM) composed of DMEM supplemented with 15% heat-inactivated ES grade fetal bovine serum (FBS) (Gibco, Invitrogen), 0.1 mM β-mercaptoethanol, 4 mM L-glutamine, 1 mM sodium pyruvate, 40 ug/mL gentamicin, and 500 U/mL LIF (ESGRO, Chemicon) at 37°C in 5% CO_2. Fresh media was added on alternate days and cells were passaged with 0.05% trypsin (Gibco, Invitrogen) every 3 days.

2.2.2 Electroporation of ES Cells

Mouse ES cells were passaged 24 hrs before electroporation. For electroporation, cells were trypsinized (0.05% trypsin for 3 min at 37^{o}C), washed with ESM (3 min, 1000 rpm, RT) and counted using a haemocytometer. 5 million or 10 million cells, for R1 and V6.4 cells respectively, were used per electroporation. Cells were resuspended in 500 uL of ESM with 10 ug of linearized subclone (targeting vector) or without any DNA for negative control. Cell were then electroporated using an electroporation cuvette (Biorad, 1 mm gap) at 125 µFarads, 0.4 kVolts. Electroporated cells were allowed to rest for 10 min at RT before being split equally into six 10 cm gelatinized tissue culture dishes containing DR4 feeder (Neomycin resistant). Cells were grown overnight in ESM before selecting for successfully recombined clones by G418 selection. A range of G418 concentrations were used for the selection process: 150 ug/mL, 200 ug/mL, 250 ug/mL, 300 ug/mL, 350 ug/mL and 400 ug/mL. Negative controls were selected at only two concentrations: 150 ug/mL and 400 ug/mL.

2.2.3 ES Cell Colony Picking

G418 selection of positive recombinants was carried out for 8-10 days. Negative control plates contained no colonies by the end of the 8th day of G418 selection. 192 colonies that survived the selection from DNA-electroporated plates were picked into two 96-well gelatinized plates containing irradiated PEFs. Colonies were allowed to grow overnight in ESM before trypsinization and being split into two 96-well replicate plates. Cells were allowed to recover for 2 days in ESM before being split again. One 96-well plate was split into two fresh 96-well plates with irradiated PEFs and grown to 80% confluency before being frozen down at -150^{o}C; the other replicate was split into 24-well plates for subsequent DNA extraction.

Colonies were then pre-screened by PCR, followed by Southern blotting using the appropriate external and internal DIG-labelled probes.

2.2.4 ES Cell Cryopreservation

ES cells in 96-well plates were frozen down using freshly prepared, filtered, freezing medium (70% DMEM, 20% FBS and 10% DMSO). ESM was first removed from the wells and replaced with 70 uL of freezing medium. Plates were sealed with cling wrap and aluminium foil, and placed in a Styrofoam box for freezing in -80^{o}C overnight before further storage at -150^{o}C. Positive clones identified from pre-screening by PCR or Southern blotting were thawed at 37^{o}C, expanded, trypsinized and washed with ESM before resuspension in 500 uL of freezing medium in cryo-vials. The cryo-vials were then placed in freezing containers (Invitrogen; allows a drop in temperature at a rate of 1^{o}C per min) and frozen slowly at -80^{o}C overnight before further storage at -150^{o}C.

2.3 ES Cell Clone Screening

2.3.1 Genomic DNA Extraction

ES cells and mouse tail tips were digested with Proteinase K (Sigma; at a final concentration of 0.5 mg/mL in PK digestion buffer (PKDB)) overnight with agitation at 37^{o}C and 57^{o}C respectively. Phenol: Chloroform extraction of genomic DNA was performed using MaXtract High Density tubes (Qiagen; Cat.#129046. Equal volume of phenol: chloroform: isoamyl alcohol (pH7.9) (Ambion; Cat# AM9732) was added to the digested sample and vortexed for 30 seconds before spinning down at 13200 rpm for 5 min to separate the organic and aqueous phases. The DNA-containing aqueous layer was decanted into a fresh 1.5 ml microcentrifuge tube and was precipitated with twice the volume of 100% ethanol and subsequently

washed with 70% ethanol. DNA pellet was air-dried before re-suspension in sterile water (molecular grade).

PK Digestion Buffer (PKDB)	
Final concentration	**Reagents**
50 mM	TRIS, pH 7.0-8.0
5 mM	EDTA
1%	SDS
0.2M	NaCl

TRIS- tris(hydroxymethyl)aminomethane ; EDTA- Ethylenediaminetetraacetic acid; SDS- Sodium dodecyl sulfate ; NaCl – sodium chloride

2.3.2 Southern Blotting

- **Southern Probe Design**

Vector NTI software (from Invitrogen) was used to identify the appropriate homology arms on the wild-type (WT) and modified allele to be used for Southern blot screening. Corresponding restriction sites were also identified for each construct. In brief, short (at least 1.5 kb long) and long homology arms (at least 7 kb long) were chosen in a repeat-free region. The UCSC genome browser was used to identify repeat regions (http://genome.ucsc.edu/cgi-bin/hgGateway). External and internal Southern probes (of 400 bp – 1000 bp long) were designed to be outside or inside the homology arms respectively, but within the selected restriction sites (Figures 7-14). External Southern probes were utilized to identify the clones that had been successfully recombined, while internal probes were designed to differentiate clones with random integration of the cassettes as well. Primer 3 (v. 0.4.0) web-based software (http://frodo.wi.mit.edu/primer3/) was used to design the primers for synthesizing the external and internal probes.

- **DIG-labelled Probe Synthesis**

Southern DIG-labelled probes were synthesized by PCR the PCR DIG Probe Synthesis kit (Roche, Cat #1636090). The kit allows the addition of non-radioactive digoxygenin (DIG) by incorporating DIG-dUTP into the nucleotide sequence during the PCR. The PCR products were column purified using QIAGEN PCR purification kit (Cat #28106) and eluted using sterile water. Probe concentration was quantitated using Nanodrop Spectrophotometer and the quality was assessed by gel electrophoresis. Probes were stored at -20°C.

- **DNA digestion and transfer**

Phenol: chloroform purified genomic DNA from the ES colonies were digested with the appropriate restriction enzymes for 16 hrs. The digested DNA (10 - 12 ug) was resolved on 0.8% TAE agarose gels (without ethidium bromide) slowly at a constant voltage of 25 V for 12 - 14 hrs. The gels were then stained with ethidium bromide (10 ug/mL) for 15 - 30 min and imaged under UV light (SYNGENE gel Bio Imaging System). The gels were subsequently denatured using 0.5 M NaOH with gentle agitation for 2 x 30 minutes. For DNA bands of more than 12 kb, depurination with 0.25 M HCl was performed for 10 min before proceeding with NaOH denaturation. Stacking of the gel for Southern blotting by capillary force was done as shown in Figure 6. The gel was watered with 0.5 M NaOH every 1 hr before sealing with cling wrap for overnight transfer.

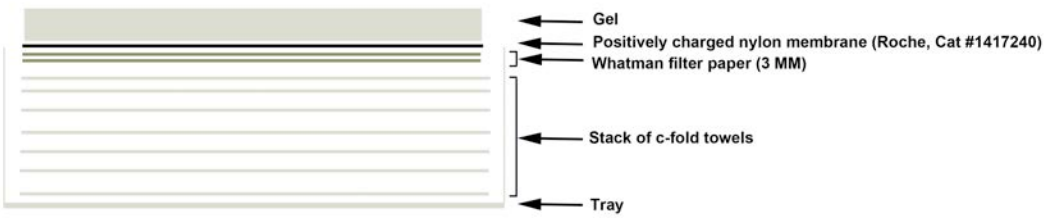


Figure 6: Illustration of stacking of the agarose gel for Southern blotting. The denatured gel is placed on a stack of C-fold towels with two sheets of Whatman 3 MM filter paper and a single positively charged nylon membrane.

- **Hybridisation and Washing**

After the overnight transfer, the nylon membrane was washed in 5X SSC for 10 min and placed in a roller bottle containing pre-warmed DIG Easy Hyb buffer (Roche, Cat #1603558; 10 mL/100 cm^2) for pre-hybridization at 42°C for 4 hrs. The DIG-labelled probes (external or internal probes) were denatured at 99°C for 3 min and quick-chilled on ice for 1 min before being added to pre-warmed DIG Easy Hyb buffer at a final concentration of 25 ng/mL. The membrane was hybridized overnight at 42°C in a rotating oven. After hybridization, the nylon membrane was washed once at RT (RT) for 10 min (2X SSC, 0.1% SDS), then twice at 60°C for 15 min each in a pre-warmed higher stringency wash buffer (0.5X SSC, 0.1% SDS) with vigorous agitation, followed by a brief rinse in 1X MABT.

- **Blocking and Detection**

The membrane was next blocked with 1% Blocking buffer (Roche Blocking Reagent, Catalog #11096176001; 1g/100mL of 1X MAB) for 30 min. The membrane was then incubated for 30 min at RT with alkaline-phosphatase (AP)-conjugated anti-DIG antibody (Roche, Catalog #11093274910) at a 1:10,000 dilution in 1% Blocking buffer with gentle agitation. Unbound antibody was washed off with 1X MABT for 15 min twice at RT with vigorous agitation. The membrane was rinsed briefly in detection buffer (0.1 M Tris-HCl, pH9.5 and 0.1 M NaCl) and incubated with CDP-*Star* chemiluminescent AP substrate (1 mL/100 cm^2; Roche Cat #12041677001) at RT for 5 min. The membrane was then warmed at 37°C for 10 min before signal detection on an X-ray film. The film was generally exposed for 1 hr in order to detect a reasonable signal.

- **Stripping and Re-probing**

Once probed with the external probe, the nylon membrane was stripped and re-probed with the internal probe. For stripping, the membrane was rinsed with deionized water for 1 min at RT, followed by washing at 37^{o}C for 2 x 15 min in pre-warmed stripping buffer (0.2 M NaCl/ 0.1% SDS) with vigorous agitation. The membrane was then washed with 2X SSC before pre-hybridization and hybridization with another probe.

Maleic acid buffer (MAB) 10x	
Final concentration	**Reagents**
0.2 M	Maleic acid
0.3 M	NaCl
Adjust pH to 7.5 with NaOH	NaOH pellets

Maleic acid buffer with Tween (MABT) 1x	
Final concentration	**Reagents**
1x	MAB
0.3%	Tween-20

2.4 Generation of Transgenic Mice

2.4.1 Ethics statement

All the animal procedures were performed according to the Singapore A*STAR Biopolis Biological Resource Center (BRC) Institutional Animal Care and Use Committee (IACUC) guidelines. The IACUC protocols were reviewed and approved by the aforementioned committee before any animal procedures were undertaken for this study (IACUC Protocol No: 110689 and 110648).

2.4.2 Microinjection of ES Cells

Correctly targeted ES cell clones (without secondary random intergration) where thawed from cryo-vials, washed with ESM and cultured for 2 days before being passaged 24 hrs prior to microinjection. On the morning of microinjection, cells

were trypsinized, washed with PBS and resuspended in M2 medium. In each 2- to 8-cell stage mouse embryos harvested from C57BL/6J mice, about 8-10 ES cells were microinjected [88]. On the same day, 6-10 microinjected embryos per oviduct were re-implanted in CD-1 pseudopregnant mice. This method of microinjection allowed the generation of high percentage germline transmitting chimeras. All the ES cell microinjections, mouse maintenance and embryo harvesting were performed by Dr Petra Kraus.

2.4.3 Breeding and Genotyping of Transgenic Mice

The male chimeras, identified by coat-colour, were crossed with female C57BL/6J mice to obtain heterozygous transgenic mice. Homozygous mice were subsequently generated by mating the heterozygotes. Mice were weaned at 3 weeks after birth and ear-tagged for identification purposes. The genotypes of the mice were determined by PCR of genomic DNA extracted from mouse tail tips.

2.5 Fluorescence – Activated Cell Sorting (FACS)

2.5.1 Dissociation of Mouse Embryonic Tissue into Single Cells

For *Pax1*, both heterozygous and homozygous null mutants are viable and fertile. Hence, the $Pax1^{-/-}$ embryos were easily obtained by intermating the homozygotes. The $Pax1^{+/-}$ embryos were obtained by mating the $Pax1^{-/-}$ mice with CD-1 WT mice. The embryos were first staged based on the M.H. Kaufmann, Theiler morphological criteria and confirmed for the presence of enhanced green fluorescent protein (EGFP) by viewing under the fluorescence microscope (LEICA M205 FA microscope). The *Pax1* WT tagged with *EGFP* ($Pax1^{+/EGFP}$) embryos were obtained by mating the $Pax1^{EGFP/EGFP}$ with CD-1 WT mice. The vertebral column which showed EGFP expression was dissected from the E12.5 or E13.5 mouse embryos in cold Leibovitz's L-15 Medium (Invitrogen, Catalog #21083027). The dissected tissue was

dissociated manually by pipetting up and down in freshly prepared and sterile filtered (Millipore, 0.2 μm filter disc) dissociation buffer (DB). The dissociated suspension was passed through a 100 um strainer placed on a 50 mL Falcon tube containing an equal volume of 20% FBS in Leibovitz medium (sterile filtered) in order to stop the enzymatic activity. Any clumps of tissue left on the strainer was aspirated with a wide-bored 1 mL pipette tip and dissociate again in more DB and passed through the same 100 um strainer and collected in the same 50 mL Falcon tube. This step was repeated until no clumps of tissue were observed on the strainer. The filtrate was then pelleted at 2000 rpm for 5 min in a 4°C centrifuge. The supernatant was discarded and the pellet was resuspended in the filtered, pre-warmed (to 37^{o}C) resuspension buffer (RB). For every vertebral column tissue from a single embryo (E12.5 or E13.5), 400 uL of RB was used. The cell suspension was passed through a 40 um strainer and aliquoted into a 5 mL polystyrene tube. In parallel, vertebral column tissue from a CD-1 WT or a littermate WT embryo of the same embyonic stage was dissociated to be used as the gating control for GFP detection threshold. The cells were then sorted by FACSAria flow cytometer (BD Biosciences) and collected into a 1.5 mL microcentrifuge tube containing 500 uL of 20% FBS. Sorted cells were pelleted at 1,400 rcf for 10 min in a 4°C tabletop centrifuge, resuspended in Trizol (Invitrogen Cat#15596-018; 100,000 cells / mL of Trizol), incubated for 5 min at RT and stored at -80°C. Cells can be stored in Trizol for a maximum of 6 months in -80 °C before RNA extraction.

Dissociation buffer (DB)	
Final concentration	**Reagents**
-	Lebovitz L-15 medium
100 U/mL	Collagenase I (Sigma)
100 U/mL	Collagenase II (Sigma)
50 U/mL	Dnase
0.05%	Trypsin

Resuspension buffer (RB)	
Final concentration	**Reagents**
2 %	FBS
2 uM	EDTA (pH 8.0)
25 mM	HEPES-KOH (pH 7.5)
50%	Accumax™
-	Lebovitz L-15 medium

2.6 Microarray Analysis of Gene Expression

2.6.1 RNA Extraction

Sorted cells that had been stored in -80°C were thawed on ice before total RNA extraction was carried out using Trizol followed by column purification with the QIAGEN RNeasy Micro kit (Cat # 74004). Briefly, 0.2 mL of chloroform was added per 1 mL of Trizol and transferred to a 2 mL MaXtract High Density tube (Qiagen; Cat. #129046). After a 15 sec vigorous shaking by hand and incubation at RT for 3 min, the tube was centrifuged at 15,800 rcf for 15 min at 4°C in a table-top centrifuge. The aqueous layer was then transferred to fresh RNase-free 1.5 mL microcentrifuge tube and precipitated with an equal volume of freshly prepared 70% ethanol (about 60% of the initial volume of Trizol sample). Precipitated sample was then passed through RNeasy MinElute column (Qiagen) and centrifuged at RT for 30 sec at 10,000 rpm. The column was washed with 350 ul wash buffer RW1 before on-column DNase treatment was done using 80 uL of 1U/uL DNaseI in buffer (QIAGEN #79254) for 15 minutes at RT. The column was washed with 350 uL buffer RW1, 500 uL RPE and then 500 uL 80% ethanol (made fresh). Thereafter, the column was spun dry at 13,200 rpm for 5 min at RT, and incubated in 14 uL of RNase-free for 1 min before elution at 13,200 rpm for 1 min in a fresh 1.5 mL centrifuge tube. RNA samples were quantified and checked for their integrity using Agilent RNA Pico 6000

Chip (#5067-1513) and Agilent 2100 Bioanalyzer software according to the manufacturer's protocol. RNA was stored at -80°C until further use.

2.6.2 RNA Amplification and Biotin Labelling

Purified RNA samples with RIN value of at least 7.0 were chosen for subsequent processing. RNA samples were amplified using the NuGEN Ovation™ RNA Amplification V2 kit (Cat # 3100-12) and biotin-labeled with NuGEN Encore™ BiotinIL Module kit (Cat # 4210-48) according to the manufacturer's protocol. 2 ng of total RNA was used to synthesize the single-stranded cDNA with a single round of linear PCR amplification using SPIA amplification system. The microgram amounts of cDNA that was prepared this way was purified using QIAGEN PCR purification columns as per manufacturer's instructions. cDNA samples were quantified using the Nanodrop spectrophotometer before 3 ug of cDNA was used for biotin-labeling using the NuGEN Encore™ BiotinIL Module kit. Labeled cDNA were then purified using QIAGEN PCR purification columns and quantified by Nanodrop and stored in -20^{o}C until further use.

2.6.3 Hybridization on Illumina Mouse WG-6 BeadChip

Biotin-labelled cDNA was used for gene expression analysis using the Illumina MouseWG-6 Expression BeadChip (Cat #BD-201-0202). Hybridization, washing and signal development were performed according to manufcaturer's protocol provided in the Illumina MouseWG-6 Expression BeadChip kit. For each sample, 1.5 ug of cDNA was resuspended in 10 uL of nuclease-free water. The sample was incubated at RT for 10 min before 20 uL of GEX-HYB (prewarmed to 58°C and cooled to RT) was added and heated at 65 °C for 5 min. Hybridization chamber and gaskets were assembled and 200 uL of GEX-HCB was added to the humidifying buffer reservoirs. The 30 uL of warmed biotin-labeled cDNA was briefly

vortexed, centrifuged at 13,200 rpm for 1 min and cooled to RT before being loaded onto the microarray chip. Biological replicates for each genotype were loaded in a randomized manner on multiple chips. The loaded BeadChips were placed horizontally inside the hybridization chamber and sealed. The chamber was placed in the pre-warmed hybridization oven and incubated for 18 hours at 48°C according to the recommendations by the NuGEN Ovation™ RNA Amplification V2 kit owing to the use of less stable cDNA/DNA pairs compared to the usual cRNA/DNA pairs. Meantime, the 1X High-Temp wash buffer (500 mL) was prewarmed overnight at 55°C.

The following day, coverseals of the hybridized BeadChips were carefully removed in the Wash E1BC solution. Beadchips were incubated in the 1X High Temp wash buffer at 58°C for 10 minutes, followed by washing in the E1BC solution for 5 min, 100% ethanol for 10 min and E1BC solution for 2 min. Blocking with the Block E1 buffer for 10 min at RT was then performed. Next, Streptavidin-Cy3 labelling was carried out in the dark (2 uL of Streptavidin-Cy3 in 2 ml of Block E1 buffer) for 10 min. Beadchips were then washed in E1BC solution for 5 min, spun dry at 275 rcf for 4 min and finally scanned with the Illumina® BeadArray Reader on the same day.

2.6.4 Gene Expression Analysis using GeneSpring GX 11.0

Illumina® BeadStudio software was used to extract the raw image data from the scanned beadchips. The gene expression data was exported as sample probe profiles in a GeneSpring GX 11.0 compatible text file format, with background subtraction but no normalization. The text file was then imported into GeneSpring GX 11.0 for further gene expression analysis.

For E12.5 and E13.5 *Pax1*-null vs WT & GFP(+) vs GFP(-) comparisons

Criteria for flagging entities were chosen as follows: "present" – detection p-value > 0.8; "absent" – detection p-value < 0.6; and "marginal" – values in between "present" and "absent". Negative raw values were set to a minimum threshold of 1.0 and a 75% percentile shift. The entities were filtered by flags, including only those that fell in the "present" and "marginal" categories in at least one out of total number of samples. A gene-level analysis was performed and the entities were further filtered by raw expression data, including those that were between 20-100th percentile in at least one out of the total number of samples. WT B1 (E12.5) biological replicate was found to be an outlier and hence removed from further analyses. Pair-wise comparison was made with unpaired Student's t-test for $Pax1^{+/E}$ (WT) vs $Pax1^{-/-}$ for E12.5 and E13.5 samples. For GFP(+) vs GFP(-) comparisons, one-way ANOVA statistical testing was performed since it involved the comparison of more than one pair of samples (GFP(-) vs WT, $Pax1^{+/-}$ and $Pax1^{-/-}$). Multiple testing correction was performed on the p-values with the Benjamini Hochberg False Discovery Rate (B-H FDR) and all entities with p-value $\leq$ 0.05 and a fold change of $\geq$ 1.5 were defined as significant.

For multiple allele KO study ($Pax1^{-/-}Pax9^{+/-}$ and $Pax1^{-/-}Pax9^{-/-}$):

For the $Pax1^{-/-}Pax9^{+/-}$ and $Pax1^{-/-}Pax9^{-/-}$ embryos, only 1.1 ug of biotin-labelled cDNA could be produced for each biological replicate from the extracted RNA. This was still within the recommended amount of cDNA for hybridization on Illumina Chips (750 ng -1.5 ug). Owing to time and fiscal constraints new $Pax1^{+/E}$ (WT) and $Pax1^{-/-}$ samples could not be collected. Hence, the cDNA of WT ($Pax1^{+/E}$) and $Pax1^{-/-}$ left-over from the biological replicates used in the prior analysis (E12.5 $Pax1^{-/-}$ vs WT) were biotin-labelled and used in the new microarray chips alongside the multiple-allele KOs, at a constant cDNA amount of 1.1 ug for each sample (3 biological

replicates each). Since the *Pax1*$^{+/E}$ (WT) and *Pax1*$^{-/-}$ samples were stored in the cDNA form at -20^{o}C for only a short period, degradation should be minimal.

The raw intensity data from the new microarray analysis (E12.5 *Pax1*$^{+/E}$, *Pax1*$^{-/-}$, *Pax1*$^{-/-}$*Pax9*$^{+/-}$ and *Pax1*$^{-/-}$*Pax9*$^{-/-}$ - at 1.1ug) and old microarray analysis (E12.5 *Pax1*$^{+/E}$, *Pax1*$^{+/-}$ and *Pax1*$^{-/-}$) were loaded onto GeneSpring GX 11.0 software and quantile normalized. The quantile normalization method successfully mitigated any batch effect and differences owing to the DNA amount used, with any difference between the new and old batch to be defined by just ~19% in the Principal Component Analysis (PCA). Also, combining both data sets improved the statistical strength of the WT and *Pax1*$^{-/-}$ samples (from 3 biological replicates to 8 biological replicates) and hence the accuracy of the results. WT B1 (E12.5) biological replicate was an outlier and was removed from all further analyses.

Again, all entities were filtered by flags (including only those that fell in the "present" and "marginal" categories). A gene-level analysis was performed and the entities were filtered by expression (including only those that were between 20-100th percentile in at least one out of the total number of samples). Multiple pair-wise comparisons were made with Welch, one-way ANOVA (unequal variance) statistical testing, with Benjamini Hochberg False Discovery Rate (B-H FDR) multiple testing correction on the p-values. All entities with p-value $\leq$ 0.05 and a fold change of $\geq$ 1.5 were defined as significant.

Functional annotation clustering was performed on the microarray results using **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) v6.7 (http://david.abcc.ncifcrf.gov/).

2.7 Chromatin Immunoprecipitation – Sequencing (ChIP-Seq)

2.7.1 Tissue Harvesting and Cross-linking

The vertebral column tissues were dissected from staged (based on M.H. Kaufmann, Theiler morphological criteria) E12.5 and E13.5 CD-1 WT mouse embryos (about 50 embryos) in ice-cold Leibovitz medium. The dissected tissues were pelleted at 1,100g for 5 min at 4^0C and resuspended in 10 mL of ice-cold 1xPBS. Tissues were then homogenized using a pre-chilled 15 mL Douncer. Tissues were then washed once with ice-cold 1xPBS and pelleted at 1,100g for 5 min at 4^0C. Pellet was weighed and resuspended in 10x volume of 1x PBS at RT. One-tenth the volume of cross-linking buffer (11% Formaldehyde) was added to the resuspended cell suspension and incubated for 10 min on a nutator at RT. Cross-linking was stopped by adding one-tenth the volume of 2.5 M of Glycine solution. Cross-linked tissue was then pelleted and washed once with ice-cold 1x PBS at 1,100g for 5 min at 4^0C. Cell suspension was homogenized again on the Douncer before one more wash in ice-cold 1xPBS. Cells were then pelleted, weighed and snap frozen in liquid nitrogen before being stored at -80^{o}C until futher use. Cross-linked tissue samples can be stored for up to 6 months in -80^{o}C.

Cross-linking buffer (11% formaldehyde)	
Final concentration	**Reagents**
100 mM	NaCl
50 mM	Hepes-KOH (pH7.5)
1 mM	EDTA (pH 8.0)
0.5 mM	EGTA (pH 8.0)
11%	Formaldehyde
-	Water

2.7.2 Binding of Antibodies to Magnetic Beads

Antibodies of interest (Pax1: SC-25407X (M116X) or Pax9: SC-25410X (H-95 X)) were conjugated to magnetic Dynabeads® Protein G (Invitrogen; Cat. 100.04D).

100 uL of dynabeads were used per immunoprecipitation reaction (IP). The beads were washed thrice in 1 mL of pre-chilled blocking solution (0.5% BSA) and incubated with 5 μg of rabbit IgG (ab46540) in 250 uL of blocking solution and rotated 360° for 6 hrs at 4°C for use in the pre-clearing step later. In parallel, another set of tubes were set up for antibody conjugation with 10 ug of the Pax1 or Pax9 antibodies for 7 - 24 hrs in a similar manner. Lysis of cross-linked cells and sonication of the chromatin was performed during the bead conjugation process.

2.7.3 Cell Lysis, Sonication, Pre-clearing and Chromatin Immunoprecipitation

- **Cell lysis**

The cross-linked samples that were stored in -80°C were thawed on ice and resuspended in 10x the volume of pellet in lysis buffer 1 (LB1; with protease inhibitors; Roche #11697498001) and nutated for 10 min at 4°C. Cell debris was cleared by centrifuging the samples at 1,350 g for 5 minutes at 4°C. Samples were then resuspended in lysis buffer 2 (LB2; with protease inhibitors) at an equal volume as LB1 and incubated for 10 min at RT. Cell nuclei were obtained by pelleting the sample at 1,350 g for 5 minutes at 4°C. Nuclei pellet was weighed again and chromatin was obtained by lysing the nuclei with 10x the volume of lysis buffer 3 (LB3; with protease inhibitors).

- **Chromatin Sonication**

The nuclear lysates were sonicated in a 15 mL bacterial culture tube with 1 mL of glass beads (BioSpec Products; #11079105) per 2-3 mL of lysate using Branson Digital Sonifier® in the cold room. Samples were kept on ice throughout the sonication process (sonication condition: ON: 15sec; OFF/REST: 30 sec; 20 cycles; amplitude: 40%; total effective sonication = 5min; total time = 15min). The chromatin

was sheared to a size range of 100-500bp, and transferred to a 2 mL microcentrifuge tube. Triton X-100 was added to a final concentration of 1% and samples were spun at 13200 rpm for 10 min at 4°C to remove any debris. The chromatin concentration was quantified using the Nanodrop spectrophotometer (Thermo Scientific).

Lysis buffer 1 (LB1) + 1x Protease inhibitors	
Final Concentration	**Reagents**
50 mM	Hepes-KOH, pH 7.5
140 mM	NaCl
1 mM	EDTA
10%	Glycerol
0.50%	Igepal CA630
0.25%	Triton X-100
-	Water

Lysis buffer 2 (LB2) + 1x Protease inhibitors	
Final Concentration	**Reagents**
10 mM	Tris-HCl, pH 8.0
200 mM	NaCl
1 mM	EDTA
0.5 mM	EGTA
-	Water

Lysis buffer 3 (LB3) + 1x Protease inhibitors	
Final Concentration	**Reagents**
10 mM	Tris-HCl, pH 8.0
100 mM	NaCl
1 mM	EDTA
0.5 mM	EGTA
0.10%	Na-Deoxycholate
0.50%	SDS
-	Water

- **Pre-clearing**

The sheared chromatin (2 mg per IP) was precleared in a pre-washed, IgG-conjugated beads for 1 hr at 4°C, rotating at 360°. The volume was kept constant for all IP reactions at 2 mL by topping up with LB3 containing 1% Triton X-100. From the pre-cleared sample, 1% of the volume was reserved as input and stored at -80°C until the de-crosslinking step.

- **Chromatin immunoprecipitation (IP)**

The pre-cleared supernatant was transferred to tubes containing the pre-washed antibody-conjugated beads. Samples were then rotated 360° overnight at 4˚C for immunoprecipitation.

2.7.4 Wash, Elution and Reverse Cross-link

After the overnight IP, the beads were washed with 1 mL of wash buffer for 5 min at 4°C with 360° rotation. The beads were then magnetized, supernatant was discarded, and the wash was repeated 6 more times. Next, the beads were washed once with 1 mL of TE buffer in the same way. The IP chromatin was eluted from the beads by incubation at 65°C for 30 min in 210 uL of elution buffer with vigorous agitation at 1,400 rpm. The eluate (supernatant) was then de-crosslinked overnight at 65°C. The 1% of input pre-cleared chromatin that was reserved earlier was also de-crosslinked in parallel.

Wash Buffer + 1x Protease inhibitors	
Final Concentration	**Reagents**
50 mM	Hepes-KOH, pH 7.6
500 mM	LiCl
1 mM	EDTA
1%	NP-40
0.70%	Na-Deoxycholate
-	Water

TE buffer	
Final Concentration	**Reagents**
10mM	TRIS pH8.0
1mM	EDTA, pH8.0
50mM	NaCl
-	Water

Elution buffer	
Final Concentration	**Reagents**
50 mM	Tris-HCl, pH 8.0
10 mM	EDTA
1%	SDS
-	Water

2.7.5 ChIP DNA Clean Up

To the de-crosslinked samples (including the input sample), 200 uL of TE buffer and RNase A (final concentration of 0.2 ug/mL) were added and incubated for 2 hrs at 37°C to remove all RNA. Samples were then incubated for another 2 hrs with Proteinase K (final concentration of 0.2 ug/mL) at 55°C to remove all protein. The samples were then transferred to MaXtract High Density tubes (QIAGEN Cat.#129046) and an equal volume of phenol: chloroform: isoamyl alcohol (pH 7.9) was added and vigorously agitated for 30 sec. The organic and aqueous phases were thus separated by centrifugation at 15,800 rcf for 5 min at RT. The aqueous layer was decanted into fresh 1.5 mL microcentrifuge tubes and NaCl (final concentration of 200 mM) and glycogen (30 ug) were added per sample. Precipitation of the samples was carried out with 800 uL of 100% ethanol for 30 min at -80°C. Thereafter, samples were spun at 13200 rpm for 10 min at 4°C, and the pellet was washed with 80% ethanol. The pellets were air-dried and resuspended in 30 uL of 10mM Tris-HCl, pH 8.0 each. The purified DNA was quantitated using Nanodrop spectrophotometer.

2.7.6 ChIP-Seq DNA Library Preparation

The quantitated DNA from IP was used for library construction suitable for Illumina's Solexa Sequencer cluster amplification and sequencing platform. Library was prepared according to the in-house protocol using the NEBNext® ChIP-Seq

Sample Prep Reagent kit (#E6200S). Briefly, 15ng of ChIP DNA fragments were end-repaired and adenosine 'A' base overhang was added at the 3' end of the fragments. Illumina's universal adaptors were added to the ends of the fragments, and amplified by PCR using adaptor-specific primers. The PCR-products were then size selected on a 2% agarose gel and gel purified using QIAGEN gel purification kit (Cat #28704). Samples were eluted with 20 uL of EB buffer. The ChIP DNA libraries were checked for their quantity and quality using Agilent Bioanalyzer DNA 1000 chip kit (Cat #5067-1504). From the sequencing results peak calling, binding site distribution and motif analyses were performed by Sun Wenjie and Hu Xiaoming from our collaborator's (Dr Shyam Prabhakar) lab.

2.8 Embryo Processing for Histology

Freshly dissected mouse embryos from the uterine horn were rinsed with cold Lebovitz medium and fixed overnight at 4°C in 4% paraformaldehyde (PFA) in PBS (DEPC-treated). The following day, embryos were washed with 1xPBS (DEPC-treated) for 10 min at RT and dehydrated through an ethanol gradient of 50% ethanol/ PBS and 70% ethanol/ PBS for 15 min each at RT. Embryos were then processed in an automated tissue processor (Leica TP 1020), embedded in paraffin and sectioned to 10 uM thickness using the microtome (Leica RM 2165).

Tissue processing program	
Reagents	Time
70% Ethanol	30 min
95% Ethanol	15 min
95% Ethanol	15 min
100% Ethanol	1 hr
100% Ethanol	2 hrs
100% Ethanol	2 hrs
Histoclear™ (National diagnostics; #HS-200)	30 min
Histoclear™	30 min
Histoclear™	30 min
Paraffin (with vacuum)	2 hrs 20 min
Paraffin (with vacuum)	3 hrs

2.9 Section In-Situ Hybridization (SISH)

- **RNA Probe Synthesis**

The DIG-labelled RNA probes for sectioned-in-situ hybridization were prepared using the DIG RNA labelling kit (Roche, Cat.#11 175 025 910) according to manufacturer's protocol. All the cDNA clones used as templates for the probe synthesis were purchased from Open BioSystems. After *in vitro* transcription, the reaction was stopped with EDTA (final concentration of 0.04 M) and purified using QUICK spin columns Sephadex G-50 (Roche cat #11274015001). The purified samples were then precipitated with NaOAc, pH 5.5 (final concentration of 0.3 M) and 100% ethanol (2.5x the total volume of reaction). The samples were precipitated in -20^{o}C for 30 min before pelleting at 13,200 rpm for 10 min at 4^{o}C and washing once with 70% ethanol (RNase-free). RNA probe pellets were air-dried and resuspended in 30 uL RNase-free water. The integrity of the probes was checked by resolving on a 0.8% agarose gel and then samples were quantified using the Nanodrop spectrophotometer.

- **Section Pre-treatment, Pre-hybridization and Hybridization**

The paraffin embedded mouse embryos were sectioned (10 uM thickeness) onto polysine-coated slides and air-dried before use. The slides were de-waxed using the Histoclear^{TM} (20 min) and rehydrated through an ethanol gradient of 5 min each (100%, 90%, 70% and 30% ethanol) before washing twice with 1xPBS (DEPC-treated). The tissues were post-fixed in 4% PFA at RT for 20 min, washed again in 1x PBS, before proteinase K digestion (final concentration of 10 ug/mL in 0.1 M Tris, pH 7.5) for 10 min at RT. The slides were washed in 1x PBS and fixed again for 10 min in 4% PFA at RT. Following post-fixing the slides washed and allowed to pre-hybridize in the pre-hybridization buffer at 67^{o}C for 2-3 hrs before an overnight

hybridization with the appropriate DIG-labelled RNA probes (concentration ranging from 50 – 1000 ng/mL based on transcript abundance) at 67^{o}C.

Pre-hybridization buffer	
Final Concentration	**Reagents**
50%	Formamide (Roche; Cat#11814310001)
5X	SSC (1st Base)
1X	50X Denhardt's (5g Ficoll, 5g BSA and 5g Polyvinylpyrrolidine in 500 mL of RNase-free water)
0.1 %	Tween20
0.1 mg/mL	Yeast tRNA (Ambion; Cat# AM7118)
0.05 mg/mL	Heparin
-	RNase-free water

- **Post-hybridization Washes and Probing with anti-DIG Antibody**

Following the overnight hybridization, the slides were washed in pre-warmed Solution 1 thrice, for 30 min each, at the same temperature as hybridization (67^{o}C). The slides were further washed in TNT buffer thrice for 5 min each at RT, once in TNT:Solution 2 (1:1 mixture) for 5 min at RT and finally in pre-warmed Solution 2, thrice for 30 min each, at 63^{o}C (4^{o}C less than the hybridization temperature). Thereafter, the slides were washed thrice in MABT for 5 min each before blocking with 2% blocking solution for 2 – 3 hrs at RT and incubation overnight with alkaline phosphatase (AP)-conjugated anti-DIG antibody (1 in 2000 dilution in 2% blocking solution) at 4^{o}C.

Solution 1	
Final Concentration	**Reagents**
50%	Formamide
5X	SSC (pH 4.5 or 7)
1%	SDS
-	RNase-free water

TNT	
Final Concentration	**Reagents**
10 mM	Tris HCl pH 7.5
0.5 M	NaCl
0.1%	Tween20
-	RNase-free water

Solution 2	
Final Concentration	**Reagents**
50%	Formamide
2X	SSC (pH 4.5 or 7)
0.2%	SDS
-	RNase-free water

- **Post-Antibody Washes and Colour Development**

The following day, the slides were washed with MABT thrice for 10 min each at RT and washed thrice more for 1 hr each. The slides were then washed with freshly prepared NTMT solution thrice for 10 min each before colour development with NBT/BCIPsubstrate (Roche; Cat#11681451001; 200 uL of stock diluted in 10 mL of 0.1 M Tris-HCl, pH9.5 and 0.1 M NaCl). The slides were incubated in the dark at RT until the desired intensity of colour had developed. The slides were then washed in 1xPBS, twice for 10 min each, before they were mounted with an aqueous-based mounting medium, glycerol gelatin (SIGMA cat # 128K6296). All sections were imaged with Zeiss Axio Imager Z1.

NTMT	
Final	**Reagents**
100mM	Tris HCl pH 9.5
50mM	$MgCl_2$
100mM	NaCl
0.1%	Tween20
-	RNase-free water

2.10 Immunohistochemistry (IHC)

Similar to SISH, the paraffin sections on the polysine-coated slides were de-waxed using the Histoclear™ (20 min) and rehydrated through an ethanol gradient of 5 min each (100%, 95%, 90% and 70% ethanol). Antigen retrieval was performed in 0.01 M sodium citrate buffer (pH 6.0) at 121°C for 15 min. After cooling the slides for 3 hrs, the slides were washed twice with 1x PBS before marking the required sections with an immuno-pen. The slides were next incubated with 0.6% hydrogen peroxide for 20 min at RT (protected from light) and washed twice with 1x PBST (0.2% tween-20) for 5 min each. Blocking of non-specific binding sites were carried out with blocking serum provided in the Vectastain® ABC kit (Vector Laboratories; cat # PK-4002) for 30 min at RT and then with 2% BSA/ 5% sheep serum in PBS for another 30 min at RT. Thereafter, the slides were incubated overnight with the primary antibodies (dilutions: Pax1: SC-25407X (M116X) – 1:200; Pax9: SC-25410X (H-95 X) – 1:200; GFP: SC-9996 (B-2) - 1:50; HA epitope: AP09230PU-N, 1:200) at 4°C.

The following day, the slides were allowed to incubate in the primary antibody at RT for another 1 hr before washing in 1xPBST (0.1% Tween20) thrice for 5 min each. The slides were then incubated with the appropriate secondary antibodies (biotinylated bovine anti-rabbit IgG-B – SC-2363 - 1:400; horse anti-mouse IgG – comes with the Vectastain® ABC kit – recommended dilution) for 2 hrs at RT.

Meantime, the A+B complex was prepared by mixing reagents A and B (1:100 dilution) provided in the Vectastain® ABC kit and incubated at 4°C for 1 hr. The slides were then washed thrice with 1xPBST (0.1% Tween20) before incubation for 1 hr with the A+B complex at RT. Thereafter, the slides were washed extensively thrice with 1xPBST (0.1% Tween20), 5 min each, 1xPBS for 10 min and 1x TBS for another 10min. Colour development was then carried out with DAB substrate (BD Biosciences Pharmigen™ DAB substrate kit, cat # 550880) at RT, until the desired signal/noise ratio was obtained. The slides were next washed in running tap water for 5 min before they were mounted with aqueous-based mounting medium (glycerol-gelatin). All the sections were imaged with Zeiss Axio Imager Z1.

2.11 Alcian Blue staining

The paraffin mouse sections were de-waxed in Histoclear™ (20 min) and re-hydrated through ethanol gradient (100%, 90%, 70% and 50% ethanol/PBS) for 3 min each. The slides were washed for 5 min in 1xPBS before a brief rinse in tap water and incubation in alcian blue staining solution for 30 min at RT before washing for 5 min in running tap water. The slides were mounted with glycerol-gelatin.

Alcian Blue staining solution	
Final Concentration	**Reagents**
1%	Alcian Blue 8GX (SIGMA – A3157-10G)
3%	Acetic acid
-	Water

CHAPTER 3 – RESULTS AND DISCUSSIONS

3. Generation of Transgenic Mouse Lines

3.1. Construct Design Strategy

To perform an *in vivo* study using the mouse as model system, there was a requirement to generate the appropriate transgenic mouse lines via gene targeting technique. Since the principal objective of this study was to perform a cell-type specific gene expression analysis to identify the target genes of *Pax1* and *Pax9*, enrichment of specific populations of cells had to be performed. This evidently called for the co-expression of the TF-of-interest with a reporter protein as that would enable us to isolate very pure populations of cells through Fluorescence Assisted Cell Sorting (FACS). In this study, EGFP was the primary choice as the reporter protein owing to its highly photostable nature and it is also known to be one of the brightest fluorescent proteins [89].

Traditionally, to disrupt a gene or to tag it with a reporter protein, the reporter protein (eg. EGFP) is simply inserted into the exon of the gene or at the end of the last exon (in frame), before the stop codon. This however leads to the production of fusion proteins. The drawback with such a strategy is that fusion proteins may become misfolded, resulting in a lower level of EGFP fluorescence expression or an absolute lack of EGFP fluorescence and a non-functional gene product. This paved the way for the use of bi- or poly-cistronic constructs whereby multiple ORFs can be concatenated to co-express multiple proteins from a single promoter, without the production of fusion proteins. The viral internal ribosome entry site (IRES) sequence and 2A oligopeptide sequences (commonly the F2A peptide) are currently the popular choices in generating polycistronic vectors [90]. Both of the elements function by different mechanisms.

When the IRES element is inserted between two open reading frames (ORFs), the translation of the first ORF is initiated by a cap-dependent mechanism, whereas the second one occurs by a cap-independent mechanism [91]. On the other hand, the 2A peptides are self-cleaving, which function by a "ribosomal skipping" mechanism. In brief, a 24 amino acid long 2A peptide sequence is inserted between two ORFs (consensus motif of 2A peptide: DxExNPG↓P- (↓) represents the position of 'skipping'). The ribosome begins translation by a cap-dependent mechanism but when it reaches the motif sequence, it fails to form the peptide bond between the glycine (G) and proline (P) residues, then continues to translate the second ORF. This "skip" results in the production of two discrete proteins [92, 93]. The schematic of F2A-peptide mechanism is shown in Figure 7.

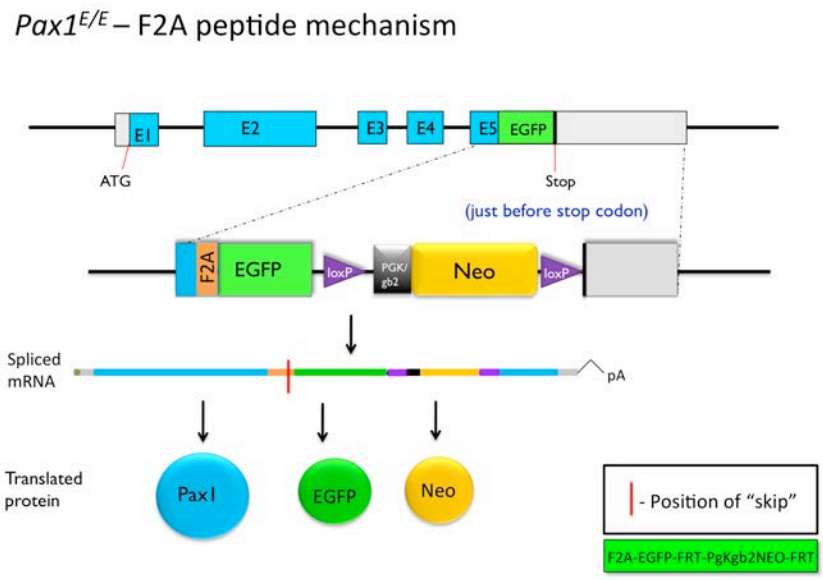


Figure 7: Illustration of F2A-peptide strategy in concatenating ORFs. *Pax1* is used as an example for this illustration. E – exon; Neo – neomycin.

While performing this study, the efficiency of the 2A-peptides and IRES elements in co-expressing concatenated proteins at equal levels were also being investigated, in an *in vivo* system, using other transgenic mouse lines (co-first authored publication; the findings have been published in [94]). Since the IRES

elements were the conventional choice in making the bi- or poly-cistronic vectors, the initial strategy was to use the IRES elements to concatenate the *Pax1* and *EGFP* cistrons (*Pax1*$^{IE/IE}$) to generate the *Pax1* WT allele tagged with *EGFP*. Subsequently, through the extensive analysis of the other mouse lines, it was discovered that both IRES element and the F2A peptide reliably co-expressed the linked proteins. However, the F2A peptide was more efficient at producing stoichiometric (equal) levels of the linked proteins compared to the IRES element. The IRES element gave a consistently lower level of EGFP expression compared to the upstream protein [94]. Since *Pax1* and *Pax9* are expressed at moderate levels, and only in a limited number of cells, the strategy that will express EGFP at sufficiently detectable levels was opted for to ensure that the *Pax1* and *Pax9*-specific cells could be efficiently enriched by FACS. Hence, for all the subsequent constructs to knock-out the *Pax1* or *Pax9* (*Pax1* KO and *Pax9* KO), the F2A peptide strategy was utilized. Notably, the use of F2A strategy will ensure the inhibition of Pax1 or Pax9 protein synthesis without disrupting any of the intronic regions in the genome, which may contain essential cis-regulatory elements. Also, to keep the study consistent, construct with the WT (*Pax1*) allele tagged with *EGFP* using the F2A peptide strategy (*Pax1*$^{E/E}$) was generated.

3.2 Generation of *Pax1* and *Pax9* tagged and knock-out mouse lines

To perform cell-type specific gene expression profiling analyses, several mouse lines were generated with the *Pax1* WT allele tagged with the *EGFP* reporter or knocked-out *Pax1*/ *Pax9* using the *EGFP* reporter.

BAC recombineering technology was employed to generate all the *Pax1* and *Pax9* constructs (listed in Table 4). The details of the construct(s) synthesis are described in the Materials & Methods section. The construct design, Southern

blotting confirmation strategy and the PCR primer pairs used for genotyping the mice are shown in Figures 8-12. Since the *Pax9* heterozygotes did not show any vertebral column defects, an analysis of *Pax9* WT vs *Pax9*$^{-/-}$ was not performed.

For all the constructs, successfully modified BACs were screened carefully for potential mutations / deletions before being subcloned into a minimal vector, linearized and electroporated into mESCs (R1 or V6.4 cells). The modified and subcloned BACs were electroporated into mouse ES cells and were selected in G418 medium for 8-10 days. Correctly targeted ES cell clones were identified by Southern blotting and/or PCR screening. External Southern probes were used to identify the clones that were correctly targeted, while internal probes were used to detect the clones which also had random integration of the cassettes (Figures 8-12). The ES clones were karyotyped and microinjected into 8-cell stage mouse embryos, which were then implanted into pseudo-pregnant mice. Chimeras were obtained and confirmed either by coat colour or PCR screening and mated with C57BL/6J (WT) to obtain the heterozygotes and subsequently the homozygotes by mating the heterozygotes. All the genotyping of the mice was first carried out by Southern blotting and subsequently confirmed by PCR screening.

Table 4: List of *Pax1* and *Pax9* constructs made by BAC recombineering technology.

	Constructs	Genotype	ES cells	Chimeras / Germline transmitting mice?
1	*Pax1*$^{E/E}$	*Pax1* WT allele tagged with EGFP (F2A)	R1 cells	Yes
2	*Pax1*$^{IE/IE}$	*Pax1* WT allele tagged with EGFP (IRES)	R1 cells	Yes
3	*Pax1*^{KO}	*Pax1* knock-out	R1 cells	Yes
4	*Pax1*^{HA3}	*Pax1* WT allele tagged with triple HA epitope	V6.4 cells	Yes
5	*Pax9*^{KO}	*Pax9* knock-out	V6.4 cells	Yes
6	*Pax9*^{HA3}	*Pax9* WT allele tagged with triple HA epitope	V6.4 cells	Yes

3.2.1 *Pax1*$^{IE/IE}$ and *Pax1*$^{E/E}$ - WT mice tagged with EGFP

To tag the WT allele with *EGFP*, either the *F2A-EGFP-FRT-PGKgb2-Neo-FRT* cassette or the *IRES-EGFP-FRT-PGKgb2-Neo-FRT* cassette was inserted by homologous recombination in the respective BAC clones. In both the cassettes, the *Neomycin (Neo)* resistance (selection marker) gene was expressed under the *PGK-gb2* promoters (a dual eukaryotic-prokaryotic promoter) to enable the selection of correctly targeted clones in both the bacterial system (for BAC modification) as well as the mammalian system (ES cell colonies). *FRT* sequences flanked the *PGKgb2-Neo* sequence to enable the removal of the *Neo* through FLPe-mediated recombination.

For the F2A-construct (*Pax1*$^{E/E}$ – Figure 8), the cassette was inserted immediately before the stop codon (in frame) in the Exon 5 of *Pax1*. As for the IRES constructs (*Pax1*$^{IE/IE}$ - Figure 9*),* the cassette was inserted at the 3'UTR (untranslated region) of the *Pax1*.

In the F2A-based cassettes (*F2A-EGFP-FRT-PGKgb2-Neo-FRT*), a *RAKR-GSG* sequence was appended before the F2A peptide sequence (i.e. *RAKR-GSG-F2A-EGFP-FRT-PGKgb2-Neo-FRT*). Since the "skipping" mechanism in the F2A peptide causes a 23 amino acid sequence to be fused to the C-terminal end of the upstream-concatenated protein, a furin protease recognition sequence (*RAKR*) was included immediately before the F2A peptide sequence. This *RAKR* sequence was used to trim the residual 2A peptide from the upstream protein, thus leaving only two additional amino acids (arginine and alanine) at the C-terminus of the Pax1 protein [95]. Moreover, a Gly-Ser-Gly (GSG) spacer was added after this *RAKR* sequence (just before the F2A sequence) in order to enhance the translational "skipping" [96].

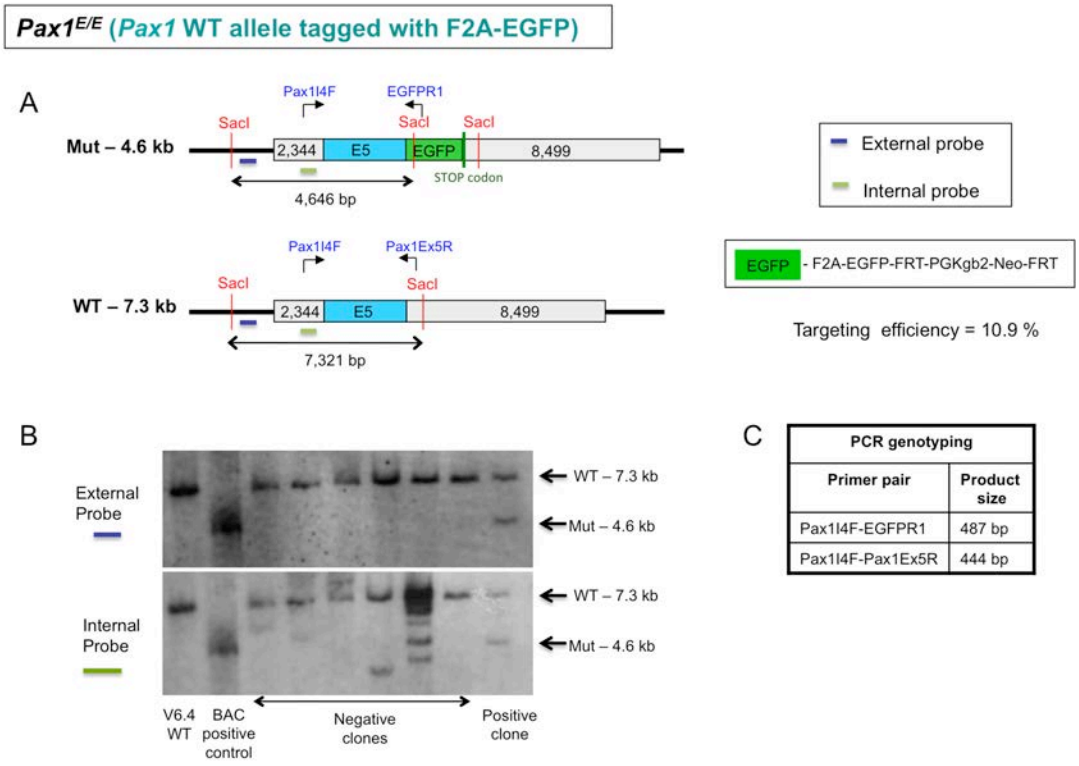


Figure 8: Construct design and confirmation strategy for *Pax1*$^{E/E}$. (A) The construct design whereby the *F2A-EGFP-FRT-PGKgb2-Neo-FRT* cassette was inserted at the 3' end immediately before the translational stop codon of *Pax1*. Short and long homology arms of 2,344 bp and 8,499 bp were used for subcloning. The Southern probes (external – outside the homology arm and internal – within the homology arm) were designed in a repeat-free region. The expected band sizes (with a *SacI* restriction enzyme digestion) were 4,646 bp for a mutant band (correctly targeted) and 7,321 bp for a WT band. Restriction enzyme digestion sites are shown in red; location of the PCR primer pairs used for genotyping are indicated with black arrows; homology arms are represented as grey boxes. The legend on the right indicates the color coding for the external and internal Southern probes and the detail of the EGFP cassette. (B) Southern blotting result of genomic DNA extracted from ES cell colonies and digested with *SacI* restriction enzyme. Top - the blot probed with the external Southern probe; bottom - the blot probed with the internal Southern probe. (C) The primer pairs used and the corresponding PCR product sizes for genotyping of the genomic DNA extracted from the mouse tail tips or the yolk sac. Targeting efficiency of the construct in R1 cells was 10.9%. WT – wild-type; Mut – mutant; E5 – exon 5.

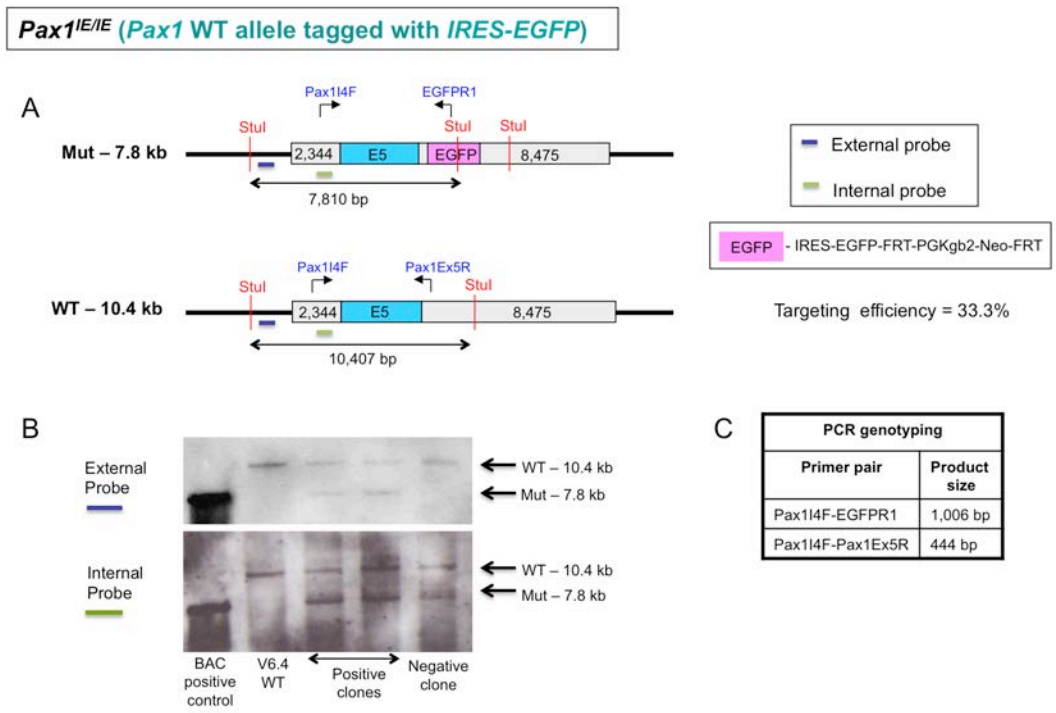


Figure 9: Construct design and confirmation strategy for *Pax1*$^{IE/IE}$. (A) The construct design whereby the *IRES-EGFP-FRT-PGKgb2-Neo-FRT* cassette was inserted at the 3' UTR of *Pax1*. Short and long homology arms of 2,344 bp and 8,475 bp were used for subcloning. The Southern probes (external – outside the homology arm and internal – within the homology arm) were designed in a repeat-free region. The expected band sizes (with a *StuI* restriction enzyme digestion) were 7,810 bp for a mutant band (correctly targeted) and 10,407 bp for a WT band. Restriction enzyme digestion sites are shown in red; location of the PCR primer pairs used for genotyping are indicated with black arrows; homology arms are represented as grey boxes. The legend on the right indicates the color coding for the external and internal Southern probes and the detail of the EGFP cassette. (B) Southern blotting result of genomic DNA extracted from ES cell colonies and digested with *StuI* restriction enzyme. Top - the blot probed with the external Southern probe; bottom - the blot probed with the internal Southern probe. (C) The primer pairs used and the corresponding PCR product sizes for genotyping of the genomic DNA extracted from the mouse tail tips or the yolk sac. Targeting efficiency of the construct in R1 cells was 33.3%. WT – wild-type; Mut – mutant; E5 – exon 5.

3.2.2 *Pax1*^{KO} and *Pax9*^{KO} mice

To knock-out the *Pax1* and *Pax9* genes, the *F2A-EGFP-FRT-PgKgb2-Neo-FRT* cassette was inserted in Exon 2 of *Pax1* and *Pax9,* 3 amino acids after the start of the second exon, by homologous recombination in the respective BAC clones (Figures 10 and 11). The Exon 2 of both *Pax1* and *Pax9* contain the conserved paired-box (DNA-binding domain) that is crucial for the function of these TFs. Also, targeting the Exon 2 will disrupt all the transcripts of *Pax1* and *Pax9* to give a full knock-out of the respective genes.

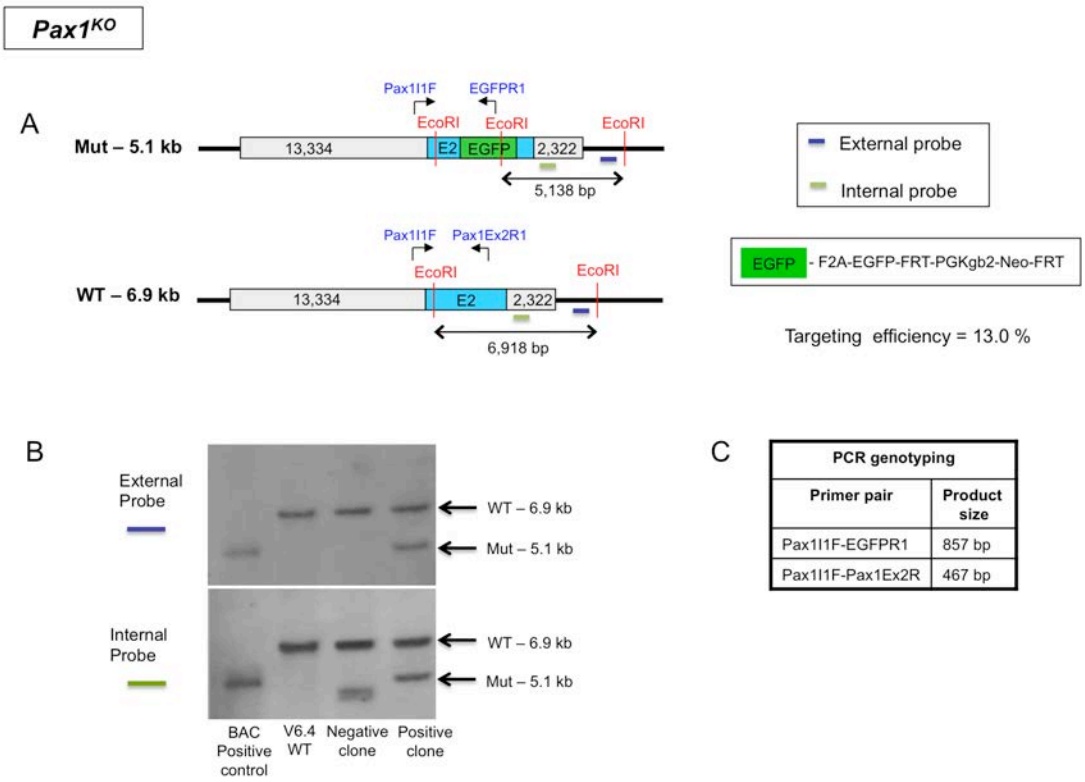
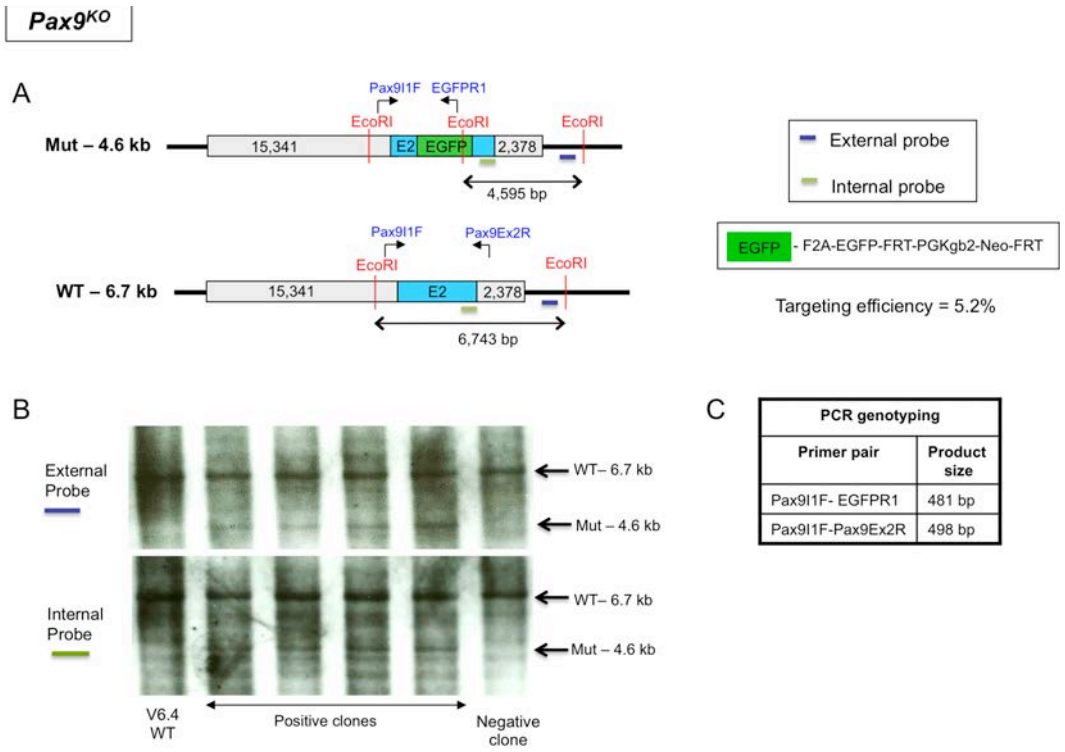


Figure 10: Construct design and confirmation strategy for *Pax1 KO* (*Pax1*$^{-/-}$). (A) The construct design whereby the *F2A-EGFP-FRT-PGKgb2-Neo-FRT* cassette was inserted in the second exon of *Pax1,* 3 amino acids after the start of the second exon. Short and long homology arms of 2,322 bp and 13,334 bp were used for subcloning. The Southern probes (external – outside the homology arm and internal – within the homology arm) were designed in a repeat-free region. The expected band sizes (with *EcoRI* restriction enzyme digestion) were 5,138 bp for a mutant band (correctly targeted) and 6,918 bp for a WT band. Restriction enzyme digestion sites are shown in red; location of the PCR primer pairs used for genotyping are indicated with black arrows; homology arms are represented as grey boxes. The legend on the right indicates the color coding for the external and internal Southern

probes and the detail of the EGFP cassette. (B) Southern blotting result of genomic DNA extracted from ES cell colonies and digested with *EcoRI* restriction enzyme. Top - the blot probed with the external Southern probe; bottom - the blot probed with the internal Southern probe. (C) The primer pairs used and the corresponding PCR product sizes for genotyping of the genomic DNA extracted from the mouse tail tips or the yolk sac. Targeting efficiency of the construct in R1 cells was 13.0%. WT – wild-type; Mut – mutant; E2 – exon 2.



PCR genotyping	
Primer pair	**Product size**
Pax9I1F- EGFPR1	481 bp
Pax9I1F-Pax9Ex2R	498 bp

Figure 11: Construct design and confirmation strategy for *Pax9 KO* (*Pax9*$^{-/-}$). (A) The construct design whereby the *F2A-EGFP-FRT-PGKgb2-Neo-FRT* cassette was inserted in the second exon of *Pax9,* 3 amino acids after the start of the second exon. Short and long homology arms of 2,378 bp and 15,341 bp were used for subcloning. The Southern probes (external – outside the homology arm and internal – within the homology arm) were designed in a repeat-free region. The expected band sizes (with *EcoRI* restriction enzyme digestion) were 4,595 bp for a mutant band (correctly targeted) and 6,743 bp for a WT band. Restriction enzyme digestion sites are shown in red; location of the PCR primer pairs used for genotyping are indicated with black arrows; homology arms are represented as grey boxes. The legend on the right indicates the color coding for the external and internal Southern probes and the detail of the EGFP cassette. (B) Southern blotting result of genomic DNA extracted from ES cell colonies and digested with *EcoRI* restriction enzyme. Top - the blot probed with the external Southern probe; bottom - the blot probed with the internal Southern probe. (C) The primer pairs used and the corresponding PCR product sizes for genotyping of the genomic DNA extracted from the mouse tail tips or the yolk sac. Targeting efficiency of the construct in V6.4 cells was 5.2%. WT – wild-type; Mut – mutant; E2 – exon 2.

3.2.3 *Pax1*^{HA3} and *Pax9*^{HA3} - WT mice tagged with triple HA epitope

For *in vivo* ChIP-seq studies, endogenous tagging of the WT *Pax1* and *Pax9* alleles was required. The HA epitope (YPYDVPDYA) is from the hemagglutinin protein which is a surface glycoprotein of the influenza virus [97, 98]. It was chosen for to its small size (9 amino acids) and the availability of good commercial HA antibodies for ChIP. Triple tandem repeats of the HA epitope were used so as to increase the sensitivity of detection.

The *3xHA-loxP-Neo-loxP* cassette was inserted just before the stop codon in Exon 5 of *Pax1* and Exon 4 of *Pax9* by homologous recombination in the respective BAC clones (Figures 13 and 14). The C-terminal end was chosen for this epitope tagging so as to minimize potential interference of the epitope with the protein folding. This way, the multiple protein-coding transcripts of the *Pax1*/*Pax9* can also be successfully tagged. For *Pax9,* both the protein-coding transcripts (*Pax9-001* and *Pax9-201*) were tagged this way, whereas for *Pax1*, only one of the two transcripts (*Pax1-001*) could be tagged. *Pax1-002* transcript could not be tagged owing to its inherent nature whereby it differs from the other transcript at both the N-terminal and C-terminal ends (Figure 12) (Ref: Ensembl genome browser). Moreover, the C-terminal was preferred over the N-terminal end since a successful detection of the tagged protein would indicate the presence of a fully translated protein.

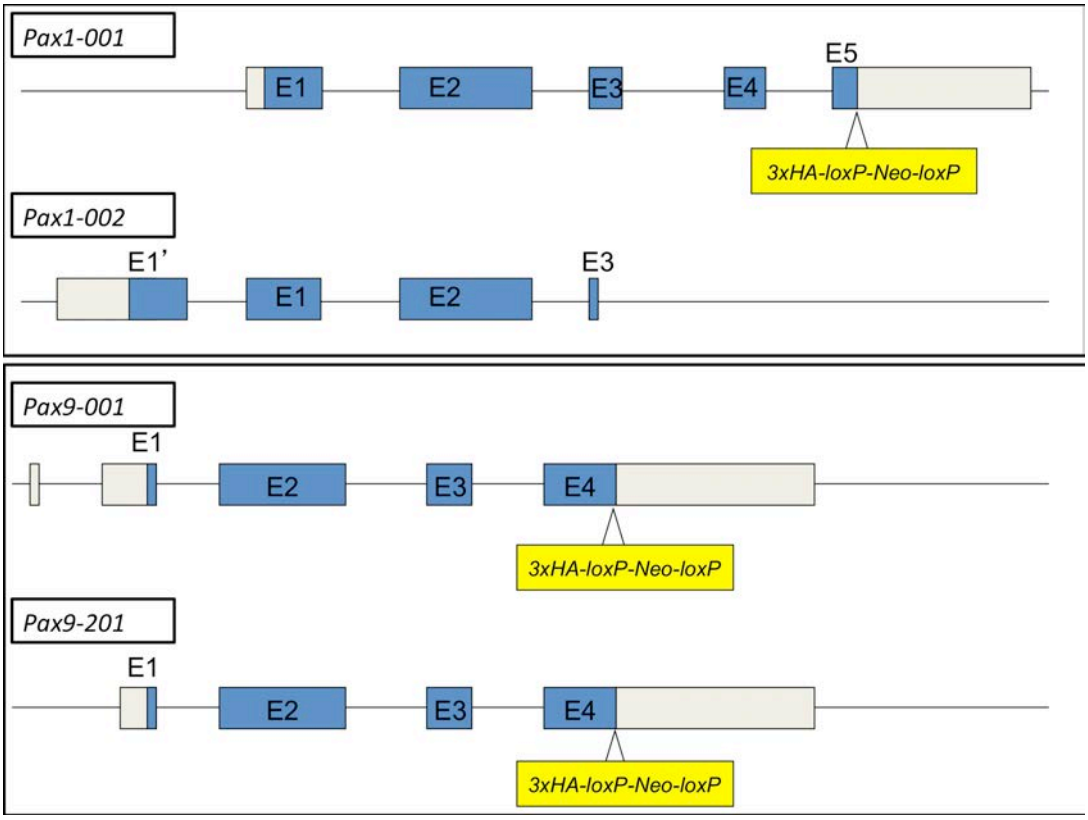


Figure 12: *Pax1* and *Pax9* transcripts targeted for triple HA epitope tagging. *Pax1-001* and *Pax1-002* transcripts differ at the N-terminal and C-terminal ends. *Pax9-001* and *Pax9-201* transcripts are identical in terms of their translational start and stop sites. Only the protein-coding transcripts for each gene are shown. Ref: Ensembl genome browser.

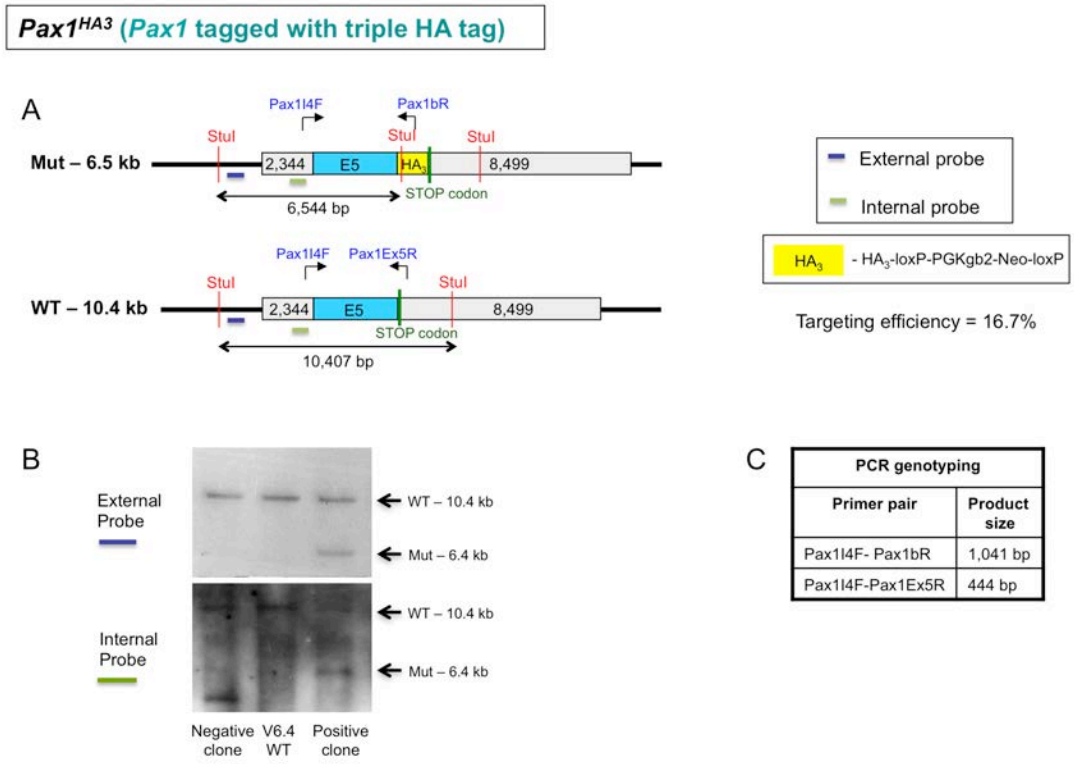


Figure 13: Construct design and confirmation strategy for *Pax1^{HA3}*. (A) The construct design whereby the *HA_3-loxP-PGKgb2-Neo-loxP* cassette was inserted at the 3' end of *Pax1,* immediately before the translational stop codon. Short and long homology arms of 2,344 bp and 8,499 bp were used for subcloning. The Southern probes (external – outside the homology arm and internal – within the homology arm) were designed in a repeat-free region. The expected band sizes (with *StuI* restriction enzyme digestion) were 6,544 bp for a mutant band (correctly targeted) and 10,407 bp for a WT band. Restriction enzyme digestion sites are shown in red; location of the PCR primer pairs used for genotyping are indicated with black arrows; homology arms are represented as grey boxes. The legend on the right indicates the color coding for the external and internal Southern probes and the detail of the HA_3 cassette. (B) Southern blotting result of genomic DNA extracted from ES cell colonies and digested with *StuI* restriction enzyme. Top - the blot probed with the external Southern probe; bottom - the blot probed with the internal Southern probe. (C) The primer pairs used and the corresponding PCR product sizes for genotyping of the genomic DNA extracted from the mouse tail tips or the yolk sac. Targeting efficiency of the construct in V6.4 cells was 16.7%. WT – wild-type; Mut – mutant; E5 – exon 5.

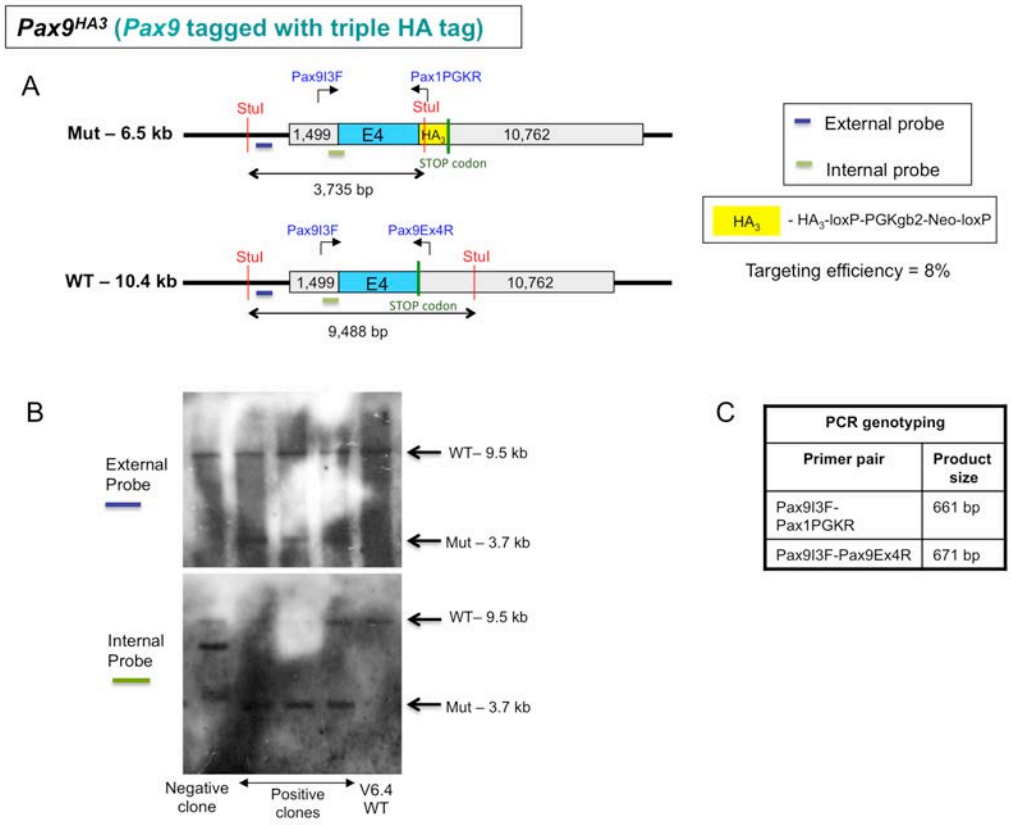


Figure 14: Construct design and confirmation strategy for *Pax9^{HA3}*. (A) The construct design whereby the *HA_3-loxP-PGKgb2-Neo-loxP* cassette was inserted at the 3' end of *Pax9,* immediately before the translational stop codon. Short and long homology arms of 1,499 bp and 10,762 bp were used for subcloning. The Southern probes (external – outside the homology arm and internal – within the homology arm) were designed in a repeat-free region. The expected band sizes (with a *StuI* restriction enzyme digestion) were 3,735 bp for a mutant band (correctly targeted) and 9,488 bp for a WT band. Restriction enzyme digestion sites are shown in red; location of the PCR primer pairs used for genotyping is indicated with black arrows; homology arms are represented as grey boxes. The legend on the right indicates the color coding for the external and internal Southern probes and the detail of the HA_3 cassette. (B) Southern blotting result of genomic DNA extracted from ES cell colonies and digested with *StuI* restriction enzyme. Top - the blot probed with the external Southern probe; bottom - the blot probed with the internal Southern probe. (C) The primer pairs used and the corresponding PCR product sizes for PCR genotyping of the genomic DNA extracted from the mouse tail tips or the yolk sac. Targeting efficiency of the construct in V6.4 cells was 8%. WT – wild-type; Mut – mutant; E4 – exon 4.

3.3 Assessment of *Pax1* and *Pax9* mouse lines

***Pax1* mice**

3.3.1 Phenotypes of the *Pax1*$^{E/E}$ and *Pax1*$^{IE/IE}$ adult mice

The phenotypes of the adult mice for *Pax1*$^{E/E}$ and *Pax1*$^{IE/IE}$ were assessed. The adult mice were of normal size and resembled the WT mice, with no observable abnormalities (Figure 15). The *Pax1*$^{E/E}$ and *Pax1*$^{IE/IE}$ mice were also viable and fertile. To flox out the *Neo*, the F1 mice were crossed to homozygous *Rosa26R*^{Flpe} mice.

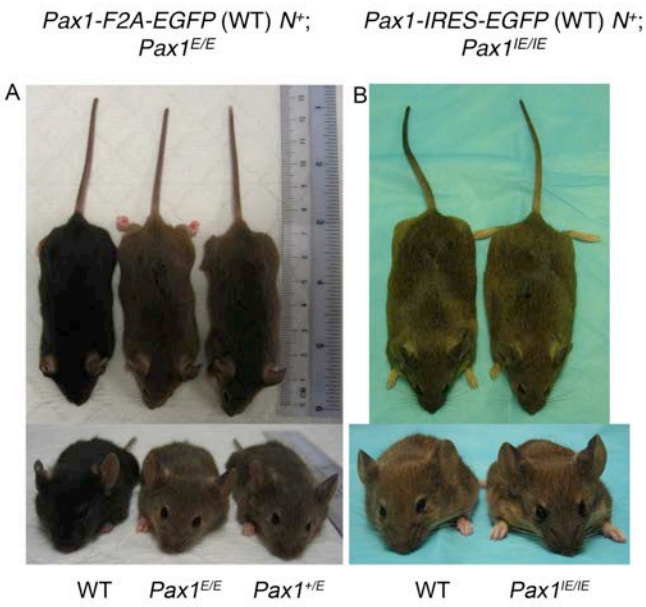


Figure 15: ***Pax1* WT mouse lines tagged with EGFP.** (A) Picture of the adult *Pax1*$^{E/E}$ heterozygous and homozygous mice compared to the WT B6 mouse; mice did not exhibit any abnormalities and were viable and fertile. (B) Picture of the *Pax1*$^{IE/IE}$ homozygous mouse compared to the WT mouse; mice were phenotypically normal, viable and fertile as expected. *N* – Neomycin; WT – wild-type.

3.3.2 Fluorescence expression pattern in the *Pax1*$^{E/E}$ and *Pax1*$^{IE/IE}$ embryos

Next, the EGFP fluorescence expression pattern was assessed at various developmental stages for the heterozygote and homozygote *Pax1*$^{E/E}$ and *Pax1*$^{IE/IE}$ embryos, with or without *Neo* (Figure 17 and Figure 18). This was to ensure that the EGFP fluorescence expression pattern recapitulated endogenous *Pax1* expression (Figure 16) and that the presence of *Neo* did not adversely affect it. As expected, all the heterozygote and homozygote embryos of *Pax1*$^{E/E}$ and *Pax1*$^{IE/IE}$ mouse lines exhibited fluorescence expression in the *Pax1*-specific domains. In general, expression was seen in pharyngeal arches, anterior proximal limb buds, sclerotome,

intervertebral disc of the vertebral column, craniofacial region, eyelid mesenchyme and thymus glands. Presence of *Neo* also did not affect the fluorescence expression patterns in the *Pax1*$^{E/E}$ embryos. All the embryos showed no defects in the vertebral column or the length of the tail.

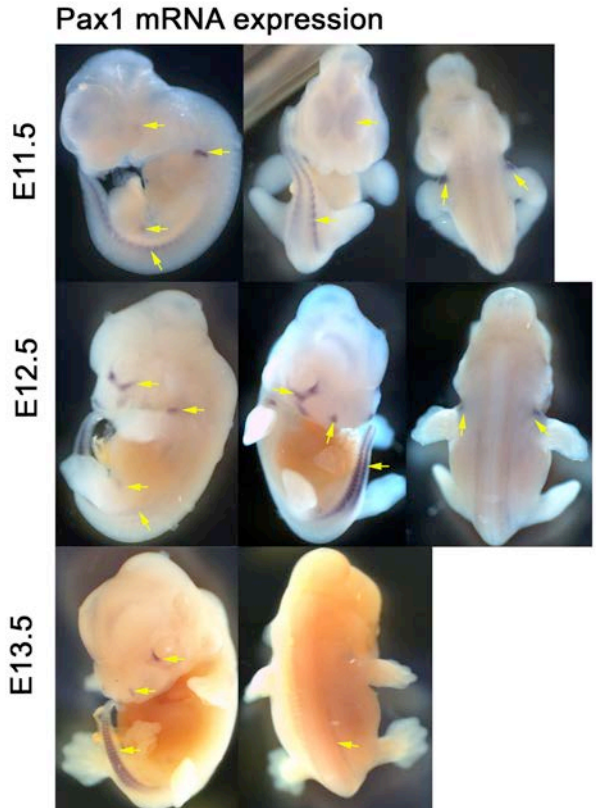


Figure 16: ***Pax1*** **mRNA expression pattern during different developmental stages.** Whole-mount in situ hybridization of E11.5 to E13.5 embryos using *Pax1* anti-sense DIG-labelled RNA probe. Top panel – embryos at E11.5 showed expression in the fore- and hind-limb buds, pharyngeal arch, the forebrain and the sclerotome. Fontal and dorsal views of the embryos are shown. Middle panel – embryos at E12.5 still exhibited expression in the fore- and hind-limb buds and the vertebral column. Expression in the craniofacial regions was also seen at this stage. Lower panel – embryos at E13.5 showed *Pax1* expression in the facial region and the vertebral column. Expression in the fore- and hind-limbs appeared weak.

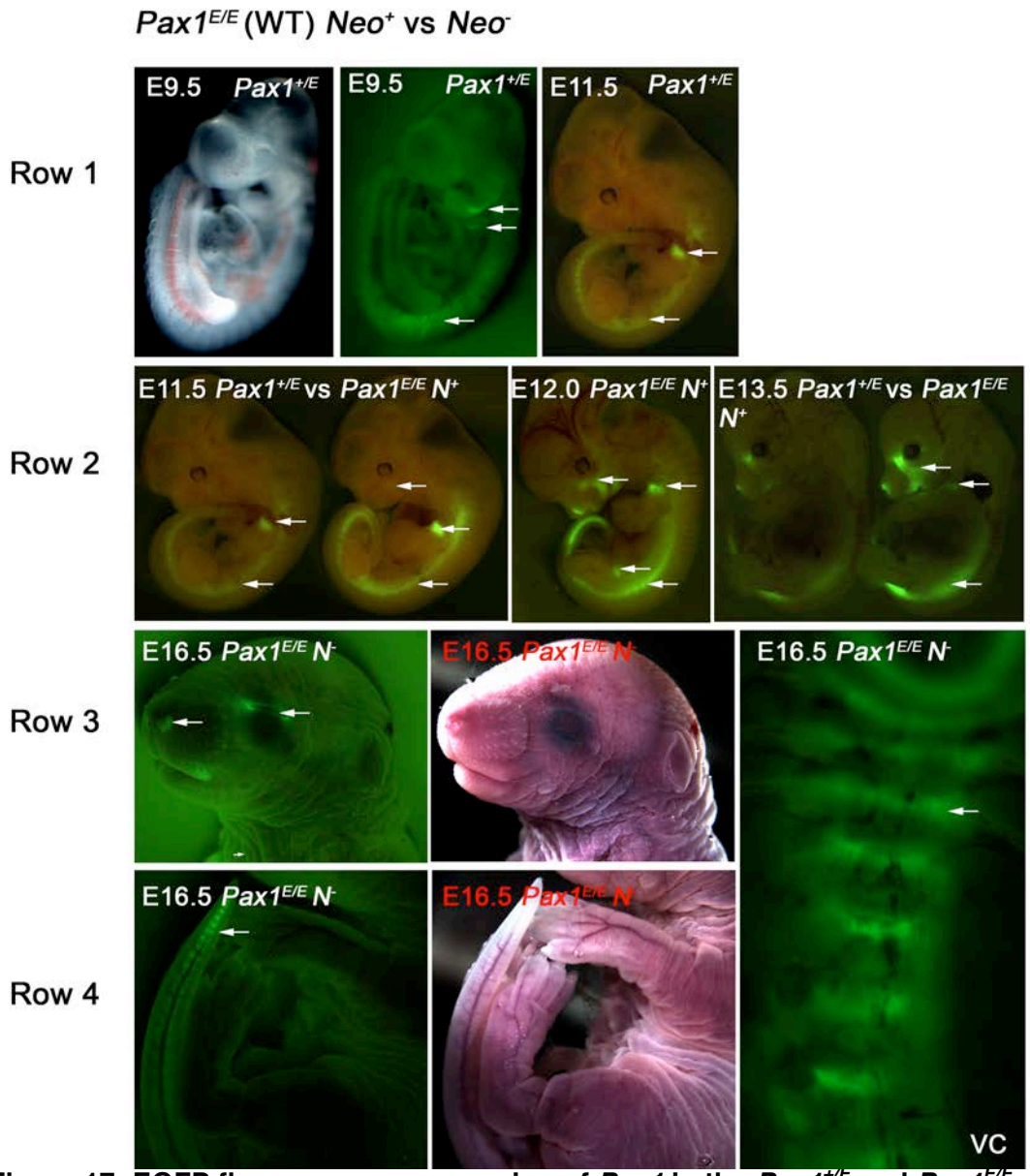


Figure 17: EGFP fluorescence expression of *Pax1* in the *Pax1*$^{+/E}$ and *Pax1*$^{E/E}$ embryos of different developmental stages, with or without *Neomycin*. Row 1: *Pax1*$^{+/E}$, *Neo*$^{+}$ embryos: E9.5 embryo (white light image on the left; fluorescence image on the right) showed expression in the pharyngeal arches and the sclerotomal cells of the somites; E11.5 embryo showed expression in the sclerotome and the fore- and hind-limb buds. Expression in the hind limb bud appears obscured owing to the angle at which the embryo was positioned. Row 2: in the E11.5 embryos, expression in the facial region was seen only in the *Pax1*$^{E/E}$, *Neo*$^{+}$ embryos, albeit weakly. E12.0 embryo showed expression in the facial mesenchyme, limb buds and the sclerotome. E13.5 embryos exhibited fluorescence expression likewise in the facial mesenchyme and the vertebral column. Expression in the limbs was seen in the *Pax1*$^{E/E}$, *Neo*$^{+}$ embryo. Rows 3 and 4: E16.5 *Pax1*$^{E/E}$, *Neo*$^{-}$ embryo showed expression in the eye lid mesenchyme, nasal region and the tail. Corresponding white light image is shown on the right. The vertebral column still expressed *Pax1* at E16.5 as shown in the dissected vertebral column. *N* – *Neomycin*; VC – vertebral column.

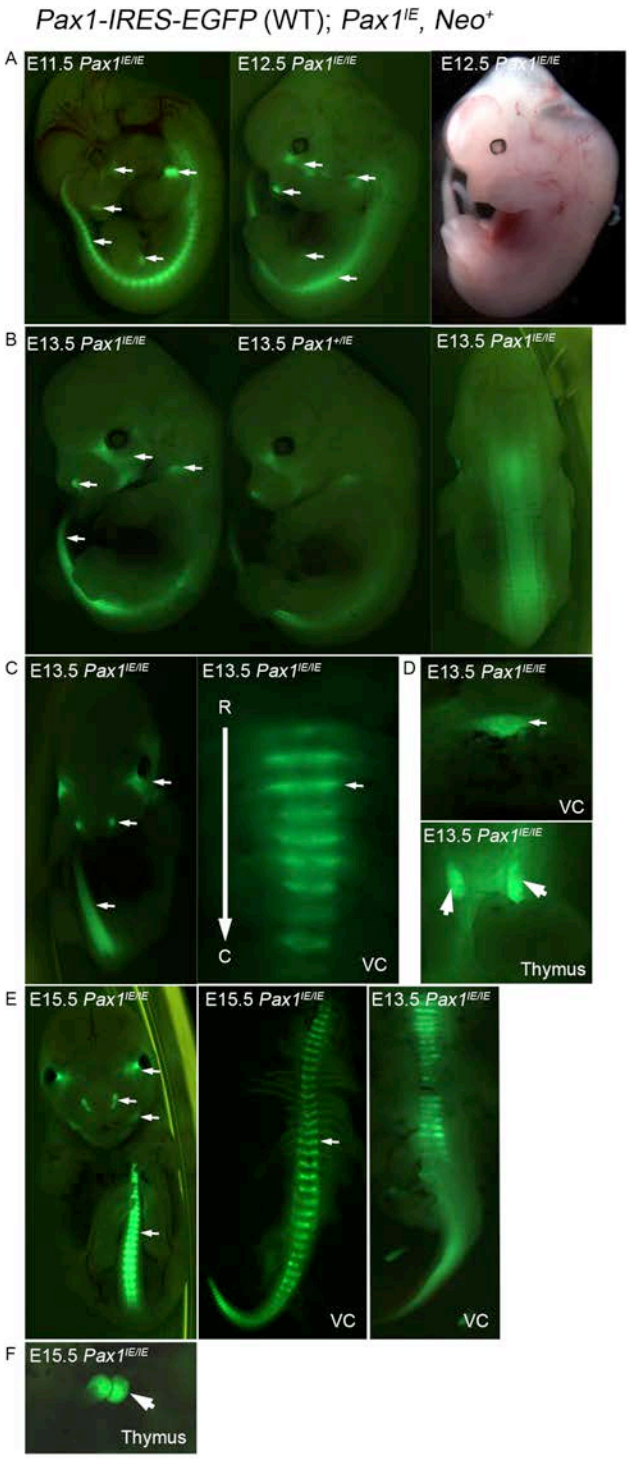


Figure 18: EGFP fluorescence expression of *Pax1* in the *Pax1*$^{+/IE}$ and *Pax1*$^{IE/IE}$, *Neo*$^{+}$ embryos of various developmental stages. (A) E11.5 and E12.5 *Pax1*$^{IE/IE}$ fluorescing embryos with white light image of E12.5 embryo on the right. Expression was observed in the facial mesenchyme, the fore- and hind-limb buds and the vertebral column. (B) E13.5 *Pax1*$^{IE/IE}$ and *Pax1*$^{+/IE}$ fluorescing embryos with dorsal view on the right. (C) Frontal view of E13.5 *Pax1*$^{IE/IE}$ embryo on the left; ventral view of dissected vertebral column of the cervical region on the right. (D) Top panel – transverse section of vertebral column; bottom panel – thymus glands showed strong fluorescence. (E) Frontal view of E15.5 *Pax1*$^{IE/IE}$ embryo on the left, showed a tail of normal length; the entire dissected vertebral column of E15.5 and E13.5 *Pax1*$^{IE/IE}$ on the right. (F) Thymus glands of E15.5 *Pax1*$^{IE/IE}$ embryo showed *Pax1* expression. Thymus glands appeared more rounded instead of the elongated appearance seen at E13.5. R – rostral; C – caudal; VC – vertebral column.

3.3.3 Pax1 and Pax9 protein expression in the *Pax1*$^{E/E}$ embryos

Besides the fluorescence expression pattern, the Pax1 and Pax9 protein expression was also assessed by immunohistochemistry in the E13.5 *Pax1*$^{+/E}$ embryos to further confirm that the mouse line resembled the WT (Figures 19 and 20). Pax1 protein was detected in the IVD anlagen (Figure 19) and Pax9 protein was also detected in the tooth and facial mesenchyme which are Pax9-specific regions (Figure 20).

E13.5 *Pax1*$^{+/E}$ embryo – anti-Pax1 ab

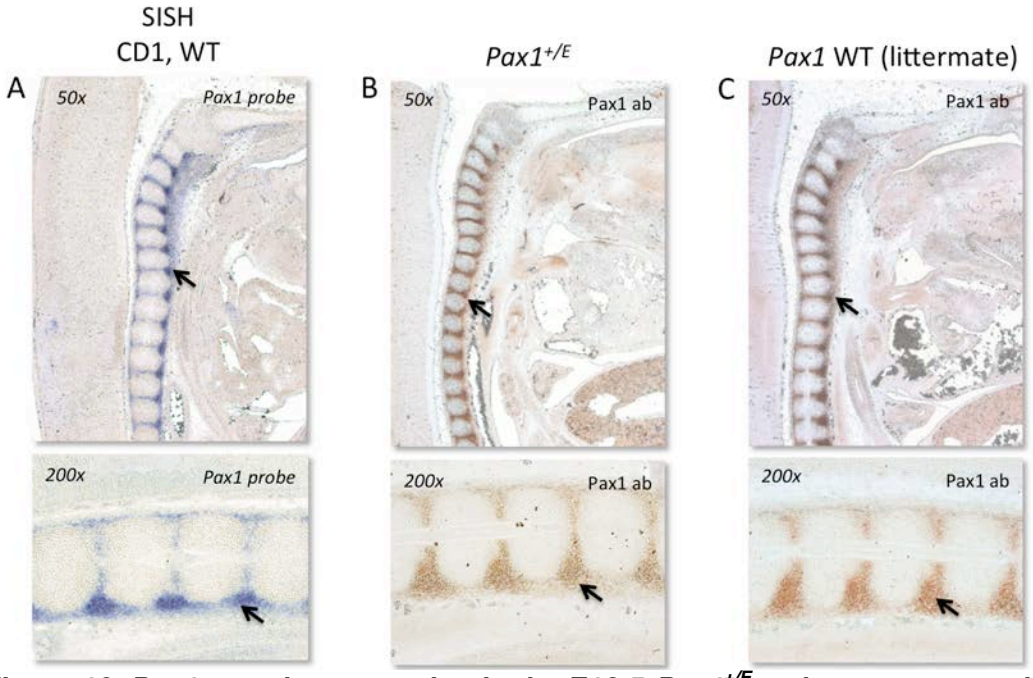


Figure 19: Pax1 protein expression in the E13.5 *Pax1*$^{+/E}$ embryos compared to the WT. (A) *Pax1* mRNA expression detected by sectioned *in situ* hybridization using anti-sense *Pax1* DIG-labelled RNA probe. *Pax1* expression was seen in the intervertebral disc anlagen; magnified image of the vertebral column is shown below the top panel. (B) Pax1 protein expression in the *Pax1*$^{+/E}$ embryo resembled expression pattern seen in the littermate WT (C). All the paraffin sections were of 10um thickness. SISH – sectioned *in situ* hybridization; ab – antibody.

E13.5 *Pax1*$^{+/E}$ embryo – anti-Pax9 ab

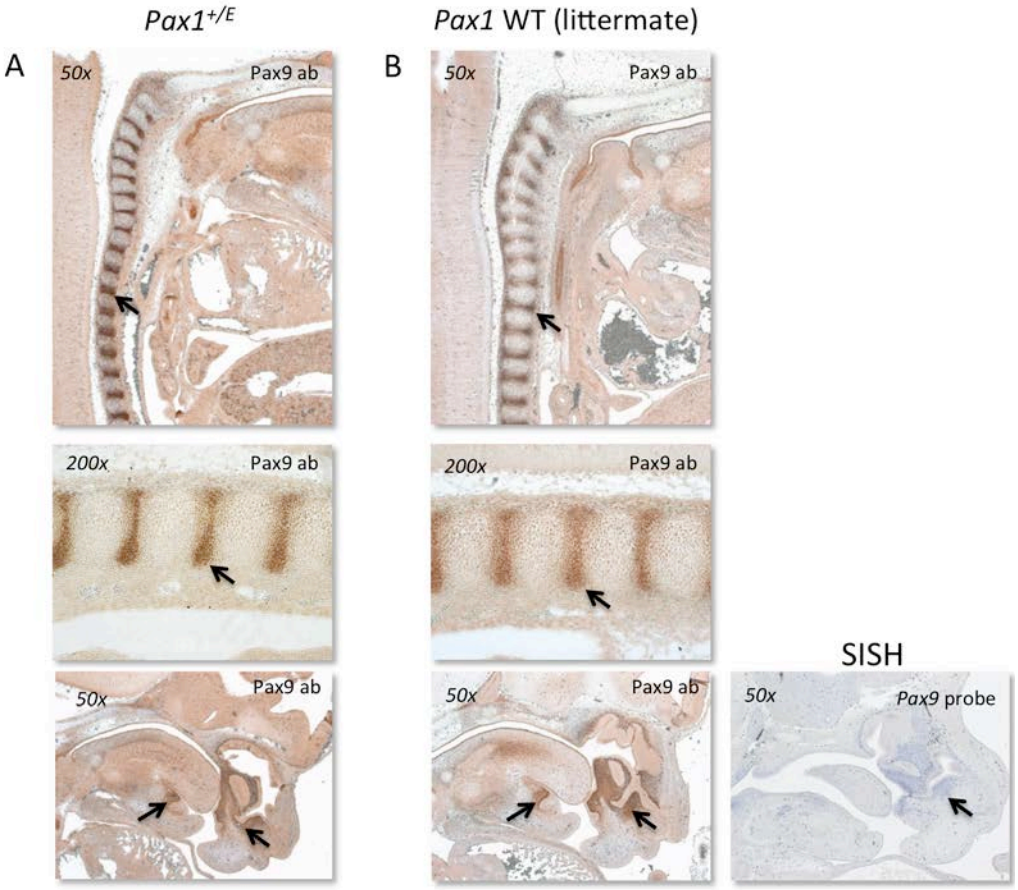


Figure 20: Pax9 protein expression in the E13.5 *Pax1*$^{+/E}$ embryos compared to the WT. (A) Pax9 protein was detected in the intervertebral disc anlagen, tooth and facial mesenchyme in the *Pax1*$^{+/E}$ embryo, which recapitulated endogenous Pax9 expression in the littermate WT shown in B; (B) *Pax9* mRNA expression detected by sectioned *in situ* hybridization using anti-sense *Pax9* DIG-labelled RNA probe shown on the side; expression was seen in the facial mesenchyme. All the paraffin sections were of 10 um thickness. SISH – sectioned *in situ* hybridization; ab – antibody.

Even though no change in the EGFP fluorescence expression was detected in the presence of *Neo*, it was further ensured that there were no adverse effects in the Pax1 and Pax9 protein expression. Therefore, Pax1 and Pax9 protein expression was also evaluated in the E13.5 *Pax1*$^{E/E}$ *Neo*$^{+}$ and *Neo*$^{-}$ embryos. All the Pax1 expression domains, such as the IVD anlagen and tail, and the Pax9-specific tooth and craniofacial mesenchyme, were expressing Pax1 and Pax9 respectively, in both the *Neo*$^{+}$ and *Neo*$^{-}$ E13.5 embryos (Figure 21). Since the presence of *Neo* did not

affect the fluorescence expression pattern, subsequent studies were all performed with the *Neo*$^{+}$ embryos.

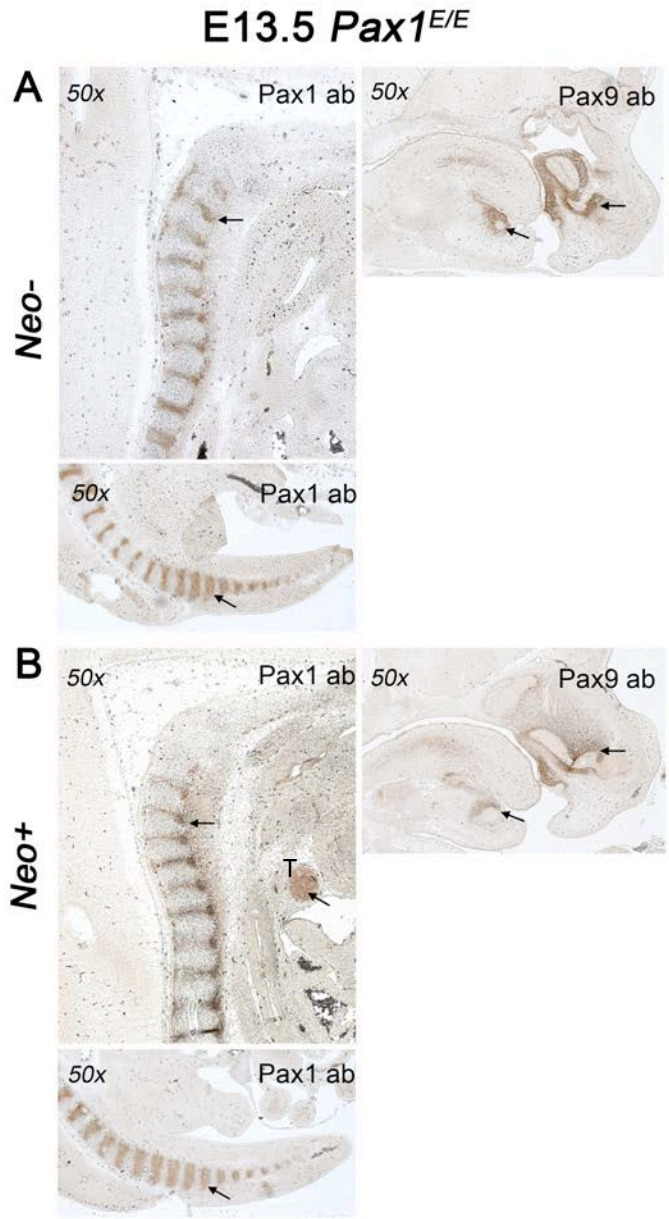


Figure 21: Pax1 and Pax9 protein expression in the E13.5 *Pax1*$^{E/E}$ *Neo*$^{-}$ *and Neo*$^{+}$ embryos. Pax1 expression was observed in the intervertebral disc anlagen and tail. Pax9 expression was seen in the tooth mesenchyme and facial mesenchyme. A: *Neo*$^{-}$ and B: *Neo*$^{+}$. All the paraffin sections were of 10um thickness. ab – antibody; T- thymus anlagen.

3.3.4 Phenotype of the *Pax1*$^{-/-}$ adult mice

The *Pax1*$^{-/-}$ adult mutant mice were viable and fertile, but slightly smaller and had short, kinked-tails compared to the WT mice, which correlated with what has been reported by Wilm *et al* (1998) (Figure 22) [73]. In all our litters, the *Pax1*$^{+/-}$ mice, however, did not show any vertebral abnormalities and appeared normal (not shown). This is in contrary to what was reported by Wilm *et al*. The authors had

reported an 88% penetrance of defects in the lumbar and sternum skeletal elements in the *Pax1*$^{+/-}$. They attributed the lack of 100% penetrance to a potential genetic background effect [73]. Another reason for this discrepancy could be the fact that Wilm *et al* had deleted the first two exons of *Pax1* including the first intron. In this method, the 2nd exon of *Pax1* was disrupted by the insertion of the *EGFP* cassette, without disrupting any intronic regions. There may be some regulatory elements present within the 1st intron which resulted in the mild skeletal phenotypes seen by the authors.

To remove the floxed *Neo* the F1 mice were further mated to homozygous *Rosa26R*^{Flpe} mice.

***Pax1*^{KO} (*Pax1*$^{-/-}$)**

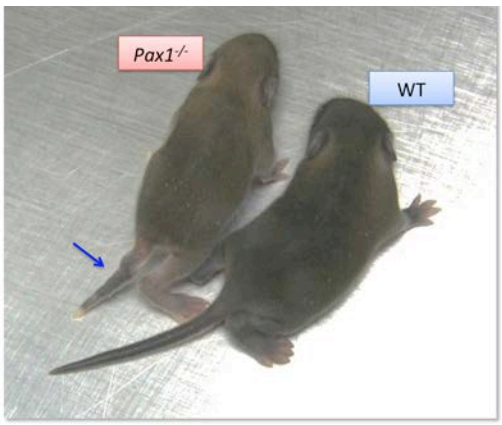


Figure 22: *Pax1*$^{-/-}$ and WT adult mice. *Pax1*$^{-/-}$ mouse was smaller, with a shortened and kinked tail (left) compared to a WT mouse (right).

3.3.5 EGFP expression pattern

The EGFP fluorescence expression pattern was assessed at various developmental stages for the *Pax1*$^{-/-}$ embryos, with or without *Neo*. The fluorescence expression pattern recapitulated endogenous *Pax1* expression (Figure 16) and the presence of *Neo* did not adversely affect it (Figure 23). The *Pax1*$^{-/-}$ embryos showed shortened tail from E13.5 onwards, with the defect becoming progressively more severe at later developmental stages (Figure 23). Since the presence of *Neo* did not

affect the fluorescence expression pattern, subsequent studies were all performed with the *Neo*$^{+}$ embryos.

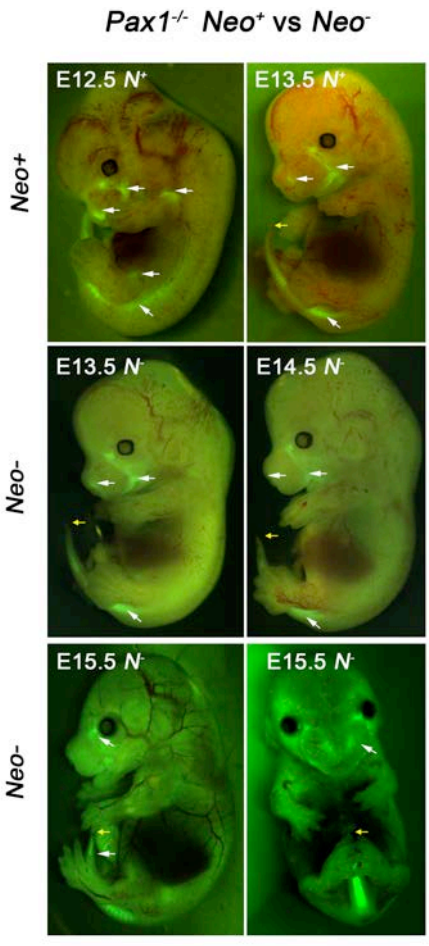


Figure 23: EGFP fluorescence expression in the *Pax1*$^{-/-}$ *Neo*$^{+}$ and *Neo*$^{-}$ embryos of various developmental stages. Top panel: Embryos with the *Neo* showed EGFP expression in the *Pax1*-domains (the facial mesenchyme, anterior proximal limb buds and the vertebral column); the tail appeared shorter in the E13.5 embryos compared to a WT (Figure 16 and 17D). Middle and lower panels: Embryos without the *Neo* also showed similar expression pattern to that of the *Neo*$^{+}$ shown in the top panel. Embryos show progressively shorter tail with increasing developmental stages. Yellow arrows indicate the tail tip. *N* – Neomycin.

3.3.6 Pax1 and Pax9 protein expression in the *Pax1*$^{-/-}$ embryos

Loss of Pax1 protein expression in the *Pax1*$^{-/-}$ embryos was confirmed by immunohistochemistry in E13.5 *Pax1*$^{-/-}$ embryos using anti-Pax1 antibodies. The *Pax1*$^{+/-}$ embryos did not show any difference in the Pax1 protein expression compared to the littermate WT (Figure 24).

Next, Pax9 protein expression was evaluated in the *Pax1*$^{+/-}$ and *Pax1*$^{-/-}$ embryos. In this study, no change was observed in the Pax9 domains of expression in both of the E13.5 *Pax1*$^{+/-}$ and *Pax1*$^{-/-}$ embryos. The expression in both the intervertebral disc anlagen and the tooth and facial mesenchyme resembled that of the WT (Figure 25).

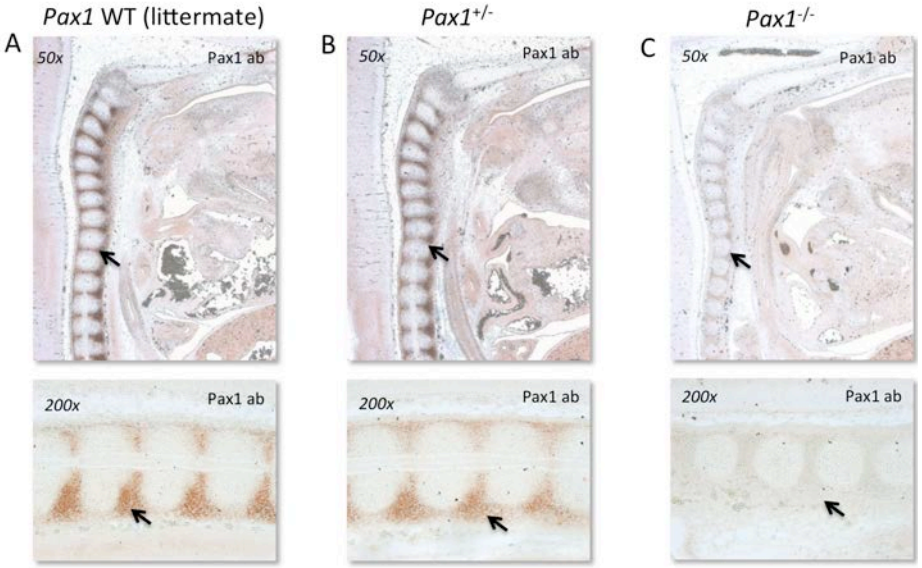


Figure 24: Pax1 protein expression in the E13.5 *Pax1*$^{+/-}$ and *Pax1*$^{-/-}$ embryos compared to the littermate WT. (A and B) Pax1 protein expression was seen in the intervertebral disc anlagen; magnified image of the vertebral column is shown below the top panel. (C) Pax1 protein expression was not detected in the *Pax1*$^{-/-}$ embryo. All the paraffin sections were of 10um thickness. ab – antibody.

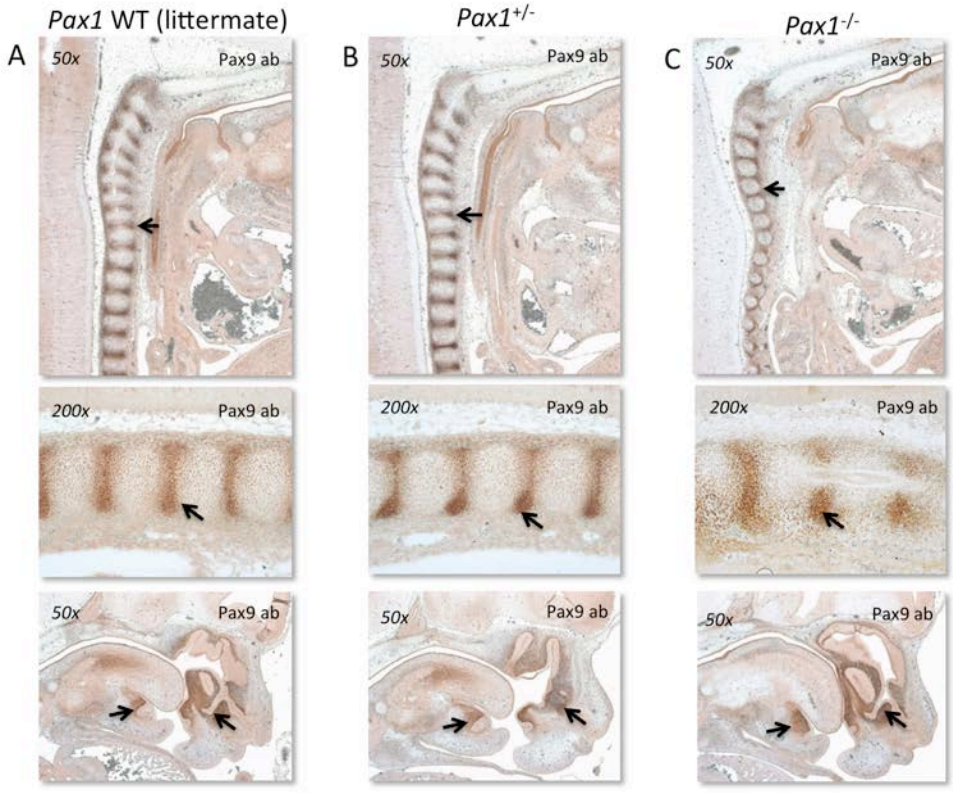


Figure 25: Pax9 protein expression in the E13.5 *Pax1*$^{+/-}$ and *Pax1*$^{-/-}$ embryos compared to the littermate WT. (A) Pax9 protein expression was seen in the intervertebral disc anlagen (magnified image below the top panel) and the tooth and facial mesenchyme. Similar expression pattern was observed in the *Pax1*$^{+/-}$ and *Pax1*$^{-/-}$ embryos (B & C). All the paraffin sections were of 10 um thickness.

3.3.7 *Pax1*$^{-/-}$ vertebral defect

Since the *Pax1*$^{-/-}$ embryos showed a shortening of the tail, the embryos were assessed by histology to investigate what caused such a vertebral defect. The E13.5 *Pax1*$^{+/-}$ and *Pax1*$^{-/-}$ embryos were checked by immunohistochemistry using anti-GFP antibodies. While no vertebral defects were seen in the *Pax1*$^{+/-}$ embryo, the *Pax1*$^{-/-}$ embryo showed a clear loss of vertebral body and IVD cells in the lumbo-sacral region (Figure 26). Alcian blue staining of the transverse sections of these embryos in the lumbo-sacral region also showed a dorso-ventral reduction in the size of the vertebral bodies (Figure 27).

This correlates with the known phenotype in the *Pax1*$^{-/-}$ newborn mice whereby, in the vertebral column, the split vertebrae defect occurred mainly in the lumbo-sacral segments. Moreover, it was reported that a decrease in proliferation and increase in apoptosis in the sclerotome of the E12.5 *Pax1*$^{-/-}$*Pax9*$^{-/-}$ embryos in the tail region [62]. Hence, a significant loss of these sclerotome cells early in development appears to be responsible for the ventro-medial reduction of the vertebral bodies as well.

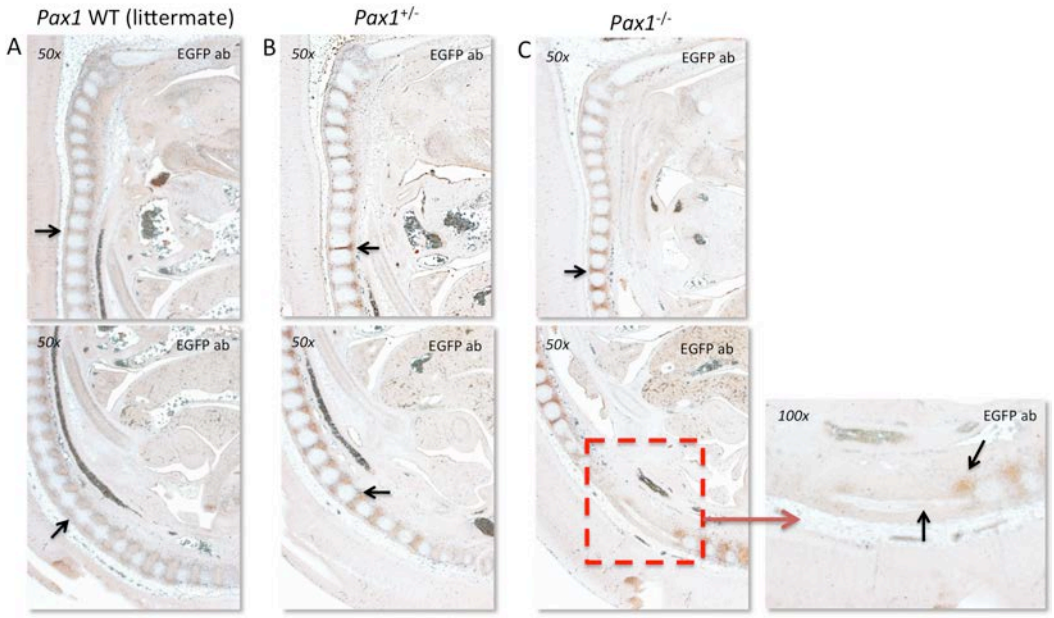


Figure 26: EGFP expression in the E13.5 *Pax1*^{KO} embryos. (A) Littermate WT embryo did not show any EGFP expression as expected. (B) The *Pax1*$^{+/-}$ embryo showed EGFP expression in the intervertebral disc anlagen. No vertebral column defects were observed. (C) The *Pax1*$^{-/-}$ embryo showed EGFP expression in the intervertebral disc anlagen. A loss of vertebral bodies and intervertebral disc cells in lumbo-sacral region was also seen (red dotted box). All the paraffin sections were of 10um thickness. Magnified image of the defective segment is shown on the right. ab – antibody.

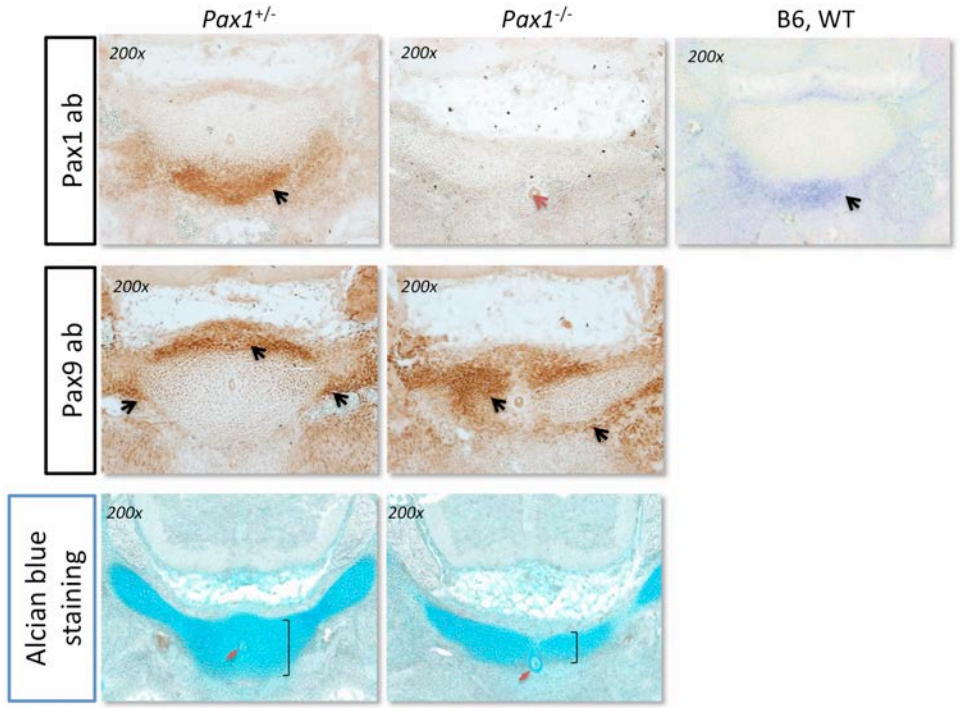


Figure 27: Histochemical analysis of E13.5 *Pax1*$^{+/-}$, *Pax1*$^{-/-}$ embryos. Top panel: Transverse section of E13.5 embryos probed with anti-Pax1 antibody showed a complete loss of Pax1 protein expression in the *Pax1*$^{-/-}$ embryo. Expression of Pax1 protein in the *Pax1*$^{+/-}$ recapitulates *Pax1* mRNA expression shown by sectioned *in situ* hybridization on the WT embryo using *Pax1* anti-sense RNA probe (right). Middle panel: Pax9 protein expression was still detected in the *Pax1*$^{+/-}$ and *Pax1*$^{-/-}$ embryos. Lower panel: Alcian blue staining (stains for the proteoglycans in the cartilage) of the *Pax1*$^{+/-}$ and *Pax1*$^{-/-}$ embryos clearly showed dorso-ventral reduction of the vertebral bodies only in the *Pax1*$^{-/-}$ embryo in the lumbo-sacral region. All the paraffin sections were of 10um thickness. Red arrows point to the notochord. ab – antibody.

***Pax9* Mice**

3.3.8 Fluorescence expression in the *Pax9*$^{-/-}$ embryos

Pax9$^{-/-}$ mice exhibited cleft secondary palate and died shortly after birth, which correlated with the published report by Peters *et al* (1998 and 1999) [62, 74]. The embryos were harvested at different developmental stages and their EGFP expression pattern was assessed. The fluorescence expression pattern recapitulated endogenous *Pax9* expression (Figure 28) – pharyngeal arches, facial mesenchyme,sclerotome and anterior proximal fore- and hind-limb buds [50]. The

expression domains in the facial and limb mesenchyme were all adjacent to the *Pax1* expression domains in these regions.

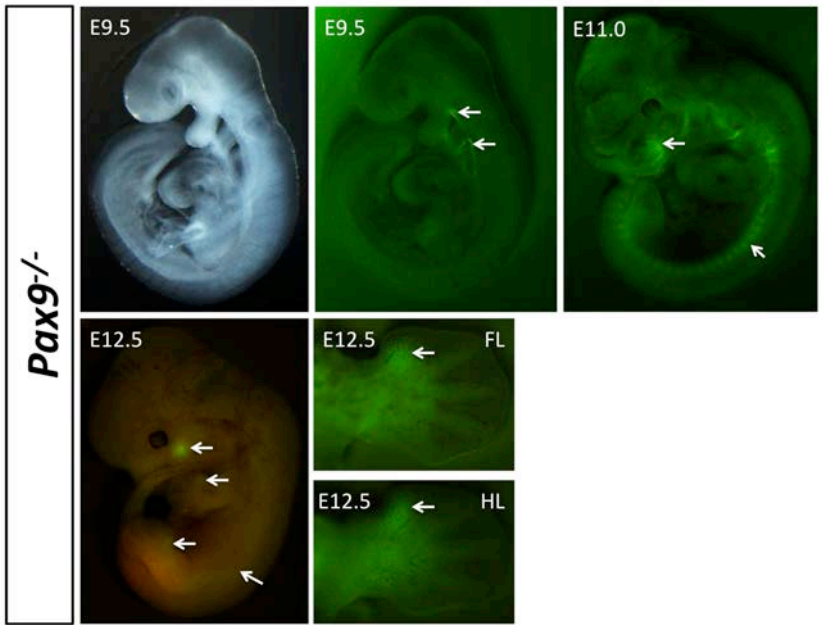


Figure 28: Fluorescence expression pattern of the *Pax9*$^{-/-}$ embryos. Embryos showed expression in the pharyngeal arches at E9.5 (white light image on the left); expression in the somites was too weak to be detected; at E11.0 expression was seen in the facial mesenchyme, pharyngeal arches and the sclerotome; at E12.5 expression was detected in the facial mesenchyme, anterior proximal fore- and hind-limb buds (domains adjacent to what was observed for *Pax1*). Magnified images of the limb buds are shown on the right. FL – fore limb; HL – hind limb.

3.3.9 Pax1 and Pax9 protein expression in the *Pax9*$^{-/-}$ embryos

Loss of Pax9 protein expression in the *Pax9*$^{-/-}$ embryos was confirmed by immunohistochemistry in the E13.5 *Pax9*$^{-/-}$ embryos using anti-Pax9 antibodies. The *Pax9*$^{-/-}$ embryos did not show any difference in the Pax1 protein expression compared to the *Pax9*$^{+/-}$ embryo. No vertebral column abnormalities were observed in the E13.5 *Pax9*$^{-/-}$ embryos, unlike in the *Pax1*$^{-/-}$ where the vertebral column was defective in the lumbo-sacral segments (Figure 29).

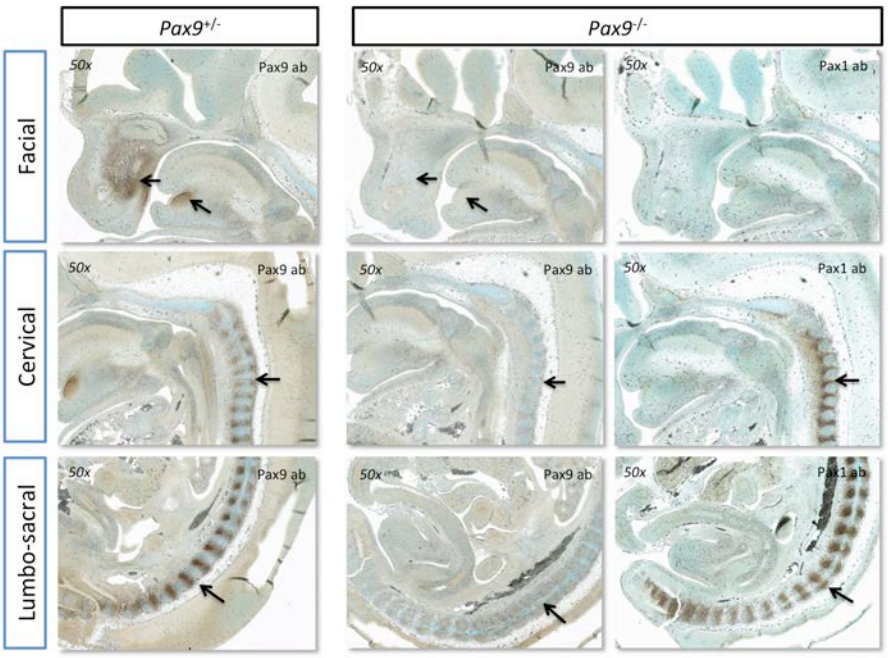


Figure 29: Immunohistochemistry of the E13.5 *Pax9*^{KO} embryos. Left panel: Pax9 protein expression was detected in the facial and tooth mesenchyme and the intervertebral disc anlagen. Middle panel: the *Pax9*$^{-/-}$ embryo was devoid of any Pax9 protein expression, while Pax1 protein was still expressed in the intervertebral disc (right panel). Pax1 protein was not detected in the facial and tooth mesenchyme - domains which are *Pax9*-specific. The vertebral column appeared to have no defects in both of the *Pax9*$^{+/-}$ and *Pax9*$^{-/-}$ embryos, including the lumbo-sacral region (all three panels). All the paraffin sections were of 10um thickness and were counter-stained with alcian blue. ab – antibody.

3.3.10 *Pax1/ Pax9* multiple allele knock-outs

Once the *Pax1*^{KO} and *Pax9*^{KO} mouse lines were confirmed to be accurate, double heterozygotes (DH = *Pax1*$^{+/-}$*Pax9*$^{+/-}$) were generated by mating the *Pax1*$^{-/-}$ mice with the *Pax9*$^{+/-}$ mice. Different combinations of allele knock-out embryos were then generated from a DH x DH mating, since the DH mice were viable and fertile (Figure 30). The *Pax1*$^{-/-}$*Pax9*$^{+/-}$ mice died post-natally (as reported in [62]), hence could not be used for mating to obtain the *Pax1/Pax9* double null embryos for gene expression profiling analysis. The embryos were checked by fluorescence microscopy for their expression patterns (Figure 30), and the yolk sacs from the

corresponding embryos were genotyped by PCR. Embryos of interest were FACS-ed individually.

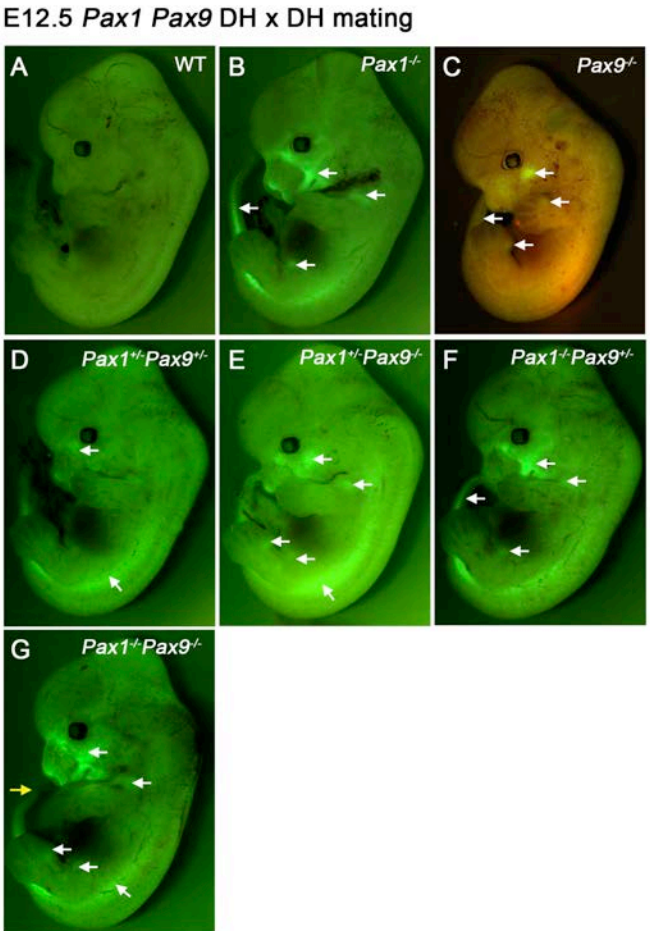


Figure 30: Different combinations of knock-out allele embryos obtained from DH x DH matings. (A-G) Genotype of each embryo is shown, and fluorescence expression sites are indicated with white arrows. The *Pax1*$^{+/-}$ and *Pax9*$^{+/-}$ embryos exhibited very weak fluorescence expression (not shown). (B) In the *Pax1*$^{-/-}$ embryo, fluorescence expression was observed in the facial mesenchyme, fore- and hind-limb buds and the vertebral column. (C) In the *Pax9*$^{-/-}$ embryo fluorescence expression was seen in the facial mesenchyme, but adjacent to the *Pax1* domain. Similarly, expression in the anterior proximal fore- and hind-limb buds was seen, including a much weaker expression in the vertebral column. (G) Expression in the double null (*Pax1*$^{-/-}$*Pax9*$^{-/-}$) embryo was a combination of both the *Pax1* and *Pax9* expression sites. The tail was significantly shortened as indicated by the yellow arrow. DH – double heterozygote (*Pax1*$^{+/-}$*Pax9*$^{+/-}$).

The *Pax1/Pax9* double-null embryos showed fluorescence expression in both the *Pax1* and *Pax9* expression domains (Figure 30). The tail was significantly shorter than the WT embryos which could be attributed to the prevalent loss of sclerotomal cells in these embryos [62]. The shortening of tail was more severe at older stages (Figure 31). Also, at E14.5, the double-null embryos showed mis-localization of the fluorescing cells. Instead of forming IVD structures medially, the cells were localized

on the lateral regions. The notochord, which is normally enclosed within the vertebral column, was exposed owing to the lack of normal vertebral body and IVD cells in the middle to surround it. Polydactyly was also observed in the hind limbs of the double-null embryos by E14.5 (Figure 31).

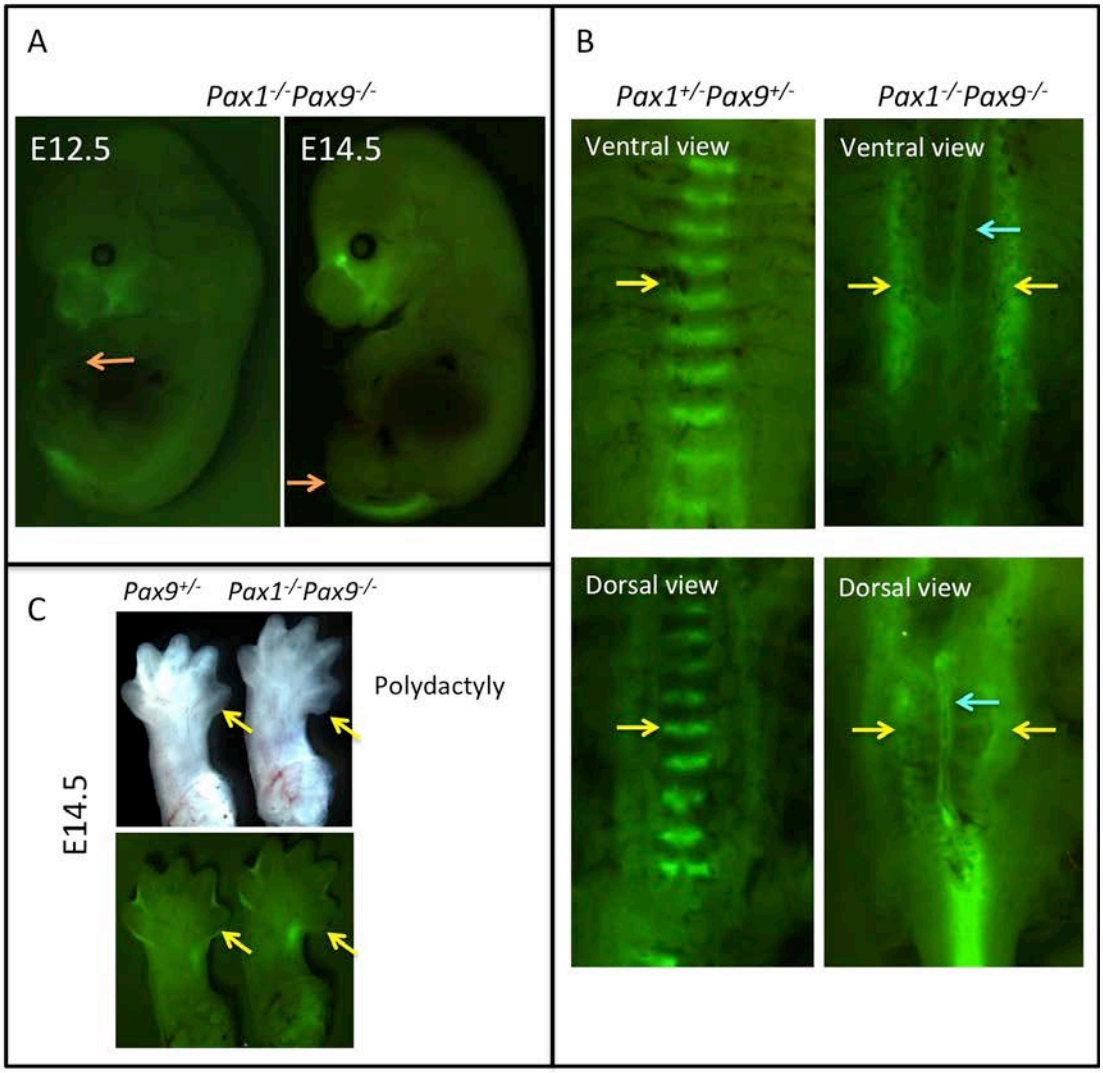


Figure 31: *Pax1/Pax9* double-null embryos. (A) Double-null embryos at E12.5 and E14.5. Tail tip is indicated with orange arrow. Tail was progressively shorter at older stage. (B) In the E14.5 *Pax1*$^{+/-}$*Pax9*$^{+/-}$ (double-heterozygote) embryo, fluorescence was seen in the IVD in a regular metameric fashion. In the double-null embryo, the fluorescing cells were mis-localized on the sides of the embryo with the notochord totally exposed in the middle (blue arrow). Bottom: dorsal view after the neural tube was removed. (C) Ventral view of hind limbs; In the *Pax9*$^{+/-}$ embryo, the anterior portion was normal; in the double-null embryo, signs of polydactyly could be seen in the anterior region.

3.4 Assessment of *Pax1* and *Pax9* mouse lines for TF mapping studies

The availability of inexpensive sequencing services and vast improvements in sequencing depth has led to the transition of technology from ChIP-Chip to ChIP-Sequencing (ChIP-Seq). ChIP-Seq has enabled a more biologically meaningful detection of TF binding sites in the DNA. While ChIP-Seq performed on cells with over-expressed TFs (cell culture) provide accurate information on the binding sites, they are not necessarily always biologically relevant and have to be assessed with several caveats in mind. For instance, critical binding sites may be missed in the absence of the appropriate co-factors in the cell culture-based assays, which is not a problem in an *in vivo* assay. Also, different co-factors may be required in a different biological context *in vivo* (eg. cell-type specific co-factors), which cannot be easily reproduced *in vitro*. Moreover, developmental-stage specific information is often lost in such cell culture experiments [5, 99].

Nonetheless, performing ChIP-Seq *in vivo* using embryonic tissues comes with limitations. In an over-expression cell culture system, one can easily use a construct whereby the TF-of-interest is tagged with a commonly used epitope, for which numerous antibodies are easily available. For an *in vivo* ChIP-Seq experiment, availability of a TF-specific ChIP-grade antibody, preferably commercially, is a pre-requisite. Raising antibodies against the TF-of-interest is time consuming and need not necessarily work in actual analytical experiments. The specificity of the antibody is also imperative. For example, since both the *Pax1* and *Pax9* TFs belong to the same subfamily, lack of cross-reactivity of the antibodies has to be verified. More importantly, even if specific antibodies are available, they may not be sensitive enough to immunoprecipitate a TF which is often expressed at moderate to low levels *in vivo* [100].

A simpler solution to these numerous impediments in studying protein function *in vivo* is to endogenously tag the protein-of-interest with a commonly used epitope. Commercial antibodies are readily available for various epitopes, such as HA, c-myc, VSV-G, FLAG, His, and also for an assortment of applications like Western blotting, immunoprecipitation, immunofluorescence/ immunohistochemistry and affinity purification [100, 101].

Such endogenous epitope-tagging of proteins by gene targeting has been widely implemented in the yeast for protein-DNA and protein-protein interaction studies [102-104]. However, it is much more challenging to perform this in the mammalian system owing to their unpredictable targeting efficiency. Use of BAC transgenes to express biotin-tagged proteins at close to endogenous levels in mouse ES cells has been carried out but it still does not enable the study of a protein function in a developmental stage-specific fashion [105].

In this study, mouse lines that stably express endogenously HA epitope tagged Pax1 and Pax9 proteins have been successfully generated (section 3.2.3). The $Pax1^{+/HA3}$ (heterozygote) and $Pax9^{HA3}$ (chimeric) embryos were assessed by immunohistochemistry using anti-HA, anti-Pax1 and anti-Pax9 antibodies (Figure 33 and 34). The HA antibodies successfully detected expression in the IVD region in both the $Pax1^{+/HA3}$ and $Pax9^{HA3}$ (chimeric) embryos and no expression was detected in the littermate WT embryos indicating the specificity of the HA antibody (Figure 33 and 34). The $Pax1^{+/HA}$ mice were also phenotypically normal, viable and fertile (Figure 32), and expressed the endogenous Pax1 and Pax9 in the correct domains as detected by the anti-Pax1 and anti-Pax9 antibodies (Figure 33). Thus, the triple HA epitope did not interfere with the biological function of the Pax1 protein.

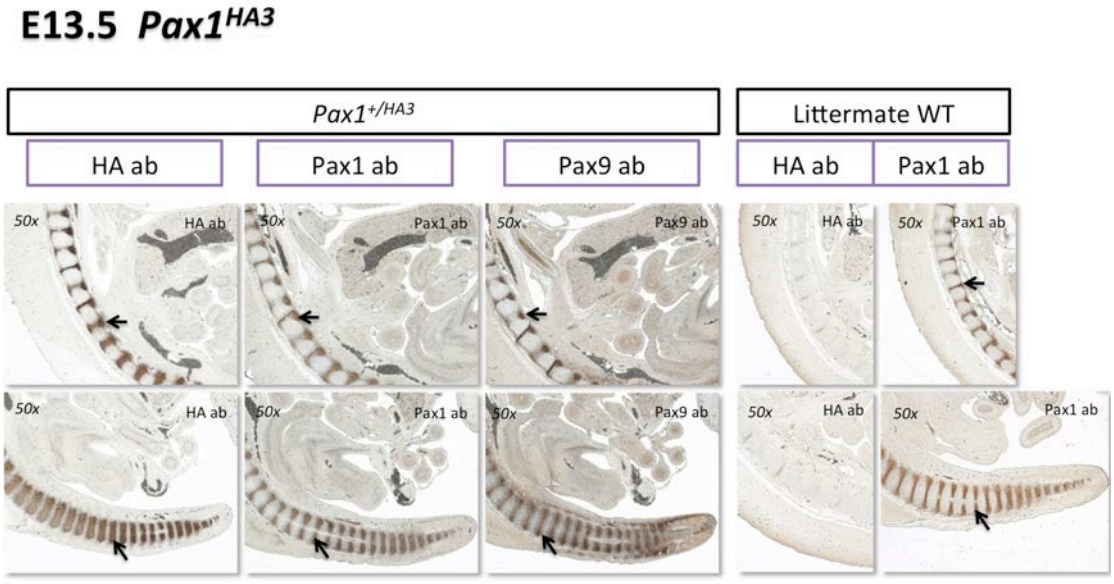
Figure 32: ***Pax1^{HA3}*** **mouse.** Picture of the adult *Pax1$^{+/HA3}$* mouse compared to the WT B6 mouse; mice did not exhibit any abnormalities and were viable and fertile. *N* – Neomycin.

Likewise, the HA-epitope was detectable in the *Pax9^{HA3}* embryos. Since the *Pax9^{HA3}* mice are still at the chimeric stage, only the high-percentage chimeric embryos were investigated. The expression of the HA epitope was patchy but in the Pax9-specific domains like the facial mesenchyme and the IVD anlagen. This was because of the chimeric nature of the embryo. Nonetheless, these embryos showed a homogenous expression of the endogenous Pax1 and Pax9 proteins as observed with the anti-Pax1 and anti-Pax9 antibodies (Figure 34). Moreover no defects were observed in the vertebral column or the facial mesenchyme. This verified the embryos to be normal.

Thus ascertained to be successfully tagged, these *Pax1^{HA3}* and *Pax9^{HA3}* mice will be profoundly useful in a myriad of proteomic experiments. For instance, this epitope tagging technique circumvents the potential cross-reactivity issue between the Pax1 and Pax9 proteins. Hence, they can be confidently used for ChIP-seq or immunoprecipitation experiments. More importantly, these proteins can now be

studied *in vivo* in a tissue-specific manner. In addition, the subcellular localization of these proteins can be studied *in vivo* through immunofluorescence coupled with confocal microscopy analyses [106, 107]. Potential changes in the protein localization or post-translational modifications in response to different treatments of the mice can also be determined [108, 109]. Besides, these mice are invaluable tools for protein-protein interaction studies. For example, the Pax1 and Pax9 proteins can be purified by affinity purification and subsequently used for mass spectrometry analysis to assess their protein partners in the different tissues (eg. vertebral column, thymus, tooth or facial mesenchyme etc) and at various developmental stages, which might assist in resolving their pleiotrophic mechanism of action.

Figure 33: Immunohistochemistry of E13.5 *Pax1^{HA3}* embryo. Left: The *Pax1$^{+/HA3}$* embryo paraffin sections of 10um thickness probed with the anti-HA antibody showed



expression in the intervertebral disc anlagen, similar to the expression pattern seen with anti-Pax1 and anti-Pax9 antibodies. Right: the littermate WT was used as a negative control to the anti-HA antibody and showed Pax1 expression when probed with the anti-Pax1 antibody. ab – antibody.

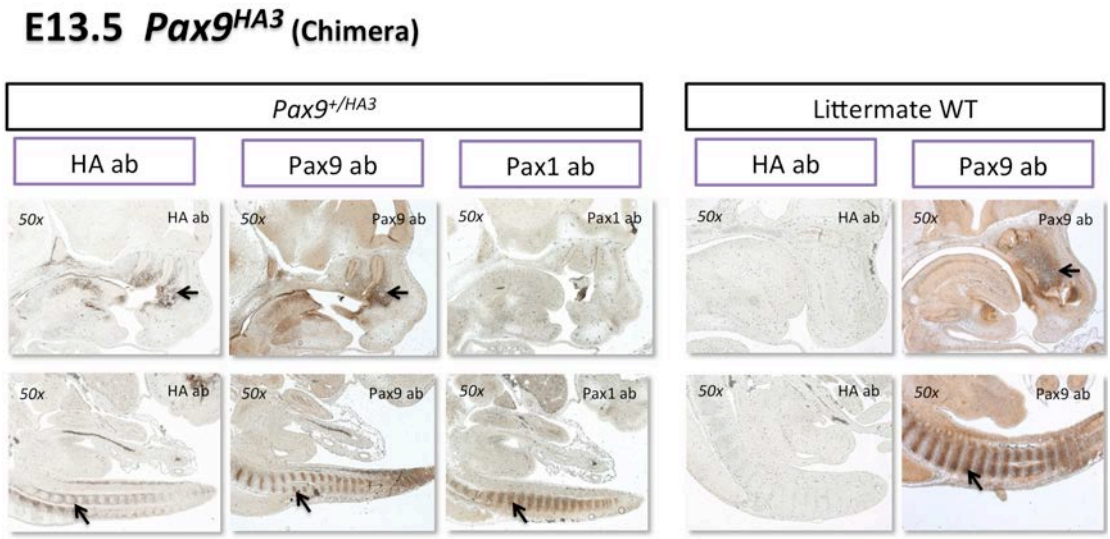


Figure 34: Immunohistochemistry of the E13.5 *Pax9*^{HA3} chimeric embryo. Left: The *Pax9*^{HA3} chimeric embryo paraffin sections of 10 um thickness probed with the anti-HA antibody showed expression in the facial mesenchyme and the intervertebral disc anlagen, similar to the expression pattern seen with anti-Pax9 antibody. Only the HA epitope expression appeared patchy owing to the chimeric nature of the embryo. Pax1 protein was seen in the intervertebral disc anlagen but not the facial mesenchyme which are Pax9-specifc regions. Right: the littermate WT was used as a negative control to the anti-HA antibody and showed Pax9 expression when probed with the anti-Pax9 antibody. ab – antibody.

3.5 Gene expression profiling - *Pax1* and *Pax9* targets in the vertebral column

Identification of the target genes of *Pax1* and *Pax9* in a cell-type specific manner necessitated the isolation of specific cell types (IVD analgen cells) from the embryonic stages. As discussed in section 3.4, mouse lines expressing EGFP under the control of *Pax1* or *Pax9* promoter were generated. From these mouse lines, fluorescing mouse embryos were harvested at the required stages.

For all the microarray gene expression profiling, fluorescing mouse embryos (*Pax1*$^{+/E}$, *Pax1*$^{+/-}$, *Pax1*$^{-/-}$, *Pax1*$^{-/-}$*Pax9*$^{+/-}$ and *Pax1*$^{-/-}$*Pax9*$^{-/-}$) were staged, dissected and dissociated into single-cells for sorting by fluorescence activated cell sorting (FACS). Only the dissected vertebral column tissue, including the tail (all the internal organs, limbs and head were discarded) was used for FACS and subsequent gene

expression analyses (microarray). Wild-type embryos of the same developmental stage, dissected and dissociated the same way as the fluorescing embryos, were used for gating during each FACS sorting experiment. Method of dissociation is described in detail in the Materials and Methods section. From the FACS-enriched pool of EGFP positive cells, RNA was extracted and converted to cDNA before a single-round of linear amplification was performed to obtain sufficient cDNA for biotinylation and hybridization on Illumina MouseWG-6 Expression BeadChips (Figure 35 for schematic diagram, Table 5 for RNA Integrity Number (RIN) values). For each array at least 3 (minimum) to 5 (maximum) biological replicates were used. Each biological replicate is comprised of cells pooled from multiple embryos (an average of 2 embryos), as very few fluorescing cells could be obtained from a single vertebral column (~ 2,000 – 4,000 cells), which are insufficient for even the downstream cDNA conversion and amplification. The Illumina bead chip was chosen for the gene expression profiling since it allows 6 samples to be assessed in parallel and it includes internal technical replicates. It also contains over 45,200 transcripts based on the latest mouse genome version, Build 36.

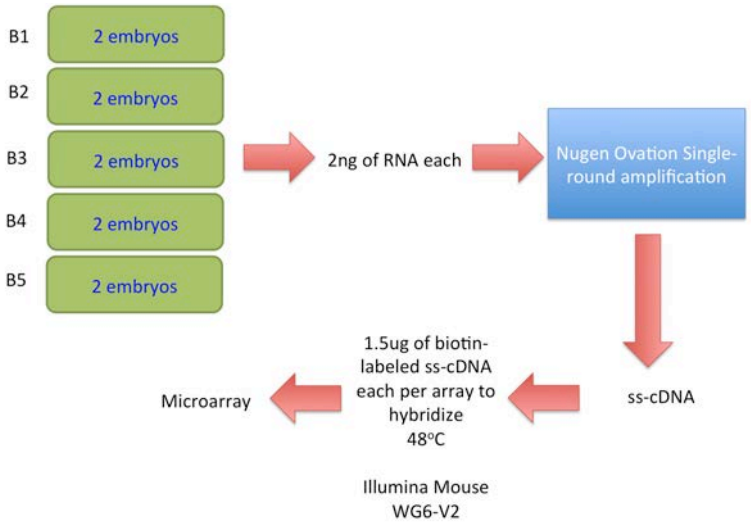


Figure 35: Schematic of FACS sorted cells used for microarray. For each biological replicate cells from an average of 2 embryos' vertebral column were used. From the extracted pool of total RNA, 2 ng was used for cDNA conversion and single-round amplification using NUGEN Ovation kit. The amplified single-stranded

DNA was then biotinylated, purified and used for hybridization on the Illumina MouseWG-6 Expression BeadChips according to the manufacturer's protocol. Bead chips were scanned by BeadScanner. B1-B5: biological replicates; ss – single-stranded.

Table 5: Quality of RNA extracted from E12.5 and E13.5 embryos for microarray

	E12.5		
Genotype	Biological replicate[#]	No. of cells	RIN
***Pax1*$^{+/E}$** **GFP(-) Control**	B1	34,269	7.4
	B2	17,051	8.2
	B3	31,628	8.4
	B4	15,793	7.5
***Pax1*$^{+/E}$** **WT**	B1	15,716	7
	B2	8,949	8.2
	B3	16,447	7.3
	B4	15,788	8.1
	B5	49,503	8.5
***Pax1*$^{+/-}$** **Het**	B1	4,614	8.7
	B2	32,892	9.4
	B3	11,738	9
	B4	Technical replicate of B2	
	B5	Technical replicate of B3	
***Pax1*$^{-/-}$** **Null**	B1	8,350	8.5
	B2	11,732	9.1
	B3	41,487	9.4
	B4	8,479	8.9
	B5	10,479	9.2
***Pax1*$^{-/-}$*Pax9*$^{-/-}$** **(Double-null)**	B1	18,882	8.2
	B2	16,030	8
	B3	12,598	8.8
***Pax1*$^{-/-}$*Pax9*$^{+/-}$** **3 allele KO**	B1	55,659	8.7
	B2	27,177	8.3
	B3	54,615	8.8
	E13.5		
***Pax1*$^{+/E}$** **GFP(-) Control**	B1	191,610	9.1
	B2	475,379	9
***Pax1*$^{+/E}$** **WT**	B1	8,173	7.7
	B2	11,228	7.3
	B3	15,554	8.2
	B4	10,463	7.8
***Pax1*$^{+/-}$** **Het**	B1	7,794	8.1
	B2	6,913	7.7
	B3	6,424	8.1
	B4	7,474	8
***Pax1*$^{-/-}$** **Null**	B1	14,008	9.6
	B2	13,968	8.7

	B3	10,763	8.2
	B4	13,643	8.2

Multiple embryos were pooled to obtain sufficient number of cells for each biological replicate; RIN – RNA Integrity Number.

3.5.1 Gene expression profile of *Pax1*-specific (GFP(+) cells) WT cells

To ensure that the FACS sorted cells were of the correct population i.e. *Pax1*-positive cells, and to further assess the set of genes that were enriched in *Pax1*-specific cells of the IVD anlagen, an expression profile on GFP(+) versus GFP(-) population was performed. Cells collected from the vertebral column of E12.5 and E13.5 $Pax1^{+/E}$ (WT), $Pax1^{+/-}$, $Pax1^{-/-}$ embryos were compared with the GFP(-) fraction of cells from the FACS sort of the corresponding developmental stages.

Using the GeneSpring software, one-way ANOVA and Benjamini-Hochberg multiple testing correction, with a fold-change cut-off of ≥1.5 fold and p-value ≤ 0.05, was applied to the gene expression data (Materials & Methods). As expected, *Pax1* was the most enriched gene in the $Pax1^{+/E}$, $Pax1^{+/-}$, $Pax1^{-/-}$ cells compared to the GFP(-) fraction of E12.5 and E13.5 stages (Table 6), indicating the success in enriching for the correct population of cells from these embryos. The *Pax1* transcripts will be detected in the microarray even in the $Pax1^{-/-}$ embryos since the method of disruption of Pax1 function is at at the translational level using the F2A peptide strategy (section 3.1). Hence, the probes on the microarray would still be able to hybridize to the undisrupted regions of *Pax1* transcript which includes any region 3 amino acids after the start of the second exon.

Table 6: *Pax1* fold enrichment compared to GFP(-) fraction of cells from E12.5 and E13.5 embryos

	Pax1 Fold enrichment		
Stage	WT vs. GFP(-)	$Pax1^{+/-}$ vs. GFP(-)	$Pax1^{-/-}$ vs. GFP(-)
E12.5	15.27	15.30	11.39
E13.5	13.64	15.83	9.52

In the E12.5 embryos GFP(+) cells compared to GFP(-) cells from $Pax1^{+/E}$ (WT) embryos, a total of 744 genes were up-regulated and 1,052 genes were down-

regulated. Similarly in E13.5 embryos, GFP(+) cells compared to GFP(-) cells from the *Pax1*$^{+/E}$ (WT) embryos, a total of 412 genes were up-regulated and 630 genes were down-regulated.

Besides *Pax1*, several other TFs were also enriched in these cells which are listed in Table 7. Of these, notable TFs are *Meox1*, *Meox2*, *Twist1*, *Sox5* and *Foxc2* (*Mfh1*) – genes which are either expressed in the IVD anlagen or known to be involved in the sclerotome and/or IVD development. In the E12.5 GFP(+) cells *Pax9* was also co-expressed with a 2.50-fold enrichment (lower than *Pax1*) and is also expressed in the IVD anlagen at E13.5 (Figure 36). Since *Foxc2* is known to genetically interact with *Pax1* in regulating the sclerotome cell proliferation, and both are co-expressed in the sclerotome up to E11.5, *Foxc2* expression pattern was checked at a later developmental stage of E13.5 (Figure 36).

Table 7: Transcription factors enriched in GFP(+) versus GFP(-) E12.5 *Pax1*$^{+/E}$ cells.

Transcription factors enriched in GFP(+) *Pax1*$^{+/E}$ vs GFP(-)	
Transcription factor activity E12.5 Enrichment score: 2.31* p-value: 0.00741	*E2f6, Sox5, Elk3, Gli2,* ***Pax1****, Wt1,* ***Pax9****, Foxf2, Etv1, Nfatc4, Tcf3, Etv5, Foxd1, Twist1, Sim2, Nfatc1, Egr1, Mafb, Foxa1, Creb5, Tbx1, Tead2, Six5, Six4, Foxp4, Foxp1, Meox2, Meox1, Trps1, Gtf2ird1, Foxc2, Foxc1, Tbx18, Nfia, Nfib*
Transcription factor activity E13.5 Enrichment score: 2.75* p-value: 0.00573	*Tshz2, Thra, E2f6, Nfix,* ***Pax1****, Gli1, Nkx6-2, Hlx, Foxf2, Nfatc4, Foxd1, Nfatc1, Mafb, Tead2, Bmyc, Creb5, Foxp4, Foxp1, Hoxd9, Hoxc10, Meox2, Meox1, Foxc2, Nfic, Nfia, Nfib, Sox5*

* - enrichment scores were derived from DAVID gene ontology analysis (ref: Materials and Methods).

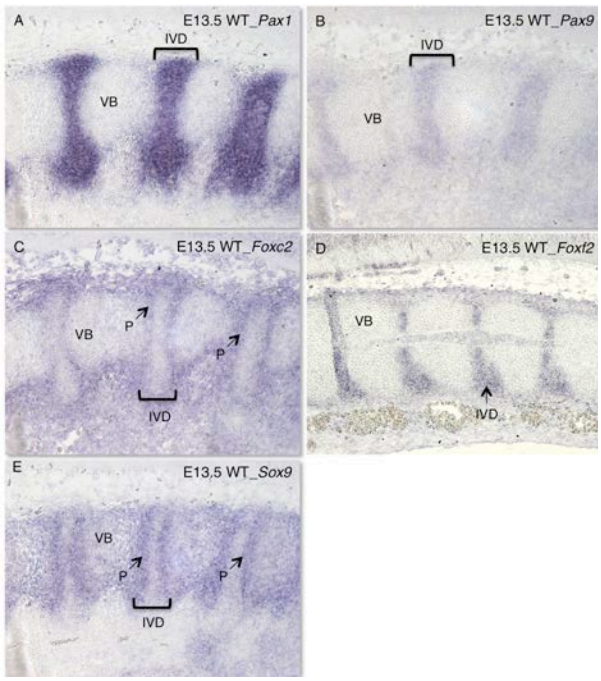


Figure 36: Expression patterns of *Pax1*, *Pax9*, *Foxc2*, *Foxf2 and Sox9* in E13.5 vertebral column sections. Paraffin sections of 10uM thickness were probed with the respective anti-sense RNA probes. Adjacent sections of the same region were used for expression analysis in A-C & E. (A-B) *Pax1* and *Pax9* expressions were seen in the IVD anlagen. (C) *Foxc2 (Mfh1)* was expressed mainly in the perichondrium of the vertebral bodies and diffusely in the tissues surrounding the vertebral column. (D) *Foxf2* expression was seen in the IVD anlagen. (E) *Sox9* was expressed in the VB and slightly more intensely in the perichondrium. VB- vertebral body; IVD – intervertebral disc anlagen; P- perichondrium. 200x magnification.

At E13.5, *Pax1* and *Pax9* were strongly expressed in the IVD anlagen, to a lesser degree in some of the vertebral body cells and in the mesenchymal cells surrounding the vertebral column. *Pax9* expression, however, was significantly weaker than *Pax1*. Likewise, *Foxf2* was also intensely expressed in the IVD anlagen. *Foxc2* (*Mfh1*) was expressed in the perichondrium and in the mesenchymal cells surrounding the vertebral column. It appeared as if *Foxc2* was outlining the borders of the vertebral bodies and the IVD. *Sox9* was expressed in the vertebral bodies, and slightly more intensely in their perichondrium. *Foxc2* and *Sox9* expression overlapped with that of *Pax1* and *Pax9* expression mainly in the perichondrium. *Sox9* did not appear to be enriched in the GFP(+) compared to the GFP(-) fraction because the vertebral body cells would have appeared in the GFP(-) fraction, which would

have expressed almost equally high levels of *Sox9* transcripts as the perichondrial cells which would be in the GFP(+) fraction. Hence, the slight difference in the expression levels between the vertebral body cells and the perichondrial cells may be below 1.5-fold and so missed by the fold-change cut-off.

In 2010, a gene expression profiling study was performed by Sohn *et al* (2010), whereby gene expression of mouse tissues from E13.5 laser-microdissected WT IVD anlagen was compared with the WT vertebral body tissues to identify IVD-enriched genes [60]. Their list of IVD-enriched genes was compared with the list of genes enriched in the GFP(+) cells. 46 genes and 23 genes were found to be overlapping with the E12.5 and E13.5 lists respectively. Some of those genes verified to be expressed in the IVD anlagen based on literature search/ gene expression database in the study by Sohn *et al* (2010) are shown in Table 8.

Table 8. List of some of the genes enriched in the E12.5 and E13.5 *Pax1*$^{+/E}$ GFP(+) cells and known to be expressed in the IVD anlagen.

Genes known to be expressed in the IVD anlagen & enriched in GFP(+) cells		
Stages	**Symbol**	**Name**
E12.5	*Col14a1*	Collagen, type XIV, alpha 1
E12.5	*Col6a1*	Collagen, type VI, alpha 1
E12.5	*Lrig1*	Leucine-rich repeats and immunoglobulin-like domains 1
E12.5	*Pax9*	Paired box gene9
E12.5	*Tgfb3*	Transforming growth factor, beta 3
E12.5	*Trps1*	Trichorhinophalangeal syndrome I (human); similar to Trps1 protein
E12.5	*Wisp1*	WNT1 inducible signalling pathway protein 1
E13.5	*Fgf18*	Fibroblast growth factor 18
E12.5 & E13.5	*Col6a2*	Collagen, type VI, alpha 2
E12.5 & E13.5	*Col6a3*	Collagen, type VI, alpha 3
E12.5 & E13.5	*Emilin3*	Elastin microfibril interfacer 3
E12.5 & E13.5	*Fmod*	Fibromodulin
E12.5 & E13.5	*Foxf2*	Forkhead box F2
E12.5 & E13.5	*Nfatc1*	Nuclear factor of activated T cells, cytoplasmic, calcineurin dependent 1
E12.5 & E13.5	*Pax1*	Paired box gene1
E12.5 & E13.5	*Vcan*	Versican

Besides this verified list of IVD-expressed genes, the entire list of *Pax1*-enriched genes clearly serves as a resource of genes that are co-expressed with

Pax1 in the IVD anlagen, and are probably involved in IVD development.

Gene Ontology analysis was performed using Database for Annotation, Visualization and Integrated Discovery (DAVID) web-based gene ontology enrichment analysis tool, to identify the set of TFs, biological processes, signalling pathways and molecular functions that were enriched in these *Pax1*-specific cells at E12.5 (only the genes up-regulated in *Pax1*-cells). A myriad of GO annotation terms were enriched and selected GO enrichment terms are shown in Figure 37. Genes were enriched for expected biological processes such as ECM, glycoprotein, somitogenesis, skeletal system development (cartilage development and osteoblast differentiation), pattern specification process, segmentation, regionalization, somitogenesis, anterior/posterior pattern formation, cell adhesion, collagen, cell migration, regulation of cell proliferation, regulation of cell size and regulation of cell-substrate adhesion among several others (Figure 37). Notably, GO terms were also enriched for face development, odontogenesis and inner ear morphogenesis, which are all associated with *Pax9* rather than *Pax1*. This could be because *Pax9* is co-expressed with *Pax1* in these cells (highlighted in red in the chart, Figure 37). At the same time, it was surprising to observe terms such as appendage development and inner ear morphogenesis to be enriched in these cells since they were obtained only from dissected vertebral column tissue. One possibility is that these genes have pleiotrophic roles, like *Pax1/Pax9*, and so are expressed in the vertebral column cells as well, potentially having a function in the axial skeleton. Evidently, not all the functions of all of the genes in the mouse genome have been completely studied and/or annotated yet, hence these genes may not yet have been discovered to play a role in the vertebral column.

The genes were also enriched for numerous intracellular signalling pathway components (receptors and ligands), such as TGF-beta signalling pathway,

Hedgehog signalling pathway, G protein coupled receptor and Wnt signalling pathway, all of which are involved in a number of cellular processes relevant to the formation of mesenchymal condensations and chondrogenesis/ osteogenesis. For instance, TGF-beta signalling pathway is mediated by secreted proteins (TGF-beta ligands, BMPs, activins, growth differentiation factors etc) and is involved in the regulation of cell differentiation, cell proliferation, cell migration, ECM production/ degradation, as well as endochondral and intramembranous ossification [110-112]. Moreover it has been shown to be critical in IVD development at the early embryonic stages [60, 61]. Hedgehog signalling components *Ptch1* (Patched receptor which represses the signalling component, Smo, by binding to it) and *Gli2* (downstream bi-potential TF, i.e. context-dependent activator or repressor) were also expressed in these *Pax1/Pax9*-positive cells indicating that they are still competent to mediate Hh signals at E12.5 [44, 113]. Similarly, Wnt signalling is also known to be vital in regulating MSC proliferation, differentiation and bone formation [114, 115].

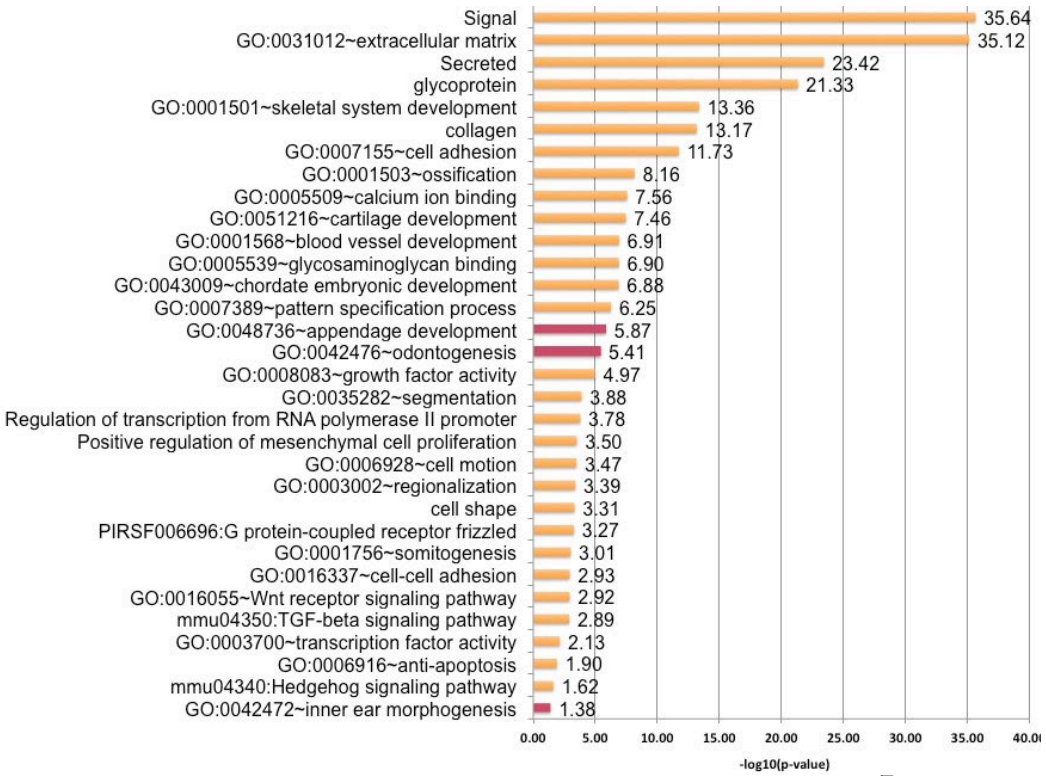


Figure 37: Gene Ontology term enrichment of E12.5 GFP(+) *Pax1*$^{+/E}$ cells. GO terms relevant to *Pax1* were enriched, including those relevant to *Pax9* (highlighted in red) since it was also co-expressed in the *Pax1*-specific cells. The -log10(p-value) are shown at the end of the bars; a value of more than 1.3 is deemed significant (-log10 transformation of a p value of 0.05 is equal to 1.3).

3.5.2 Genes regulated by *Pax1* – a temporal study

Confirming that the correct population of cells were enriched from the E12.5 and E13.5 embryos, *Pax1* regulated targets were then identified. First, a temporal study was performed by analysing the *Pax1* WT (*Pax1*$^{+/E}$) with the *Pax1*$^{-/-}$ for E12.5 and E13.5 stages. These stages were chosen since that is when IVD anlagen are first formed.

Only the WT and *Pax1*$^{-/-}$ (null) samples were compared using Student's unpaired T-test, p-value $\leq$ 0.05 and a FC $\geq$ 1.5, for E12.5 and E13.5. Since the *Pax1*$^{+/-}$ embryos did not show any defect in the vertebral column and the adult mice were reported to have very mild defects only (Table 3), they were excluded from further analysis [73]. A total of 130 genes (50 down-regulated and 80 up-regulated in

null) and 122 genes (47 down-regulated and 75 up-regulated in null) were differentially expressed at E12.5 and E13.5 respectively (Figure 38). Some of the down-regulated targets were randomly chosen and validated by sectioned *in situ* hybridization (Figure 39). Up-regulation is harder to show by sectioned *in situ* hybridization owing to limitations in the technique's sensitivity and saturation of signal.

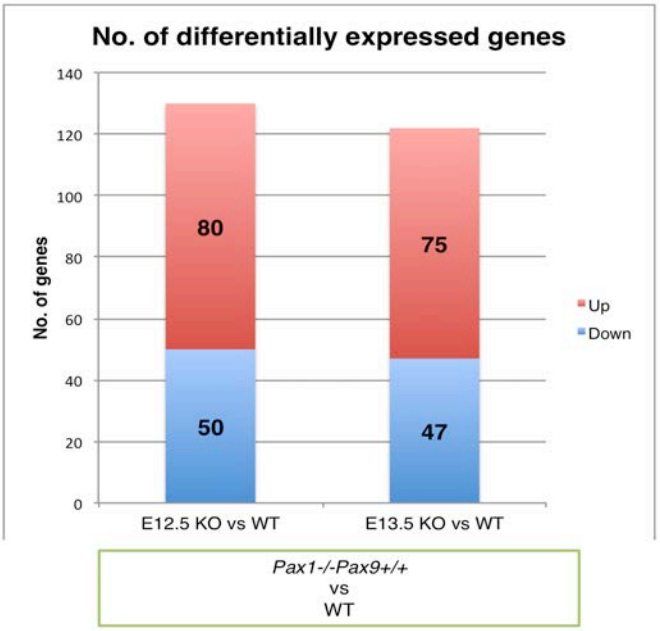


Figure 38: Total number of differentially expressed genes in E12.5 and E13.5 embryos (*WT* vs *Pax1*$^{-/-}$). The numbers of differentially expressed genes are shown for both the developmental stages – E12.5 & E13.5. Up refers to up-regulation in *Pax1*-null; Down refers to down-regulation in *Pax1*-null (KO); WT - *Pax1*$^{+/E}$. KO – knock-out.

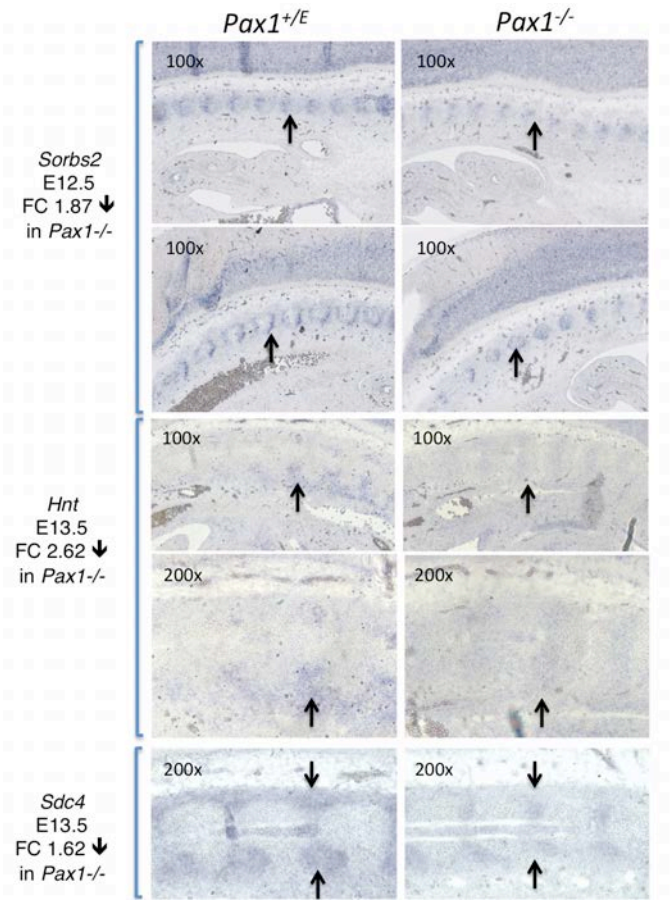


Figure 39: Gene expression profiling target validation for E12.5 and E13.5 by sectioned *in situ* hybridization. Sagital paraffin sections of E12.5 and E13.5 *Pax1*$^{+/E}$ (WT) and *Pax1*$^{-/-}$ embryos, at 10 uM thickness, were probed with *Sorbs2*, *Hnt* and *Sdc4* to validate their down-regulation seen in the (microarray) gene expression profiling. Pairs of sections for comparison were probed at the same time with the same concentration of probe and signals were developed for the same amount of time in the same regions of the embryos. FC – fold change. Dorsal region facing the top of the page; ventral towards bottom of page.

Gene Ontology analysis was then performed on the set of differentially expressed genes at E12.5 and E13.5 using DAVID**.** At E12.5, GO terms relevant to *Pax1* function were enriched (p-value $\leq$ 0.05) such as cell adhesion, cell junction, transcriptional regulation, transcription factor activity and macromolecule biosynthesis process (selected GO terms shown in Figure 40). At E13.5, the up- and down-regulated genes were enriched for several more GO terms like skeletal system development, collagen (triple helix), ECM, glycoprotein, regulation of apoptosis, ECM-receptor interaction, focal adhesion, cell-cell signalling, transcriptional regulation and transcription factor activity (selected GO terms shown in Figure 41).

TFs, cell adhesion and apoptosis related genes that were differentially expressed at both the stages are shown in Table 10.

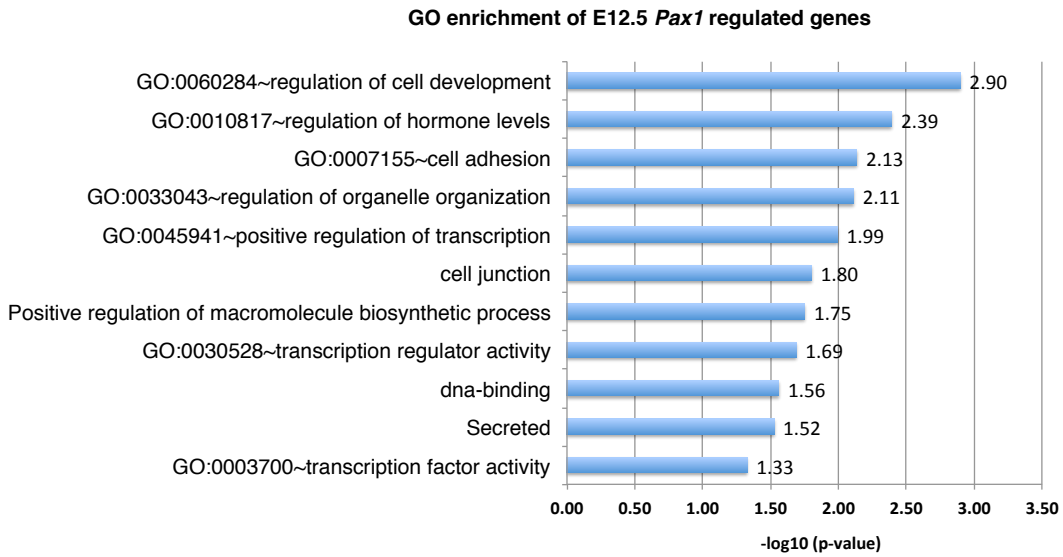


Figure 40: Gene Ontology term enrichment of E12.5 *Pax1* differentially expressed genes. GO terms relevant to *Pax1* were enriched like cell adhesion, macromolecule biosynthesis and transcriptional regulation. The -log10(p-value) are shown at the end of the bars; a value of more than 1.3 is deemed significant (-log10 transformation of a p-value of 0.05 is equal to 1.3).

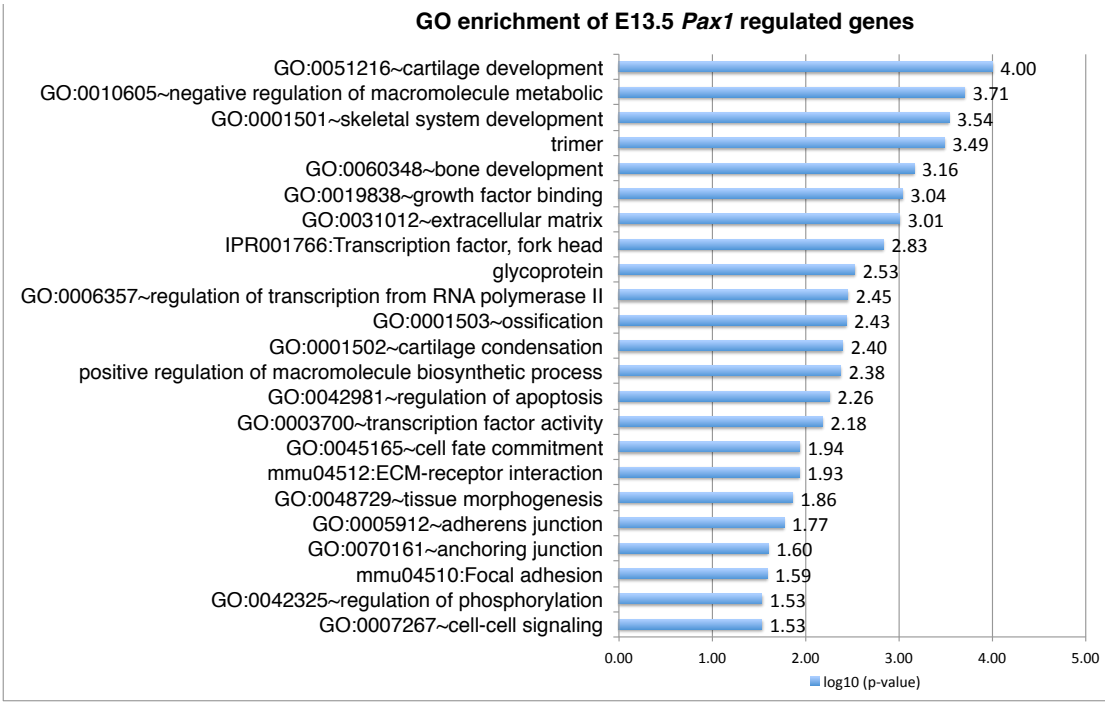


Figure 41: Gene Ontology term enrichment of E13.5 *Pax1* differentially expressed genes. GO terms relevant to *Pax1* were enriched such as cartilage development, skeletal system development, focal adhesion, ECM, glycoproteins and transcriptional regulation among several others. The -log10 (p-value) are shown at the end of the bars; a value of more than 1.3 is deemed significant (-log10 transformation of a p-value of 0.05 is equal to 1.3).

The differentially expressed genes of both stages were then overlapped. Surprisingly, only 7 genes (~ 5.56%) were common at both stages. They showed a consistent directionality and an increasing down-regulation or up-regulation (based on fold-change) with increasing developmental stage as expected (Table 9).

Table 9: Common genes differentially expressed in E12.5 & E13.5 embryos (*Pax1*$^{+/E}$ vs *Pax1*$^{-/-}$)

Common genes	E12.5*	E13.5*
Tbx3 (a TF)	Down (1.78)	Down (1.89)
Mela	Down (1.57)	Down (1.54)
Igfbp3	Down (1.51)	Down (1.77)
Trpc4ap	Down (1.54)	Down (1.59)
Crh	Up (2.51)	Up (2.85)
Foxa1 (a TF)	Up (1.56)	Up (1.83)
2010011I20Rik (Fam210b)	Up (1.50)	Up (2.99)

*Directionality - down in null; up in null; Fold change values are shown in parenthesis; TF – transcription factor.

Table 10: TFs, cell adhesion, apoptosis, migration, proliferation & ECM genes differentially expressed in *Pax1*$^{-/-}$

Genes differentially expressed		
Term	E12.5	E13.5
Transcription factor activity (GO:0003700)	*Egr1, Maf, Tbx3, Irx2, Foxa1, Emx2, Foxo1, Lef1, Olig1*	*Tbx3, Jun, Foxa1, Tbx20, Foxf2, Sox5, Klf16, Foxc2, Rarb, Sox9, Grhl2, Foxp2*
Cell adhesion (GO:0007155)	*Igsf11, Lama4, Pcdhb6, Lpp, Tek, Nlgn1, Pcdh12, Lef1, Col11a1*	*Col9a1, Ctgf, Tnc, Hspg2, Col2a1, Cdh1, Sox9*
Cell proliferation (GO:0042127 & GO:0010941)	*Tbx3, Myocd, Tek, Foxo1, Igfbp3*	*Tbx20, Foxc2, Sox9, Myh10*
Regulation of apoptosis (GO:0042981)	*Xrcc2, Tbx3, Foxo1, Igfbp3*	*Tbx3, Grid2, Foxc2, Col2a1, Cdh1, Ngfr, Rarb, Sox9, Igfbp3, Card10*
Cell migration (GO:0016477)	*Ednrb, Shroom2, Emx2*	*Tns3, Ctgf, Myh10, Tes*
Extracellular matrix (ECM) (GO:0031012)	*Lama4, Kera, Col11a1, Mmp24*	*Gpc4, Col9a1, Ctgf, Tnc, Thsd4, Col1a2, Hspg2, Col2a1*

3.5.3 Discussion

(I) *Pax1* plays a role in the mesenchymal condensation process

Based on these results, it appears that *Pax1* regulates genes involved in processes essential for condensation processes – cell adhesion (focal adhesion, adherens/ cell junctions), cell-cell signalling, ECM (glycoproteins) components and cartilage condensation. Condensation of mesenchymal cells is a pre-requisite for the formation of skeletal elements via endochondral ossification [116]. This in essence, requires sufficient number of cells to aggregate, form adhesions between each other to prevent them from migrating away, and to communicate information through intercellular signalling. Cell-matrix interactions are also vital in this communication between cells. The ECM (composed of glycoproteins and glycosaminoglycans, GAGs) acts as a medium for soluble factors and conveys signals from the surrounding region to the cells (or *vice versa*), which in turn activates intracellular signalling pathways. They also provide an attachment surface for cells during

migration. Furthermore, components of ECM like collagen and aggrecan enable water retention, and this osmotic swelling pressure in turn confers resistance to compressive forces. Boundaries also need to be established in condensations. Tenascin, an extracellular glycoprotein and its cell-surface receptor, syndecan, are known to help define the boundaries of condensations, potentially through cell-cell and cell-ECM interactions [116-118]. Both *Sdc4* (*Syndecan 4*) (1.62-fold down in *Pax1*$^{-/-}$) and *Tnc* (*Tenascin C*) (1.63-fold up in *Pax1*$^{-/-}$) were found to be differentially expressed in the E13.5 *Pax1*$^{-/-}$ embryos.

Pax1, along with *Pax9*, is known to promote sclerotome proliferation [62]. In the analysis of *Pax1*-null cells, genes involved in proliferation were found, but not enriched. This could be because the prior study was performed in *Pax1*$^{-/-}$*Pax9*$^{-/-}$ double-null mutants which have a more severe phenotype, compared to this analysis whereby only the *Pax1*$^{-/-}$ embryos were analyzed. The analysis of double-null mutants may reveal an enrichment for genes involved in proliferation.

In the study of *Pax1/Pax9* double-null embryos, the authors also observed an increase in apoptosis. However, the authors claimed apoptosis might not be a true function of these two *Pax* genes, and the increase in apoptosis might be a causative effect of drastically reduced cell numbers in the mesenchymal condensations. In the *Pax1*$^{-/-}$, genes involved in apoptosis were also found to be differentially expressed but not enriched. Identification of the direct targets of *Pax1* and *Pax9* might help to reveal if these *Pax* genes truly regulate apoptotic processes or if it just a secondary effect as the authors had predicted [62]. The direct targets of *Pax1 and Pax9* are discussed in later sections (section 3.6.4).

Thus, the hypothesis that *Pax1* regulates processes involved in mesenchymal condensation formation and differentiation appears valid and some of the genes

involved in these processes have been identified in this study through a genome-wide approach [62, 76].

(II) Different sets of genes are regulated at different time-points

Although some of the processes regulated by *Pax1* at both stages are similar, the genes involved are quite different. For instance, genes involved in transcriptional regulation, cell adhesion, proliferation, apoptosis, migration and ECM were found at both stages, but the genes corresponding to each of these functions were different (Table 10). One reason for the lack of a large overlap could be because of the inherent differences and dynamic changes in the developmental processes itself. It is between E12.5 and E13.5 that the IVD anlagen is formed and those cells are in the process of further differentiation to give rise to a hyaline cartilage for the inner annulus and a fibrous outer annulus by E14.5 [63] (See [119] for histological sections of mouse IVD from E12.5 to E15.5 & Figure 64). Another possibility is that besides *Pax9*, other TFs involved in the same or parallel pathway were also compensating for the loss of *Pax1*. For example, *Foxc2* and *Sox9* were up-regulated at E13.5 in the *Pax1*$^{-/-}$ mutants. *Foxc2* (*Mfh1*) and *Pax1* are known to cooperatively regulate the mitotic activity of sclerotome cells [41]. Likewise, *Sox9* is essential for chondrocyte differentiation and production of cartilage-specific ECM proteins like collagen and aggrecan, which are major components of ECM [52, 120-123]. Indeed, several of the known targets of *Sox9* like *Col2a1* [121], *Wwp2* [124] and *Sox5* [125] were up-regulated in the E13.5 *Pax1*$^{-/-}$ as well (Table 12). At E12.5, the IVD anlagen are first formed. Hence, in response to the cells sensing the dysregulation of some of the genes in E12.5 *Pax1*$^{-/-}$, compensation may have been kick-started slightly later, the effects of which are observed only at E13.5. Then again, it is not clear at this stage if

Pax1 was instead repressing *Foxc2* and *Sox9* (in WT embryos). Analysis of E13.5 *Pax1/Pax9* double-null mutants may reveal their true regulation.

(III) A small number of genes are differentially expressed in the *Pax1*$^{-/-}$ mutants

It appears that the knock-out of *Pax1* alone affects only a small population of genes. The total number of genes changing at each stage (E12.5 and E13.5) was also about the same. This was not too surprising since at these developmental stages, the only morphological defects seen in the *Pax1*$^{-/-}$ embryos were a loss of vertebral body and IVD cells, mainly in the lumbo-sacral region (Figures 26 & 27; similar to what has been reported before [73, 76]). It is well known that a critical number of cells are needed to form a condensation, and that the size of the condensation in turn determines the size of the skeletal element formed later [116, 118]. Thus, the loss of cells from particular vertebral segments probably translates to the absence of or small-sized vertebrae in the lumbo-sacral region of the adults, which in turn leads to a misaligned vertebral column giving rise to a kinked-tail phenotype.

Another reason for the small number of differentially expressed genes could be compensation by *Pax9*, thus masking most of the genes regulated by *Pax1* in an analysis of just *Pax1*$^{-/-}$ with WT. Indeed, the paralogous genes *Pax1* and *Pax9* are known to synergistically regulate axial skeleton development [62]. Therefore, to uncover these masked genes (genes compensated by *Pax1/Pax9*), it is essential to analyze the double KO embryos.

3.5.4 Genes regulated by both *Pax1* and *Pax9*

As it was evident that majority of the genes regulated by *Pax1* was masked by *Pax9*, and a gene-dosage effect of *Pax1* and *Pax9* on axial skeleton formation has

been reported before, multiple allele KO embryos were collected for gene expression analysis. This included the *Pax1*$^{-/-}$*Pax9*$^{+/-}$ (3 allele KO) and *Pax1*$^{-/-}$*Pax9*$^{-/-}$ (4 allele KO, i.e. double-null) mutants. A rate-limiting step for this analysis was obtaining sufficient number of cells for gene expression profiling as the chances of getting a 3 allele KO and 4 allele KO mutants from mating the double heterozygotes (*Pax1*$^{+/-}$*Pax9*$^{+/-}$) are 2 in 16 and 1 in 16 respectively (based on Mendellian ratio of inheritance). Also, the 3 allele KO mutants do not survive (*Pax1*$^{-/-}$*Pax9*$^{+/-}$ and *Pax1*$^{+/-}$*Pax9*$^{-/-}$ undergo postnatal lethality), which precluded them for use in mating. Hence, from mating a pair of double heterozygotes, each of the embryos had to be sorted separately and subsequently their genotypes were determined by PCR genotyping. Therefore, considering the cost-, time- and labour-intensiveness of this study, the analysis was restricted to just *Pax1*$^{-/-}$*Pax9*$^{+/-}$ and *Pax1*$^{-/-}$*Pax9*$^{-/-}$, and the *Pax1*$^{+/-}$*Pax9*$^{-/-}$ could not be analyzed. In the double-null embryos, the first signs of abnormalities in cell proliferation and apoptosis of sclerotomal cells and an arrest of chondrogenesis was observed at E12.5, so the gene expression profiling was performed for embryos at E12.5 [62].

In total, 3 biological replicates (2 embryos each) worth of cells were collected for each of the *Pax1*$^{-/-}$*Pax9*$^{+/-}$ and *Pax1*$^{-/-}$*Pax9*$^{-/-}$ genotypes (Table 4 for RIN values). Owing to time and fiscal constraints new *Pax1*$^{+/E}$ (WT) and *Pax1*$^{-/-}$ samples could not be collected. Hence, cDNA from the *Pax1*$^{+/E}$ (WT) and *Pax1*$^{-/-}$ embryos, left-over from the biological replicates used in the prior analysis (E12.5 *Pax1*$^{-/-}$ vs WT), were biotin-labelled and used in the new microarray chips alongside the multiple-allele KOs (3 biological replicates each). Since the *Pax1*$^{+/E}$ (WT) and *Pax1*$^{-/-}$ samples were stored in the cDNA form at -20^{o}C, degradation ought to be minimal. Nevertheless, to improve the accuracy and statistical strength of the results, the new gene expression data set was combined with the old data set. That is, the raw intensity data from

these new multiple allele KO gene expression results and the old gene expression results (E12.5 *Pax1*$^{+/E}$ (WT), *Pax1*$^{+/-}$ and *Pax1*$^{-/-}$) were loaded on to GeneSpring GX™ 11.0 software and quantile normalized. Any differences between both batches of data were successfully mitigated by quantile normalization. This allowed the data to be comparable across all of the new and old samples. A gene-level analysis and a more stringent statistical testing (one-way, Welch ANOVA; i.e. unequal variance) was performed with Benjamini-Hochberg multiple testing correction (protects against false-positives) and Student Newman-Keuls (SNK) post-hoc test. Only genes with p-value $\leq$ 0.05 and FC $\geq$ 1.5 were deemed differentially expressed.

3.5.4.1 Differential gene expression analysis of multiple allele knock-out

The various comparisons made in the multiple-allele KO analysis are shown schematically in Figure 42. The assumptions for these comparisons are as follows: the comparison of *Pax1*$^{-/-}$ with WT would reveal targets uniquely regulated by *Pax1*, direct or indirect. Likewise, the targets regulated by both *Pax1* and *Pax9* can be revealed by a comparison of WT with the 3 allele KO and 4 allele KO. The comparison of *Pax1*$^{-/-}$ with a 4 allele KO would give the targets regulated by 2 copies of *Pax9* in the absence of *Pax1*, while the comparison of the 3-allele KO with the 4 allele KO would reveal the targets regulated by a single copy of *Pax9* in the absence of *Pax1*. By default, these targets are assumed to be *Pax1* targets as well since in the *Pax9*$^{+/-}$ and *Pax9*$^{-/-}$ embryos, no vertebral column defect is observed, and *Pax1* is believed to compensate for loss of *Pax9*.

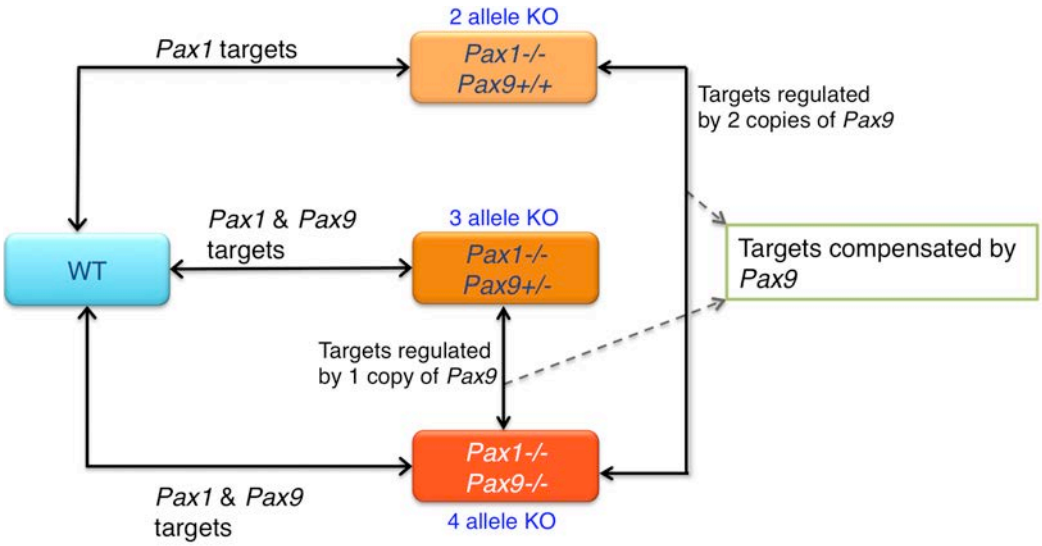


Figure 42: Schematic of multiple allele knock-out comparisons and the potential targets they would reveal. WT – wild-type; KO – knock-out.

In the double-null vs WT analysis, a total of 599 genes were differentially expressed (326 genes down- and 273 genes up-regulated in the double-null embryos). This is about six times more than what was seen in the E12.5 *Pax1*$^{-/-}$ vs WT (130 genes). With the *Pax1*$^{-/-}$*Pax9*$^{+/-}$ (3 allele KO) vs WT, a total of 467 genes were differentially expressed. Also, in the comparison of *Pax1*$^{-/-}$ vs double-null (genes regulated by 2 copies of *Pax9*), 844 genes were differentially expressed. Whereas the comparison of *Pax1*$^{-/-}$*Pax9*$^{+/-}$ (3 allele KO) vs *Pax1*$^{-/-}$*Pax9*$^{-/-}$ (4 allele KO) (genes regulated by 1 copy of *Pax9*) showed only 120 genes to be differentially expressed. It was quite unexpected that 2 copies of *Pax9* could be regulating many more genes than *Pax1* and *Pax9* combined (844 genes vs 599 genes), when *Pax1* is believed to play a bigger role in axial skeletogenesis. However, the actual gene regulation might be more complex and not necessarily have a linear change in terms of number of genes that are differentially expressed. Nevertheless, these results indicate that the comparison of the *Pax1*$^{-/-}$ with the 3 alleles or 4 allele KO can reveal genes regulated by *Pax9* in the absence of *Pax1* (genes masked by redundancy) (Figure 43).

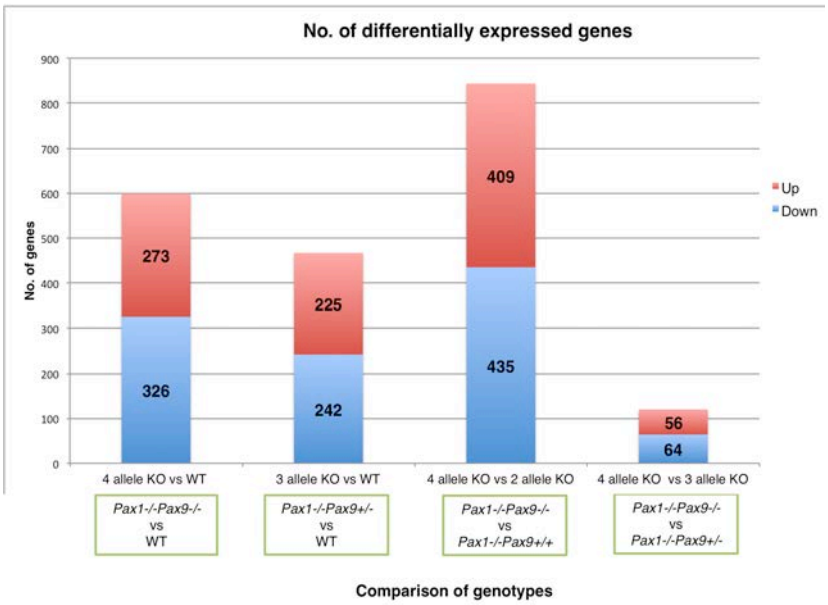
Gene Ontology enrichment analysis was performed on the set of differentially expressed genes in the E12.5 WT vs double-null comparison (599 genes) using DAVID. As expected, several more GO terms were enriched compared to what was seen in the E12.5 *Pax1*$^{-/-}$, such as ECM, cartilage development, skeletal system development, cell adhesion and apoptosis. Interestingly, these terms were also enriched in the E13.5 *Pax1*$^{-/-}$. GO terms that were not found in either of the E12.5 or E13.5 *Pax1*$^{-/-}$, but were only enriched in the double-null were collagen fibril organization, pattern binding, thyroid metabolic process, blood vessel development, cell motion, TGF-beta binding, platelet derived growth factor binding, appendage development and epithelial cell proliferation (highlighted in red in Figure 43 B; details of the genes involved are shown in Table 11). Validation of some of the targets was performed by sectioned *in situ* hybridization on the *Pax1*$^{-/-}$*Pax9*$^{+/-}$ embryos as more of the *Pax1*$^{-/-}$*Pax9*$^{-/-}$ embryos could not be obtained (Figure 44).

Table 11: Genes enriched for selected GO terms in the *Pax1*$^{-/-}$*Pax9*$^{-/-}$ mutants

GO terms	Genes
Collagen fibril organization	***Col3a1***, ***Acan**** (↓), ***Col2a1**** (↓), *Lox*, ***Col11a1**** (↓), ***Col5a2****, ***Dpt*** (↓), ***Atp7a*** (↓)
Cell motion genes	***Psen1**** (↓), ***Ctgf**** (↓), ***Etv1*** (↓), ***Reln***, *Strbp* (↓), ***Nr2f2, Boc, Foxd1,*** *Plxna2* (↓), *Evl*, *Cdh4* (↓), ***Alcam***, ***Sema6a*** (↓), ***App, Rpl24**** (↓), ***Nrp,*** *Gpx1* (↓), *Pafah1b2*
TGF-beta binding	***Ltbp1,*** *Fbn1**, ***Fbn2***
PDGF-binding	*Col4a1* (↓), ***Col3a1, Col2a1**** (↓)
Epithelial cell proliferation	***Bmp4*** (↓), ***Psen1**** (↓), ***Col8a1,*** *Col8a2*
Cartilage development	***Bmp4*** (↓), ***Atp7a*** (↓), *Mef2c*, *Lect1*, ***Ctgf**** (↓), ***Acan**** (↓), ***Sox5**** (↓), ***Hspg2**** (↓), *Gnas*, ***Col2a1**** (↓), ***Col11a1**** (↓)
Skeletal system morphogenesis	***Bmp4*** (↓), *Mef2c*, ***Hspg2**** (↓), ***Col2a1**** (↓), ***Pax1**** (↓), ***Hoxc9****, ***Psen1**** (↓), ***Ctgf**** (↓), *Ankrd11**, ***Hoxd4**** (↓), ***Acan**** (↓), *Gnas*, ***Col11a1**** (↓)

Genes in bold are known to be expressed in the IVD anlagen (inner or outer annulus fibrosus) or somites; genes which are down-regulated in the *Pax1*$^{-/-}$*Pax9*$^{-/-}$ mutants are indicated by the blue arrows. * - genes with associated axial skeleton defects.

A



B

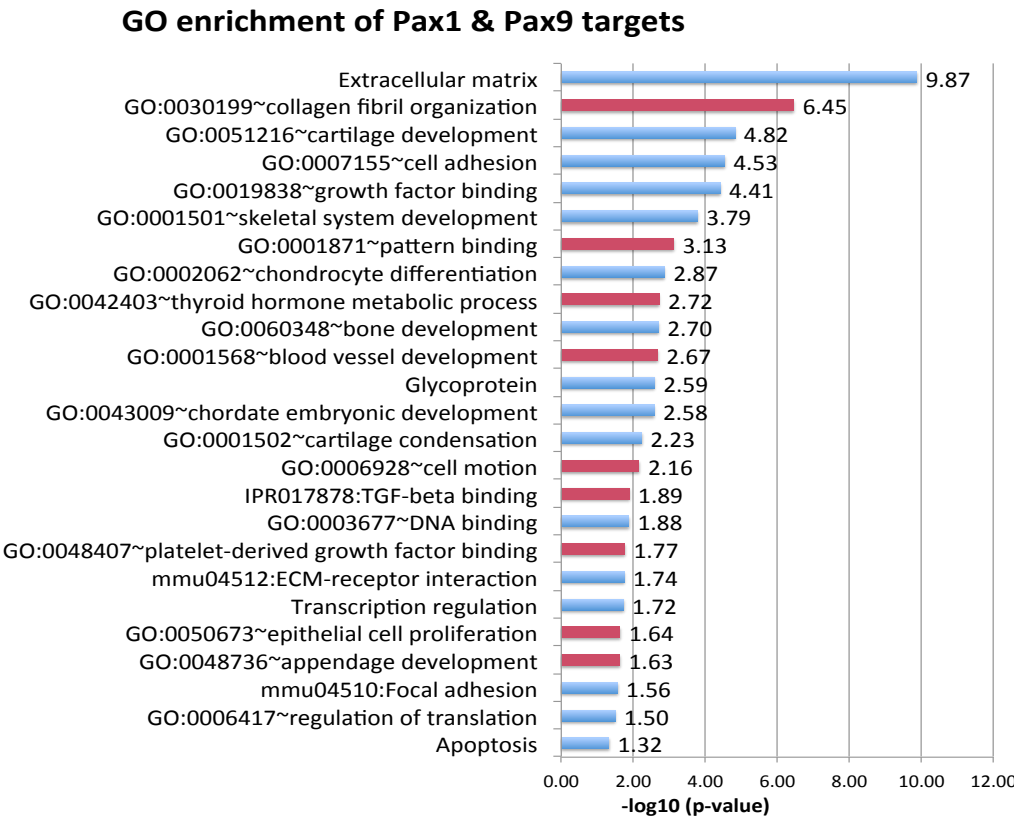


Figure 43: Number of differentially expressed genes in multiple allele KO comparisons and the GO enrichment. (A) Bar graph of all the differentially expressed genes. (B) GO enrichment of differentially expressed genes from the *Pax1*$^{-/-}$*Pax9*$^{-/-}$ vs WT comparison (*Pax1* and *Pax9* targets). New GO terms that were not found in the E12.5 or E13.5 *Pax1*$^{-/-}$ vs WT comparisons are highlighted in red. The -log10 (p-value) are shown at the end of the bars; a value of more than 1.3 is deemed significant (-log10 transformation of a p-value of 0.05 is equal to 1.3).The directionality is always in relation to the genotype shown first. i.e. in *Pax1*$^{-/-}$*Pax9*$^{-/-}$ vs WT, Up refers to genes up-regulated in *Pax1*$^{-/-}$*Pax9*$^{-/-}$ compared to WT. WT – wild-type; KO – knock-out.

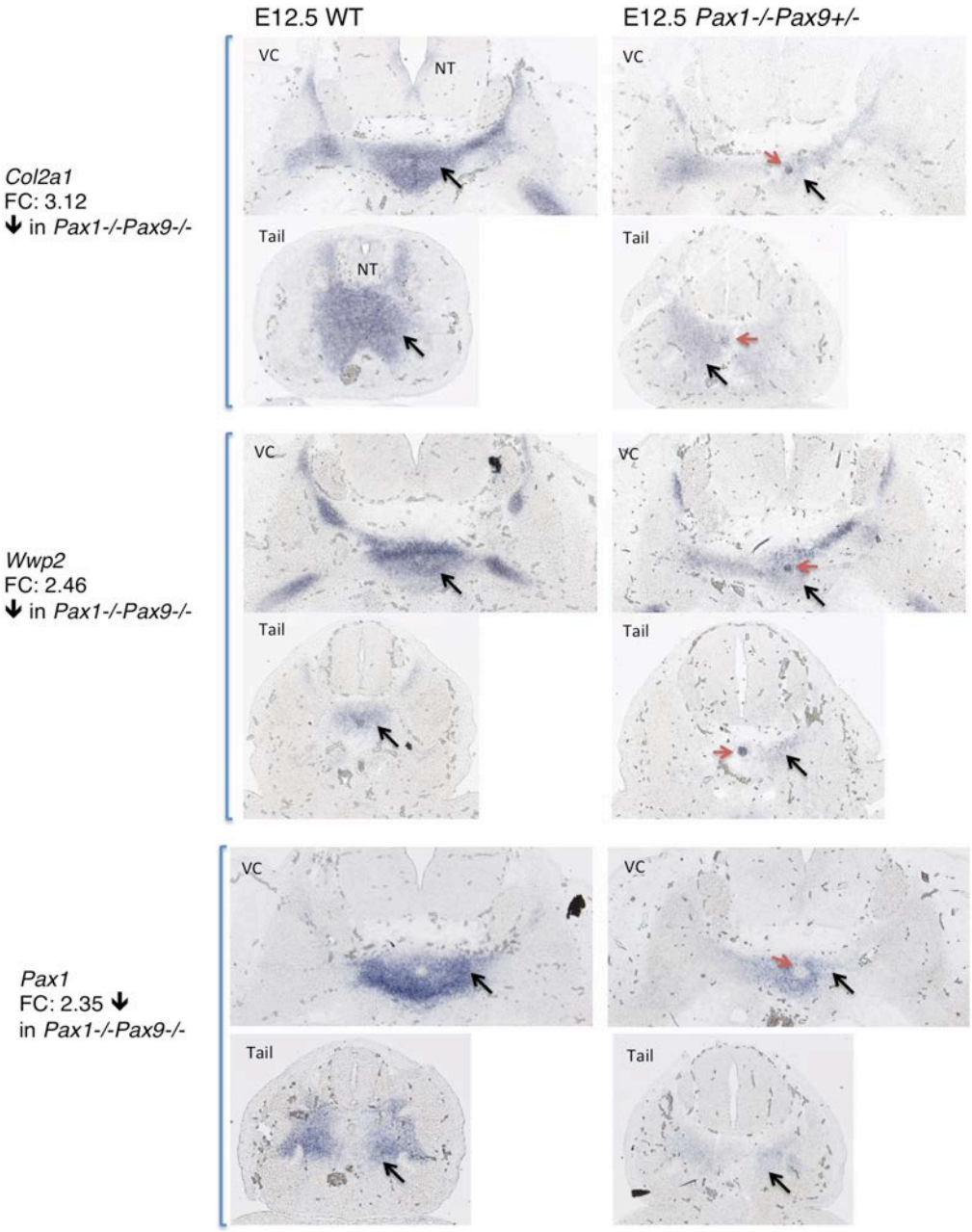


Figure 44: Validation of selected targets by sectioned *in situ* hybridization. Respective anti-sense probes were hybridized on 10 uM thick paraffin sections of E12.5 WT and *Pax1*$^{-/-}$*Pax9*$^{+/-}$ embryos. Adjacent transverse sections from lumbo-sacral regions were used. For *Col2a1* and *Wwp2*, the expression in the notochord is unaffected since *Pax1* and *Pax9* are not expressed there. Red arrows point to the notochord. All the pictures were taken at a magnification of 100x. NT – neural tube; FC – fold change.

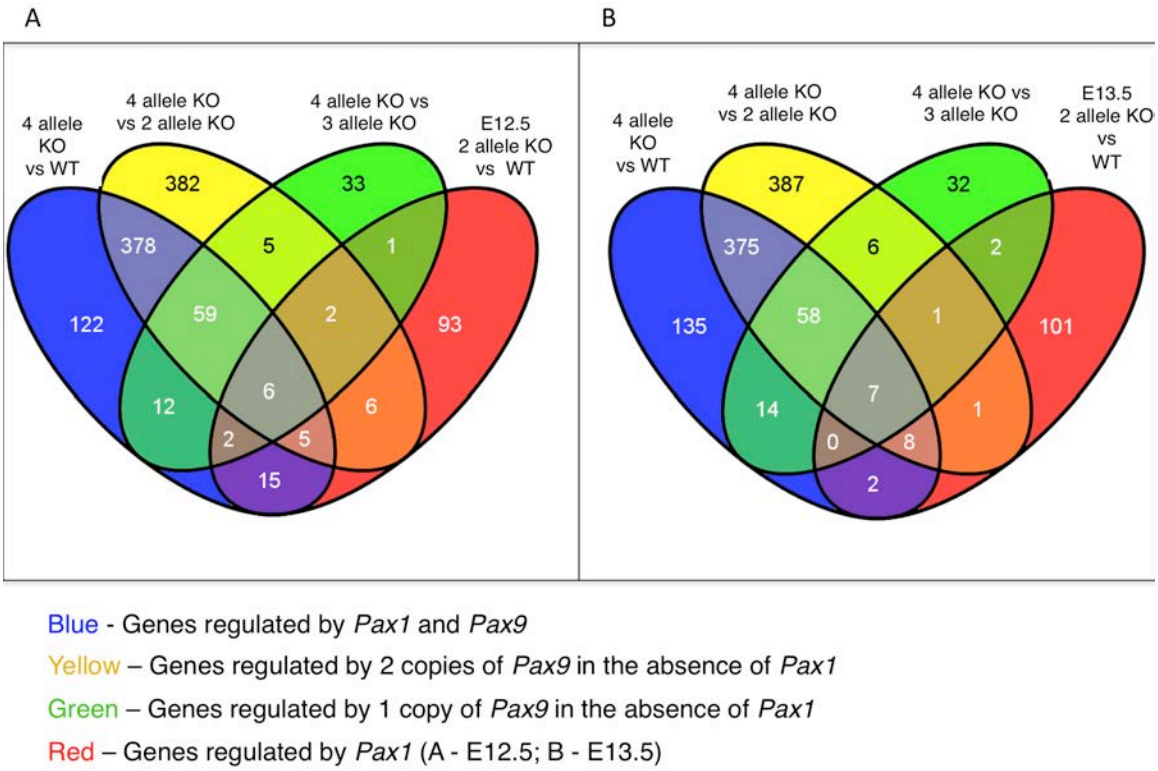


Figure 45: Venn diagram of overlap of genes from the different genotype comparisons. 4-allele KO – *Pax1*$^{-/-}$*Pax9*$^{-/-}$; 3-allele KO – *Pax1*$^{-/-}$*Pax9*$^{+/-}$; 2-allele KO – *Pax1*$^{-/-}$. WT – wild-type; KO – knock-out.

The differentially expressed genes were then overlapped for the different microarray gene expression comparisons to elucidate the trend in the regulation of genes. The various degrees of overlaps can be seen from the Venn diagram in Figure 45. The biggest overlap was between blue and yellow segments - genes regulated by both *Pax1* and *Pax9* (4 allele KO vs WT - blue) and genes regulated by 2 copies of *Pax9* in the absence of *Pax1* (4 allele KO vs 2 allele KO - yellow). These overlapping set of genes (448 genes) are probably those *Pax9* was regulating to compensate for the loss of *Pax1*. Likewise, 79 of the genes regulated by a single copy of *Pax9* in the absence of *Pax1* (green), overlapped with the 4 allele KO vs WT (blue). These are the genes which change when the last copy of *Pax9* is also removed in the absence of *Pax1*. Besides these, a total of 28 genes (21.5%) from E12.5 *Pax1*$^{-/-}$ and 17 genes (13.9%) from E13.5 *Pax1*$^{-/-}$ were overlapping with the

genes regulated in the 4 allele KO (blue overlap red). Of the 28 genes from E12.5 *Pax1*$^{-/-}$, 27 of them showed the same directionality with the 4 allele KO. On the contrary, from the 17 genes in the E13.5 *Pax1*$^{-/-}$ only 4 showed the same directionality, while 13 of them showed opposite directionality. 11 of the 13 genes were up-regulated in the E13.5 *Pax1*$^{-/-}$ but down-regulated in the E12.5 *Pax1*$^{-/-}$*Pax9*$^{-/-}$ (Table 12). Also, those set of genes were not differentially expressed in the E12.5 *Pax1*$^{-/-}$. The fold change values and direct binding sites of these genes are provided in section 3.6.4, Table 13.

Table 12: Genes with opposite or same directionality in double-null and *Pax1*$^{-/-}$

Gene symbol	E12.5 *Pax1*$^{-/-}$*Pax9*$^{-/-}$ vs. WT	E13.5 *Pax1*$^{-/-}$ vs. WT
Col2a1	Down (3.12)	Up (2.21)
Wwp2	Down (2.46)	Up (1.60)
Sox5 / A730017D01Rik	Down (2.23)	Up (1.63)
Extl1	Down (1.78)	Up (3.08)
Ctgf	Down (1.57)	Up (1.74)
Nnat	Down (2.62)	Up (1.78)
Greb1	Down (1.85)	Up (1.54)
Cdc25c	Down (1.79)	Up (1.95)
Csrp2	Down (1.75)	Up (1.52)
Tle1	Down (1.68)	Up (1.63)
Sorl1	Down (1.54)	Up (1.57)
Anxa11	Up (1.87)	Down (1.55)
Plagl1	Up (1.54)	Down (2.16)
4833440I11Rik	Down (2.23)	Down (1.95)
Mgst2	Down (2.39)	Down (2.39)
Hspg2	Down (2.37)	Down (1.66)
Crh	Up (6.04)	Up (2.85)
Gene symbol	**E12.5 *Pax1*$^{-/-}$*Pax9*$^{-/-}$ vs. WT**	**E12.5 *Pax1*$^{-/-}$ vs. WT**
Crh	Up (6.04)	Up (2.51)
Cxcl13	Up (4.07)	Up (1.91)
C030002B11Rik/ Ppm1h	Up (2.39)	Up (1.66)
AI593442	Up (3.34)	Up (2.08)
Tek	Up (3.29)	Up (1.65)
Cbln2	Up (2.97)	Up (1.86)
Myocd	Up (3.95)	Up (1.86)
Tcfec	Up (2.88)	Up (1.79)
Vstm2	Up (1.77)	Up (1.59)

Maf	Up (2.76)	Up (2.17)
Plekhd1/ 3830431G21Rik	Up (2.12)	Up (1.65)
Luzp2	Up (2.10)	Up (1.95)
Mid1	Up (2.01)	Up (1.75)
Tmem87a	Up (1.96)	Up (2.17)
Tmem106c/D15Ertd405e	Up (1.89)	Up (1.56)
Kdt1	Up (1.69)	Up (1.72)
Pdzrn4/1110017D07Rik	Up (1.68)	Up (2.05)
Nlgn1	Up (1.58)	Up (1.64)
6820402O20Rik/ Phf20	Down (2.14)	Down (1.52)
Kcnrg	Down (2.11)	Down (1.59)
Crym	Up (1.68)	Down (1.56)
Sorbs2/ 9430041O17Rik	Down (2.08)	Down (1.87)
Kcna1	Down (2.33)	Down (1.76)
Col11a1	Down (1.85)	Down (2.07)
D330048F12Rik	Down (1.79)	Down (1.58)
Fibin/ 1110018M03Rik	Down (1.76)	Down (1.91)
Mmp24	Down (1.53)	Down (1.72)
C130072C03Rik	Down (1.53)	Down (2.07)

3.5.4.2 Discussion

The analysis of *Pax1/Pax9* double-null embryos has clearly revealed several more genes regulated by both *Pax* genes. These genes were also enriched for relevant processes like collagen fibril organization, cartilage development, skeletal system morphogenesis, cell motion and proliferation, ECM proteins and cell adhesion. How these processes are relevant to *Pax1* and *Pax9* function in IVD development are discussed in this section.

(I) Collagen fibril organization

Collagen fibril organization refers to processes that order the collagen fibrils within the ECM and so is vital to maintain the structural integrity of the IVD [126]. Several genes involved in this process were dysregulated in the double-null embryos

(Figure 43; Tables 11 and 12). Disorganization of collagen fibres is known to impede their ability to resist the osmotic swelling pressure generated by the water retention of the proteoglycans (eg. aggrecan) in the cartilage ECM. This often results in abnormal vertebral bodies and IVDs that lead to misalignments and even degeneration [127]. Targeted knock-out or mutations in several of the genes involved in this process, such as *Acan* [128], *Col2a1* [127, 129] and *Col11a1* [130], are associated with various skeletal defects. More importantly, they were all down-regulated in the *Pax1/Pax9* double-null mutants.

Mouse mutants of *Acan*, *Col2a1* and *Col11a1* exhibit remarkably similar phenotypes such as craniofacial defects (cleft palate, short snout), shortened limbs, abnormal/ misaligned vertebral bodies which give rise to kyphosis, dwarfism and disc degeneration/ herniation at post-natal stages [127, 128, 130]. The highly similar phenotypes may be because they are all integral components of the cartilage ECM and play a collaborative structural function. Importantly, cleft palate and vertebral abnormalities are characteristic features of the *Pax1/Pax9* double-null mutants as well [62]. Moreover, certain forms of kyphoscoliosis in humans have been linked to the *PAX1* gene [84-86].

The *Acan* gene encodes for the large proteoglycan - aggrecan, which is a major component of the cartilage ECM [128]. Mutations in the human *ACAN* gene have been linked to spondyloepiphyseal dysplasia (SEMD) and osteoarthritis [131, 132].

Similarly, collagen II also forms a key component of the ECM. It is a homotrimer composed of αI (II) chains which are encoded by *Col2a1* gene. In *Col2a1*-null mice, the IVDs fail to form owing to disorganized collagen fibrils which results in a failure of notochord dismantling in the vertebral bodies as well as its expansion into the IVD regions [127]. In humans, mutations of *COL2A1* gene are

responsible for various type II collagenopathies, certain forms of SEMD and rare skeletal dysplasia like dysspondyloenchondromatosis (DSC) [129]. Besides, *Col3a1* and *Col5a1* were up-regulated in the double-null mutants. This was not surprising since ectopic expression of other types of collagens is known to occur in the mutant cartilage lacking collagen II. For instance, in the *Col2a1*-null mice, collagen I and III were found to be ectopically expressed in the cartilage. Hence, a similar mechanism might be involved in the double-null mutants, which may be forming abnormal heterotypic collagen fibrils owing to the reduction in the expression of αI (II) collagen chains [127].

Col11a1 also encodes fibrillar collagen (type XI), but unlike collagen II, it is only a minor component of the ECM. It copolymerizes with collagen type II molecules and is postulated to regulate collagen fibril diameter [130]. Similar to *ACAN*, polymorphisms in human *COL11A1* have been associated with susceptibility to lumbar disc herniation (LDH) [133-135]. It is noteworthy that the biochemical hallmarks of IVD degeneration are altered proteoglycan and collagen content of the IVD ECM. The loss of structural integrity of the IVD components (annulus fibrosus and nucleus pulposus) is responsible for the reduced load-bearing ability of the disc, which leads to progressive degeneration [29, 136, 137]. Thus, *Acan*, *Col2a1*, *Col11a1*, as well as the other genes like *Dpt*, *Col3a1* etc (Tables 10 and 11) which have essential roles in collagen fibril organization, all play important structural functions in the IVD anlagen.

Thus, these observations reiterate that *Pax1* and *Pax9* have essential roles in IVD morphogenesis. More importantly, the identification of *Pax1* and *Pax9* to be regulators of these genes in the IVD is a novel finding.

(II) Cartilage and skeletal system morphogenesis

In the *Pax1/Pax9* double-null mutants, numerous genes involved in cartilage development and skeletal system morphogenesis were enriched (Figure 43; Tables 10 and 11). Besides *Col2a1* and *Acan*, *Sox5* and *BMP4* were also differentially expressed, which are well known for their importance in cartilage development.

Sox5 (SRY-box containing gene 5) encodes a HMG-domain containing TF. *Sox5* and *Sox6* play redundant but important roles in the regulating the proliferation of chondroblasts and up-regulating the cartilage matrix genes like *Acan* [138] and *Col2a1* [139] for the timely maturation of chondroblasts. While *Sox5* and *Sox6* single-null mutants possess very mild skeletal defects, the *Sox5/Sox6* double-null mutants are characterized by chondrodysplasia owing to defect in the differentiation of the chondroblasts [140]. Moreover, their ability to regulate *Fmod*, an IVD anlagen marker, and the impairment of IVD formation in the *Sox5/Sox6* double-null mutant, suggest they have essential roles in IVD development [119]. It is also noteworthy that *Sox5* was down-regulated along with *Acan* and *Col2a1* in the *Pax1/Pax9* double-null mutants [119, 140]. This indicates that the down-regulation of *Acan* and *Col2a1* might potentially be because of *Sox5* down-regulation rather than a direct effect of the *Pax*es. The ChIP-Seq analysis of Pax1 and Pax9 would confirm if they are direct or indirect targets (discussed in section 3.6.4, Table 13).

Bone morphogenetic protein 4 (encoded by *Bmp4*) belongs to the TGF-beta superfamily of secreted signalling proteins. As the name suggests, BMPs are growth factors important in endochondral bone formation [111]. During the very early stages of patterning and compartmentalization of the somites, antagonism of BMP signal is necessary for sclerotome specification [42]. But *in vivo* rescue experiments [141] and *in vitro* experiments [142] on cultured explants showed that at later stages, BMP

signalling promotes ECM production and chondrocyte proliferation which are essential functions in cartilage development.

(III) Growth factor binding ECM proteins

Hspg2 encodes the large heparan sulfate proteoglycan, perlecan, which accumulates significantly in the ECM during cartilage development. It has important roles in cell adhesion and growth factor binding (eg. FGFs and PDGF-B), whereby the latter function modulates growth factor bioavailability [143]. Moreover, the targeted null mutants showed decreased chondrocyte proliferation, decreased glysosaminoglycan content and abnormal collagen fibril organization in the ECM, which resulted in chondrodysplasia, thus emphasizing the importance of perlecan [144, 145].

Similarly, *Cspg2* (*versican*) encodes a larger chondroitin sulfate proteoglycan of the ECM. It is important in localizing the TGF-beta molecules, thus regulating its signals for joint development. Considering that TGF-beta signals have been shown to be important in the IVD development (a form of joint in the vertebral column), and versican is also expressed in the IVD anlagen, it might be playing similar roles in tethering TGF-beta (expression database: www.eurexpress.org) [146].

Ctgf, connective tissue growth factor, encodes a secretory protein that is associated with the ECM. It is a well known for its multiple properties and functions like mitogen, cell adhesion, migration, ECM remodelling, ECM production and chondrogenic differentiation [147]. It mediates TGF-beta and BMP signals by directly interacting with them [148]. Upon induction by TGF-beta it promotes mesenchymal cell condensation and ECM production [149, 150]. It also binds BMP7 and prevents its inhibition of TGF-beta signalling [151]. Also, in the *Ctgf*$^{-/-}$ mice, endochondral

ossification is delayed and they show cleft palate and enlarged vertebral bodies in the axial skeleton [152].

Thus, *Pax1* and *Pax9* activate genes which encode ECM proteins that are essential components of cell signalling as they interact with the growth factors to stimulate the cells to grow or proliferate (Figure 43; Tables 10 and 11).

(IV) Cell motion

Genes involved in cell motion were also enriched in the *Pax1/Pax9* double-null mutants (Figure 43). During the early stages of paraxial mesoderm formation, *Pax1* expressing cells from the ventro-medial somites are known to migrate toward the notochord and surround it [153]. Also, *Pax9* is expressed in neural crest-derived tissue, whereby neural crest cells are known to have migratory properties [74]. Moreover, in the double-null mutants, the fluorescing cells were found to be mis-localized in the lateral regions and failed to surround the notochord medially (Figure 31). Hence, cell motion is likely an essential property of these *Pax1*- and *Pax9*-expressing cells. Although these genes are often associated with axon guidance, they also show regionally restricted expression in the somites at E9.5 and/or IVD at E14.5 (eg. *Alcam, Sema6a, App, Psen1, Ctgf, Etv1, Reln, Nr2f2, Boc and Foxd1*; ref - Expression databases: eurexpress.org, MGI and EMAGE). Of these, *Psen1* and *Ctgf* are known to have skeleton defects in the targeted null mutants, with *Psen1* mutants exhibiting a phenotype strikingly similar to *Pax1/Pax9* double-null mutants (Tables 10 and 11).

Psen1 encodes a transmembrane protein, presenilin 1, which is involved in regulating cell motion and cell proliferation, mainly identified through neuronal studies [154]. While it has been strongly related to the Alzheimer's disease, the targeted null mutants of *Psen1* also show a dramatic axial skeleton defect characterized by fused

vertebrae throughout the entire vertebral column, lack of sacral elements and floating ribs. The mice exhibit very short, curled tails [154]. Some of these skeletal defects are similar to the *Pax1/Pax9* double-null mutants which also exhibit a strong skeletal phenotype with absent sacral elements and the proximal parts of ribs [62]. Also, somitogenesis and sclerotome formation are unaffected in both *Pax1/Pax9* double-null mutants and the *Psen1* mutants [62, 154]. The lack of sufficient studies on the skeleton phenotype in the *Psen1*$^{-/-}$ mutants hinders the identification of the actual roles of *Psen1* in axial skeletogenesis. Nonetheless, their migratory, proliferation and cell-adhesion roles in the neurons can probably be extrapolated to the skeletal cells as well.

(V) Cell proliferation

Likewise, genes involved in proliferation were also differentially expressed (*Bmp4, Psen1, Col8a1* and *Col8a2*). While the GO term is referred to as "epithelial cell proliferation", these genes are also expressed in the IVD anlagen, hence probably have similar functions in IVD development (Expression databases: eurexpress.org, MGI and EMAGE) (Figure 43 and Table 11). Besides these, others like *Ctgf, Sox5* and *Hspg2* which were discussed above, are also known to promote proliferation.

(VI) *Pax1* and *Pax9* expression levels

Intriguingly, *Pax1* was down-regulated in the *Pax1/Pax9* double-null mutants. The decrease was evident with multiple copies of *Pax9* being lost in the *Pax1*$^{-/-}$ background (i.e. double-null vs *Pax1*$^{-/-}$: 1.50-fold down; and double-null vs WT: 2.35-fold down). It is not clear if *Pax1* or *Pax9* or both are regulating it directly, or if the

down-regulation is an indirect consequence. ChIP-seq experiments on Pax1 and Pax9 TFs may illuminate the regulation of *Pax1* (section 3.6.4).

It is notable that with the strategy employed in this study to knock-out *Pax1* or *Pax9*, the full length transcript would be produced. It is during translation the protein product is not made. Hence, the decrease seen in *Pax1* transcript level is not because of truncation of the transcript, but a true down-regulation.

Peters *et al* (1999) reported that the *Pax9* domain of expression was spatially expanded and up-regulated in the anterior segment of the sclerotome in the $Pax1^{-/-}$ $Pax9^{+/-}$ and $Pax1^{-/-}Pax9^{-/-}$ embryos at E10.5 [62]. This was proposed by the authors as a potential mechanism by which *Pax9* partially rescues *Pax1*-deficiency. Contrary to what has been reported before, in this gene expression profiling study, *Pax9* was not differentially expressed in any of the knock-out mutants (i.e. 4 allele KO, 3 allele KO or 2 allele KOs). The Pax9 protein expression resembled that of the WT in both heterozygote and homozygote *Pax1* mutants (Figure 25). In the study by Peters *et al*, 1999, the authors had knocked-out *Pax9* by the insertion of a *lacZ* cassette in the exon of *Pax9* [107]. Hence, the mode by which the authors had investigated *Pax9* expression levels was through X-gal staining, which is not an accurate method of assessment of the actual Pax9 protein levels. Prolonged X-gal staining is known to give rise to spurious background staining [107]. With our current method of assessment using immunohistochemistry, neither a change in Pax9 expression level nor a spatial expansion was observed in the E13.5 $Pax1^{+/-}$ and $Pax1^{-/-}$ embryos.

Then again, it is possible that such spatial expansion might occur only at E10.5, the stage which the authors had investigated. The *Pax9* expressing cells may have moved into *Pax1* regions to compensate. *Pax1* expression is found in all the sclerotomal cells initially, and later it becomes restricted to posterior ventro-medial regions. *Pax9* is expressed in the posterior regions [62]. Hence, in the absence of

Pax1, *Pax9*-expressing cells may have occupied the anterior regions to compensate, and once the patterning was complete, they may have been subjected to the same regulatory controls from the neighbouring tissues as *Pax1*-cells normally would have and become restricted to the posterior regions. Hence, such a spatial expansion may have been missed by our investigation in E13.5 embryos. Therefore, while the increase in transcript level does not explain the compensation, spatial expansion theory still remains a possibility. Investigation of the $Pax1^{-/-}$ and $Pax1^{-/-}Pax9^{+/-}$ embryos at E10.5 might help clarify this doubt.

(VII) Genes with opposite directionality – a consequence of compensation?

The set of genes exhibiting opposite directionality in regulation were particularly interesting since they were up-regulated only at E13.5 in the $Pax1^{-/-}$ (unchanged in E12.5 $Pax1^{-/-}$), but down-regulated once both copies of *Pax9* were lost. *Sox5, Col2a1, Wwp2,* and *Ctgf* were among those set of genes (Table 12). The basis for this trend might be because *Pax9* was compensating for the loss of *Pax1*. Also, as mentioned earlier (section 3.5.3), *Sox5, Col2a1, Wwp2* and *Ctgf* are known to be down-stream targets of *Sox9* [125]. Hence, *Sox9*, which was concomitantly up-regulated in E13.5 $Pax1^{-/-}$, may be cooperating with *Pax9* in activating these targets in the absence of *Pax1*, resulting in an up-regulation of these targets beyond normal levels. The time delay might be because sufficient time was needed for the transcripts to be up-regulated to significant levels.

Then again, one might argue that a similar trend would be observed if *Pax1* was repressing while *Pax9* was activating these genes. However, if *Pax1* and *Pax9* were functioning antagonistically, one would expect the transcript levels of these genes to return to normal once both the activator and repressor were removed. On the contrary, a significant down-regulation is observed. Moreover, *Sox9* was not

differentially expressed in the E12.5 *Pax1/Pax9* double-null mutants. This precludes *Sox9* from being responsible for the observed down-regulation. Furthermore, *Pax1* and *Pax9* are paralogous genes and are known to function synergistically [62]. Thus, it is highly possible that *Pax1* and *Pax9* were not functioning antagonistically. Only through further experimentation can we determine which hypothesis holds true.

One way to assess the validity of these postulations is to similarly investigate the differential gene expression pattern in the *Pax9*$^{-/-}$ and *Pax1*$^{+/-}$*Pax9*$^{-/-}$ mutants (with respect to WT). Taking *Col2a1* as an example, if a similar spike in expression level is seen in the *Pax9*$^{-/-}$ and a subsequent down-regulation in the *Pax1*$^{+/-}$*Pax9*$^{-/-}$ mutants, the second hypothesis would be valid, i.e. both *Pax1* and *Pax9* are activators. Performing an *in vitro* assay is complicated because there are no sclerotomal cell lines available. Chondrogenic cell lines may be used instead, like differentiated ATDC5 cells or C3H10T1/2 (multipotent murine mesenchymal stem cell-line), but the differentiation process must be optimized since *Pax1* and *Pax9* are known to be down-regulated upon chondrogenesis and are not expressed in the fully differentiated mature cartilage of the vertebral bodies [62, 71]. Assuming that the differentiation conditions are optimized, the cells can then be transfected with over-expression vectors of *Pax1* and/or *Pax9*. The endogenous *Col2a1* transcript levels can be assessed at different concentrations of the *Pax1* and/or *Pax9*. If *Pax1* was a repressor and Pax9 an activator, *Col2a1* level should go up with decreasing concentrations of *Pax1* or increasing concentrations of *Pax9*. One caveat of this method is that it is not clear if endogenous *Sox9* might complicate the results. However, considering that it is an over-expression assay, the effects of endogenous *Sox9* may be minimal.

In conclusion, all the target genes discussed here exhibit phenotypes or defects of the axial skeleton some of which are strikingly similar to that of the

Pax1/Pax9 double-null mutants. It has to be noted though that the expression of these genes are not totally abrogated in the double-null mutants. Therefore, no one gene could be responsible for the phenotype observed in the *Pax1/Pax9* double-null mutants. Instead, they prove the point that they are all linked in a common pathway and it must be a collective effect of the dysregulation of these genes which results in the skeletal defect seen in the *Pax1* and *Pax9* mutants. Moreover, the analysis of the single-null mutant and double-null mutant revealed genes with opposite directionality in regulation, indicating potential compensation effects. It is not clear if the compensation is caused by *Pax9* or in conjunction with other genes like *Sox9*.

3.6 Genome-wide binding site mapping of Pax1 and Pax9

Genome-wide mapping of TF binding site is essential to complement the differential gene expression data derived for *Pax1* and *Pax9*. The differential gene expression analyses reveal the genes regulated directly by *Pax1/ Pax9* as well as those that are affected indirectly owing to the altered cascade of events triggered by the loss of *Pax1/Pax9*. Hence, the direct binding targets of Pax1 and Pax9 in the WT tissue identified through TF mapping and overlapping this data with their corresponding differential gene expression data, would reveal which genes are directly regulated (differentially expressed as well as have Pax1/Pax9 binding sites) and indirectly regulated (differentially expressed but lack corresponding Pax1/Pax9 binding site). This would add another dimension to comprehending *Pax1/Pax9* transcriptional regulation of these genes. Chromatin immunoprecipitation coupled with high-throughput sequencing allows the identification of binding sites of TFs on a genome scale, and with a high resolution.

While the HA-tagged mouse lines were being generated for the ChIP-Seq studies, in the meantime, TF mapping was performed using gene-specific antibodies on WT embryos. At the time this study was performed, there were only two

commercially available ChIP-grade antibodies each for Pax1 and Pax9. Even though *Pax1* and *Pax9* belonged to the same subfamily their proteins were highly divergent at their C-terminal ends. Hence, to avoid cross-reactivity, these commercial antibodies were raised against the C-terminal fragments of the proteins (amino acids 246-361 of mouse Pax1 and amino acids 247-341 of human PAX9). To ensure that these antibodies indeed did not cross-react, their specificity was first checked by immunohistochemistry using the ChIP-grade antibodies (Figure 46).

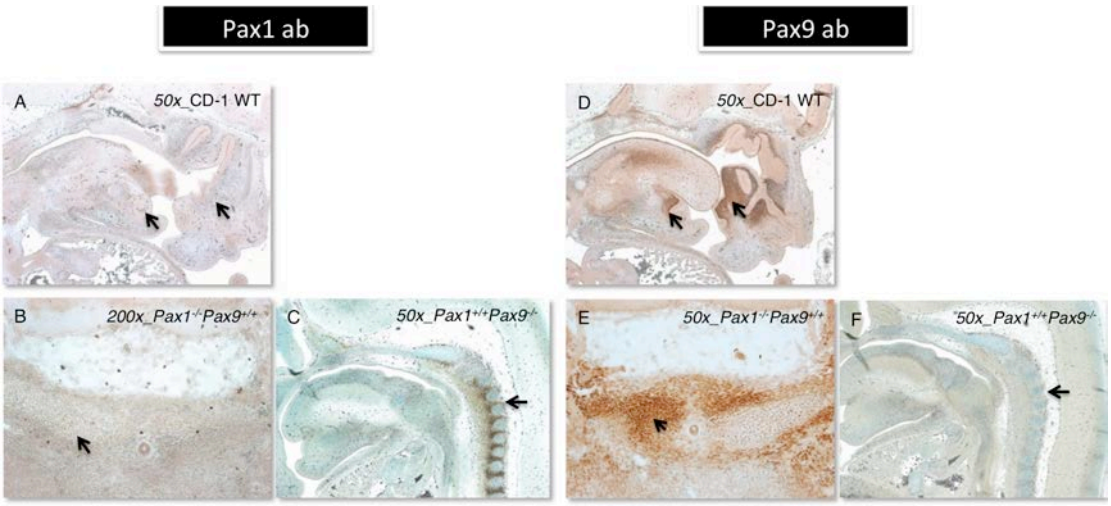


Figure 46: Pax1 and Pax9 ChIP-grade commercial antibodies were specific. (A & D) E13.5 WT embryos probed with anti-Pax1 and anti-Pax9 antibodies respectively. Only anti-Pax9 antibody detected expression in the tooth and nasal mesenchyme, which are *Pax9*-specific regions. (B & E) Transverse sections of E13.5 *Pax1*-null embryos probed with anti-Pax1 and anti-Pax9 antibodies respectively. Anti-Pax1 antibody did not detect the Pax9 protein (B) which was still expressed in the *Pax1*-null embryos (E). (C & F) E13.5 *Pax9*-null embryos probed with anti-Pax1 and anti-Pax9 antibodies respectively. Anti-Pax9 antibody did not detect Pax1 protein (F) which was still expressed in the *Pax9*-null embryos (C). Paraffin sections of 10 uM thickness were used for immunohistochemistry. Anti-Pax1 ab – Santa Cruz M116X; anti-Pax9 ab - Santa Cruz H95X. ab – antibody.

The ChIP-grade antibodies Santa Cruz Pax1 M-116X (sc-25407x) and Santa Cruz Pax9 H-95X (sc-25410x) were tested on E13.5 mouse embryo sections by immunohistochemistry for sensitivity and specificity. Anti-Pax1 antibody was sensitive enough to detect endogenous Pax1 protein (Figure 46C). It also did not

cross-react with Pax9 evident from the lack of signal in the tooth and nasal mesenchyme which are *Pax9*-specific regions (Figure 45A). In addition, the anti-Pax1 antibody did not show any signal in the *Pax1*-null embryo, despite the presence of endogenous Pax9 protein (Figure 46B). Likewise, anti-Pax9 antibody was able to detect the Pax9 protein specifically in the tooth and nasal mesenchyme in the WT embryo (Figure 46D), as well as the vertebral column in the *Pax1*-null embryo (Figure 46E). It did not detect any signal in the *Pax9*-null embryos which still expressed Pax1 protein (Figure 46F). Thus, both antibodies proved to be sensitive and specific in detecting the right proteins. Therefore, I proceeded to perform the ChIP on the CD1 WT, embryonic vertebral column-enriched tissues using these antibodies.

For immunoprecipitation, only the dissected vertebral column tissues (discarded were head, limbs and internal organs) were used. After immunoprecipitation according to the procedure mentioned in the Materials & Methods section, the library was prepared from the ChIP-DNA (Illumina ChIP-Seq DNA Prep kit) and was subsequently size-selected (200-300bp), purified, quantitated by Agilent Bioanalyzer (DNA chip) and sent for sequencing by Illumina's Solexa sequencing platform (Figure 47). The input chromatin was used to prepare the control library, which was used as the background control. Any region bound by the Pax1 or Pax9 proteins solely due to the open chromatin state (i.e. background) can be identified using this input library and accordingly subtracted from the signals detected in the Pax1- and Pax9-libraries. The ChIP-Seq was performed on E13.5 and E12.5 embryonic tissues for Pax1 and Pax9 respectively. While E12.5 Pax1 ChIP-Seq and E13.5 Pax9 ChIP-Seq were meant to be performed subsequently, owing to unforeseen fiscal constraints, those experiments could not be performed.

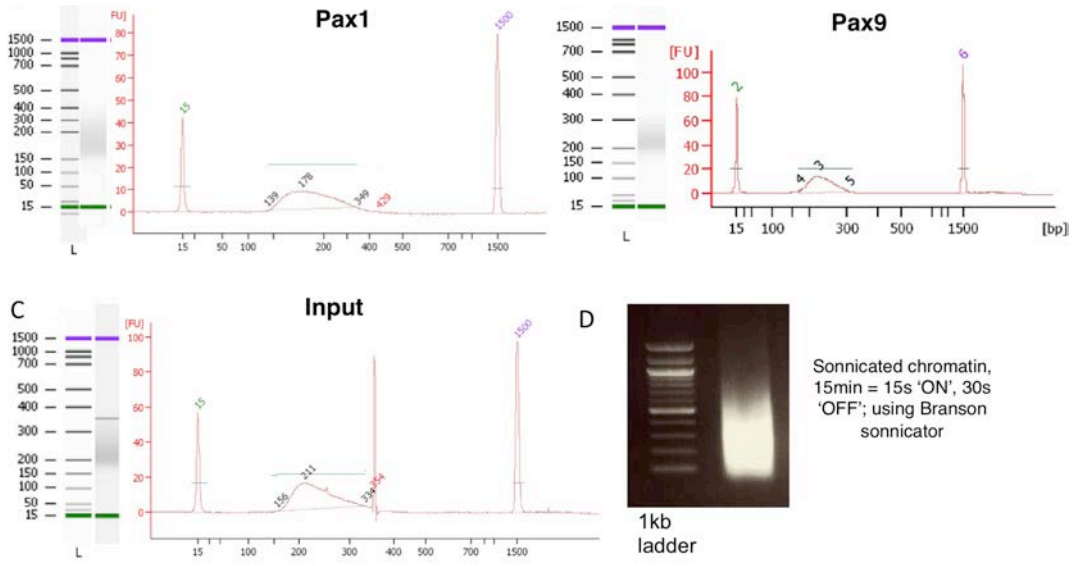


Figure 47: ChIP-Seq libraries for sequencing. (A) Pax1 ChIP-Seq library analyzed by Agilent Bioanalyzer; (B) Pax9 ChIP-Seq library; (C) input library prepared from sonicated chromatin. (D) Representative picture of sonicated chromatin resolved on 1% agarose gel.

3.6.1 Binding site distribution of Pax1 and Pax9

After high-throughput sequencing of the short tags of DNA in the libraries, the sequence reads were mapped to the mouse genome (NCBI build 37/mm9) to identify the genomic binding sites of Pax1 and Pax9. The peaks were called using Model-based Analysis of ChIP-Seq (MACS) algorithm [155]. Binding sites can be considered biologically relevant only if the tag density reads (peaks) in those regions are significantly enriched above the background levels (input control). In this study, a stringent fold-enrichment p-value cut off of a minimum 1.00e-5 was applied to identify significant peaks.

For Pax1 and Pax9 libraries, after background subtraction and statistical analysis, a total of 10,203 peaks (p-value ≤ 1.00e-5) and 11,333 peaks (p-value ≤ 1.00e-10) were identified for Pax1 and Pax9 respectively. The binding site distribution for the number of peaks within the TSS, promoter, intragenic, proximal,

distal and beyond 100 kb of a gene was assessed for Pax1 and Pax9. The percentage of binding site distribution is shown as a pie-chart in Figure 48, including the criteria set for defining the TSS/ promoter/ intragenic/ proximal/ distal/ others regions.

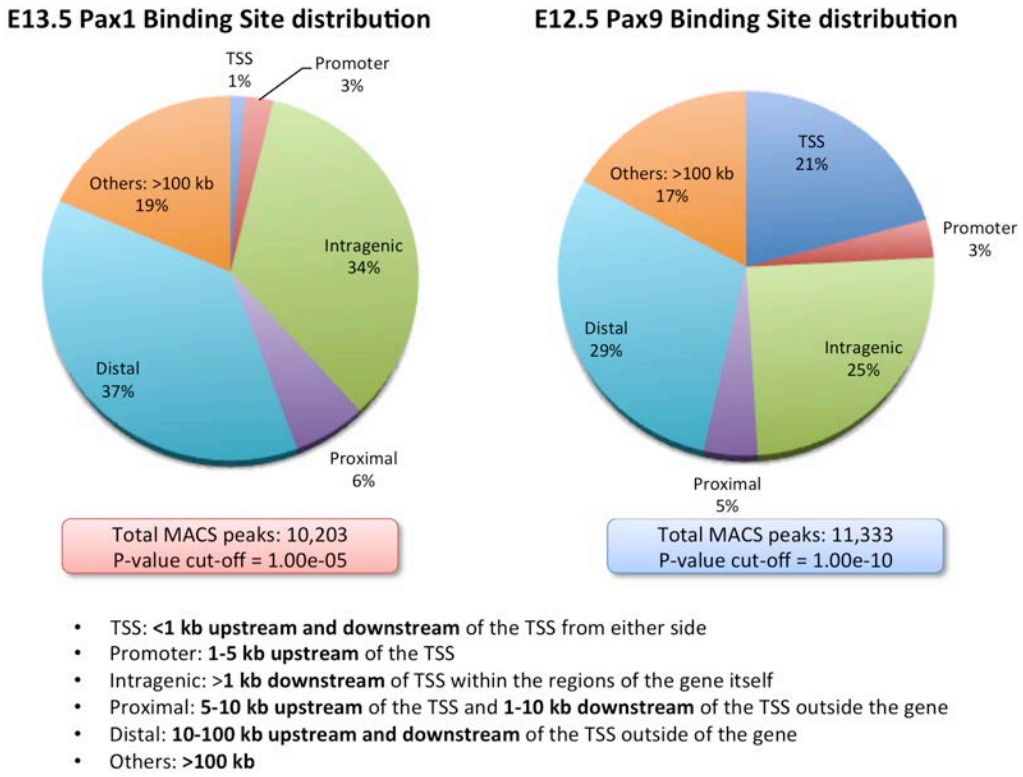


Figure 48: Binding site distribution for Pax1 and Pax9. Left: Pax1 had total MACS peaks of 10,203, whereby 37% fell in distal, 34% in intragenic, 19% at > 100kb, 6% in proximal, 3% in promoter and 1% at TSS. Right: Pax9 had a total MACS peaks of 11,333, where 29% fell in distal, 25% in intragenic, 21% in TSS, 17% at >100kb, 5% in proximal and 3% in promoter. The p-value cut off was 1.00e-5 and 1.00e-10 for Pax1 and Pax9 respectively.

From the binding site distribution, it was evident that majority of the binding sites fell in a distal region, followed by intragenic sites for both Pax1 and Pax9. This indicates that Pax1 and Pax9 were regulating most of their target genes from distal or intragenic sites. In fact, gene regulation from distal cis-regulatory elements of > 50 kb, up to even few hundred kilobases, are not uncommon. For example expression of *Shh* in the Zone of Polarizing Activity (ZPA) in the limbs is driven by an enhancer element that resides 1 MB away from *Shh* locus, in the intron of a neighbouring gene,

Lmbr1 [156]. Likewise, enhancers that drive *Sox9* expression in the mandible, second branchial arch, paraxial mesoderm of ribs and cervical vertebrae all lie in a 350 kb - 600 kb segment upstream of *Sox9* locus [157]. Furthermore, multiple independent enhancers may be involved in regulating the expression of a particular gene in a specific tissue and time [158]. Chromatin looping brings the TFs bound to distal enhancer regions close to the promoter where it interacts with the rest of the transcriptional machinery poised for transcription. These distal cis-regulatory elements could be enhancer or repressor elements, that control spatio-temporal and even the level of expression of a particular gene [159].

In general, both Pax1 and Pax9 showed a similar distribution within proximal (5-6%) and promoter (3%) regions. Remarkably, a striking difference was seen in the proportion of peaks within the TSS, whereby it was 21% for Pax9 but only 1% for Pax1. Clearly, at E12.5-E13.5, Pax1 was regulating most of its genes from distal or intragenic sites whereas Pax9 was regulating a significant proportion from the TSS as well.

3.6.2 Motif discovery in Pax1 and Pax9 binding sites

Pax proteins are known to execute their transcriptional regulation function by binding to the DNA via their paired domain and/or homeodomain. Pax1 and Pax9 belong to subfamily 1, both of which possess only the paired domain and an octapeptide motif but not the homeodomain. Hence, their entire binding specificity is largely reliant upon the paired box. The Pax proteins are also known to bind the DNA as monomers [68]. The paired domain is bipartite in nature, with the N-terminal and C-terminal sub domains, each possessing a helix-turn-helix motif (HTH) and DNA-binding ability [67, 160]. However, the two subdomains are interdependent and cannot bind the DNA autonomously [160]. The two subdomains of the paired box

recognize two half-sites positioned on adjacent major grooves of the DNA. The 5' half site and 3' half sites were identified using Pax5. The 3' half site is recognized by the N-terminal subdomain while the 5' half site is recognized by the C-terminal subdomain. Moreover, the Pax1 proteins bound to Pax5 recognition sequences, albeit with lower affinity than Pax5. Binding to both half sites are needed to confer greater affinity and stability in DNA-binding *in vitro*. Yet base changes in one can be compensated by a perfect match on the other. It is believed that such modular properties of the HTH motifs may generate diversity in the binding specificities of the Pax proteins [67, 160].

The paired domain of Pax1 is known to recognize a 24 bp sequence. Chalepakis *et al* 1991 [68], through *in vitro* assays, had identified two pentanucleotide core motifs instrumental in the DNA-binding affinity of Pax1: GTTCC and TAGAT. The known motifs for Pax1 and Pax9 (verified *in vitro* / predicted) are shown in Figure 49.

Predicted motifs/ motifs in database

		Best possible match	
Motif	**Width**	**(+)**	**(-)**
V_PAX1_B	18	CCGTTCCGCTCTAGATAT	ATATCTAGAGCGGAACGG
V_PAX9_B	24	GAGACGCAGCGAGGAGTGACCACC	GGTGGTCACTCCTCGCTGCGTCTC

Figure 49. Pax1 and Pax9 motifs found in motif databases. Motifs for Pax1 and Pax9 found in Transfac, Matrix and MEME are shown. For Pax1, both GTTCC and TAGAT pentanucleotide core motifs mentioned in the literature are underlined. For Pax9, the 3' half site of the paired domain (GTGACC), also identified in our Pax9 TF mapping, is underlined.

From the binding sites obtained for Pax1 and Pax9, binding sequence motif was identified using Weeder, a motif-finding program (Figure 50). The motif identified to be enriched in the Pax1 library did not show the Pax1 motif found in the TRANSFAC database. Instead, it showed a motif for Zscan4c, which did not contain

either of the pentanucleotide core motif sequences. However, it resembles more of the “GTG” based 3’ half site of Pax5 subfamily proteins. It However, when probed for motifs enriched up to 250 bp surrounding the binding sites using Centrimo, Pax5 motif was found. It is noteworthy that Centrimo uses JASPAR motif database which does not contain the motifs for Pax1 or Pax9. Hence, any Pax motif the program finds, it would not be able to label it as “Pax1” or “Pax9” specifically.

For Pax9, the 3’ half site identified in the Pax5 subfamily (also present in the predicted motif for Pax9), was found to be enriched (Figures 49 and 50). Just like Pax1, when the 250 bp surrounding the Pax9 binding sites were checked for enriched motifs, Pax5 motif, as well as Pax2 and Pax6 motifs were found.

A

Library	Motif detected and possible TFs	Centrimo motif enrichment +/- 250bp near the binding sites
Pax1	Gm397 = Zscan4c 	Pax5, PHD1, SOK2,Gm397, etc
Pax9	Pax2 	Pax2, Pax5, Pax6, Hic1, Max, HIF1A, etc.

B

Consensus sequence for **Pax5 subfamily** (Pax2/5/8) **3’:**

- Positions 16 to 24: GA GG / GCGTG ACCA

Figure 50: Motif discovery results for Pax1 and Pax9. Weeder motif-discovery program was used to assess if Pax motifs were enriched in the Pax1 and Pax9 libraries. (A) Table of Pax libraries and the corresponding most significant motif, and motifs enriched within 250 bp of the binding sites. Pax1 library was enriched for Zscan4c motif, while Pax9 showed a Pax2 paired box 3’ half-site motif (boxed in red). (B) Known consensus 3’ half-site motif (boxed in red) for paired domain in the Pax5 subfamily that also encompasses Pax2 and Pax8.

Even though the expected MGI expression, mouse phenotype and biological processes were enriched in the Pax1 ChIP-seq library (discussed in section 3.6.3), it was surprising that the known consensus motif was not enriched (Figures 51 and 52 in the next section). It is possible that such *in vitro* assay-derived motifs may not be

the only sequences Pax1 truly bound to *in vivo*. In fact, the known Pax1 motif was derived solely based on various mutant oligonucleotide sequences derived from a single target sequence - the e5 site from *Drosophila even skipped* promoter. Even *in vitro*, Pax1 protein bound to the native e5 sequence only weakly [68]. Hence, for Pax1, there is a possibility that it may have alternative motifs in an actual *in vivo* scenario.

Besides, even though only a half site, the expected motif was enriched in the Pax9 ChIP-Seq library which provides confidence in the dataset.

3.6.3 Gene Ontology analysis of Pax1 and Pax9 binding sites

Gene Ontology (GO) enrichment analysis is often performed on the binding sites to assess the quality of the ChIP-Seq and also to obtain a broader understanding of the functions of the genes involved. Hence, the set of GO terms enriched (in the Pax1 and Pax9 binding sites) for MGI expression, mouse phenotype, biological processes and signalling pathways were assessed using the Genomic Regions Enrichment of Annotations Tool **(**GREAT) web-based tool. In brief, the BED file of chromosome regions with peaks (identified by MACS) for Pax1 or Pax9 is uploaded to GREAT tool. For each gene GREAT assigns a regulatory domain based on user-defined criteria. The genomic regions (binding sites) are then associated with the genes whose regulatory domains overlap with that of the binding sites. The criteria assigned to define the regulatory domain for this analysis was 5.0 kb upstream and 1.0 kb downstream of the TSS and a 1000 kb (1MB) extension in both directions to the next closest gene's TSS but a maximum extension in only one direction (Figure 51). This is also the recommended settings by GREAT [161].

Figure 51. Genomic regulatory domain assignment criteria in GREAT. Arrows represent the TSS of a gene. Regulatory domain for each gene was defined as 5.0 kb upstream and 1.0 kb downstream of the TSS and a 1 MB extension in both directions to the next closest gene's TSS but a maximum extension in only one direction. Picture adapted from GREAT (http://great.stanford.edu/public/html/input.php).

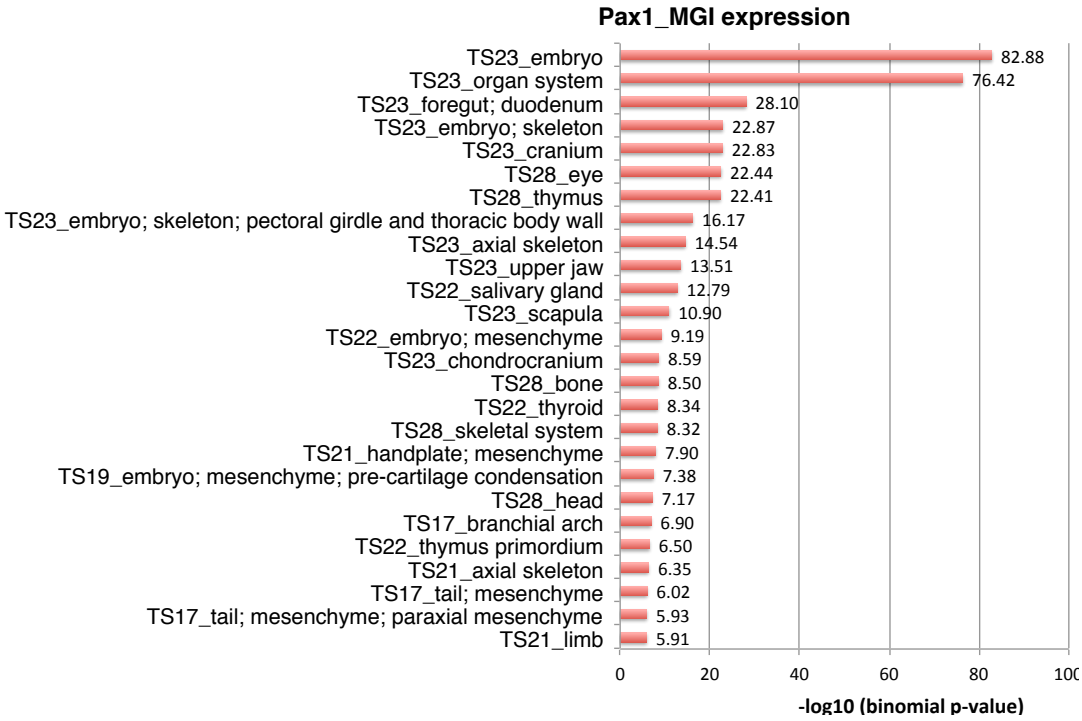


Figure 52: GO enrichment for MGI expression pattern in Pax1 TF mapping. Binding sites were enriched for genes expressed in known Pax1 expression sites like embryo skeleton, foregut, cranium, scapula, thymus primordium, tail and paraxial mesenchyme and limb. The -log10 (binomial p value) are shown at the end of the bars; a value of more than 1.3 is deemed significant (-log10 transformation of a p value of 0.05 is equal to 1.3).

In the GREAT analysis expected *Pax1*-expression sites were significantly enriched (log transformed p-value $\geq$ 1.3) in the Pax1 ChIP-Seq library. Genes expressed in the embryo skeleton, foregut, thymus, pectoral girdle, axial skeleton, scapula, pre-cartilage mesenchyme, branchial arch, limbs and tail and paraxial mesenchyme among others were associated with the binding sites (Figure 52; Table

2 for *Pax1* expression sites). Although the ChIP was performed only on dissected vertebral column tissues, expression sites like limbs, thymus, scapula were also enriched. This is possibly because the set of genes involved in axial skeleton are also involved in the development of limbs, thymus etc. owing to pleiotrophic function.

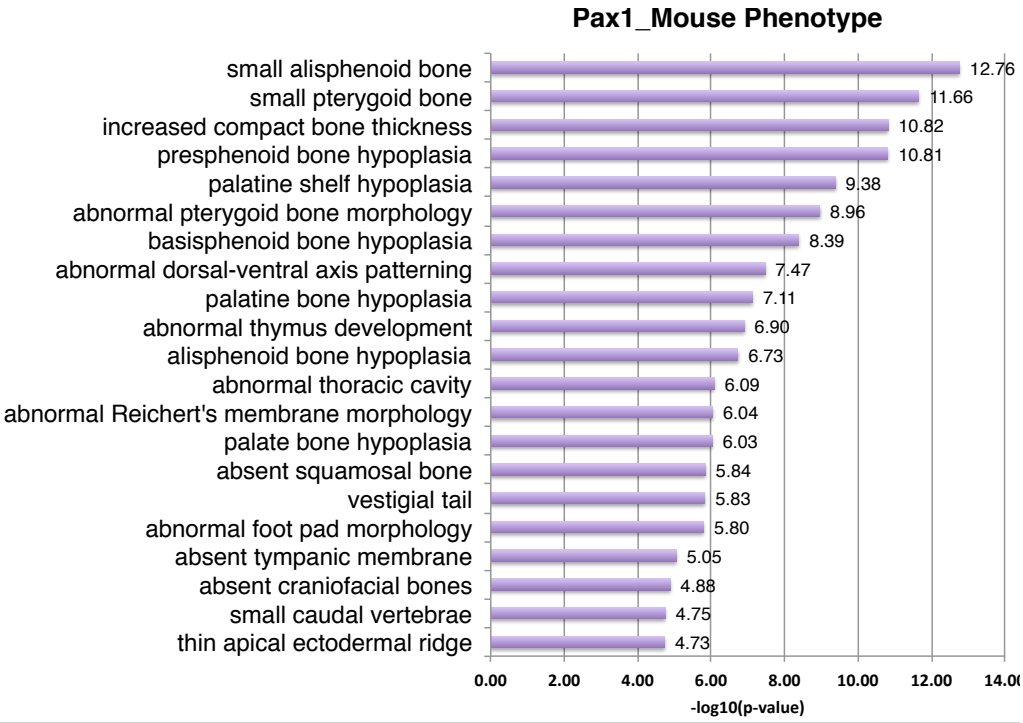


Figure 53: GO enrichment for Mouse phenotype for Pax1 binding sites. Pax1 binding sites were enriched for genes associated with mouse phenotypes like abnormal bone, dorsal-ventral axis patterning, thymus development, caudal vertebrae etc. which are all relevant to *Pax1* function. The -log10 (binomial p value) are shown at the end of the bars; a value of more than 1.3 is deemed significant (-log10 transformation of a p value of 0.05 is equal to 1.3).

As expected, the Pax1 binding sites were also associated with genes involved in numerous mouse phenotypes relevant to *Pax1* function. The genes associated with small/ abnormal bone morphology, abnormal thymus development, abnormal thoracic cavity and small caudal vertebrae were enriched, all of which are defects observed in *Pax1*$^{-/-}$ mouse mutants (Figure 53) [73].

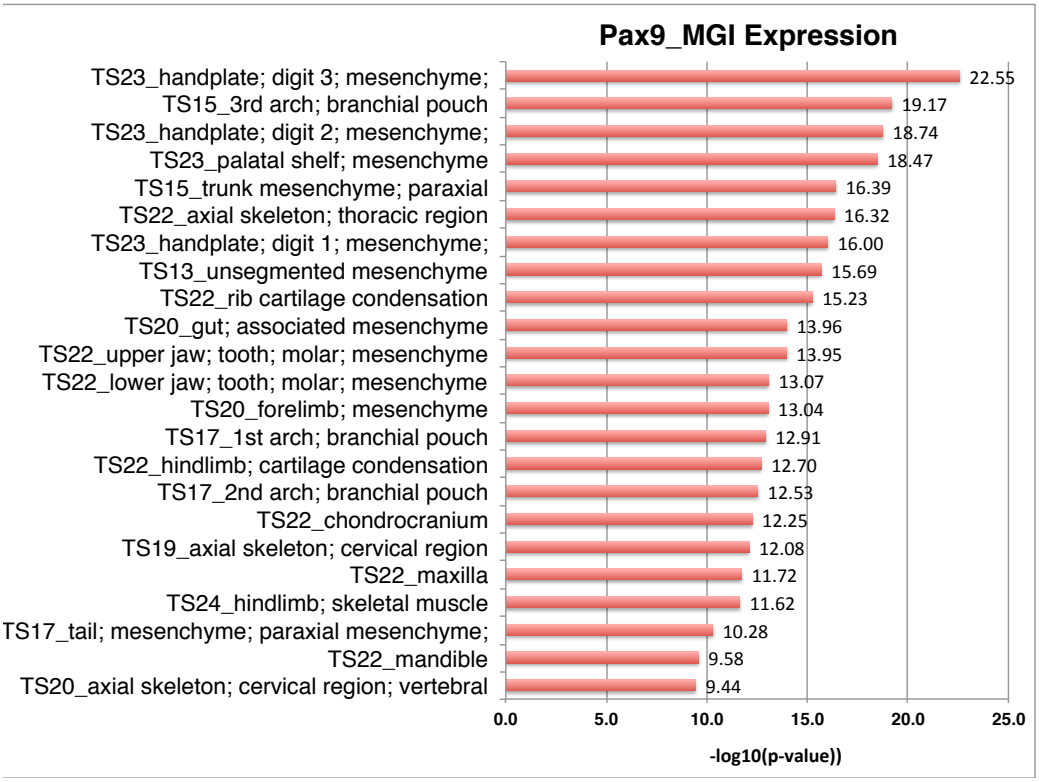


Figure 54: GO enrichment for MGI expression pattern in Pax9 TF mapping. Binding sites were enriched for genes expressed in known Pax9 expression sites like branchial pouch, forelimb mesenchyme, paraxial and tail mesenchyme, cartilage condensations, tooth mesenchyme, mandible and maxilla, chondrocranium and axial skeleton. The -log10 (binomial p value) are shown at the end of the bars; a value of more than 1.3 is deemed significant (-log10 transformation of a p value of 0.05 is equal to 1.3).

Similar to Pax1, Pax9 ChIP-Seq binding sites were also assessed by GREAT to assess if relevant genes were associated with the binding sites. As expected, the Pax9 binding sites were associated with genes expressed in the *Pax9*-specific regions like forelimb mesenchyme, branchial pouch, paraxial and tail mesenchyme, maxilla and mandible, vertebral cartilage condensations, chondrocranium and tooth mesenchyme (Figure 54; Table 2 for *Pax9* expression sites). Moreover, the genes were associated with known *Pax9*-null mouse mutant phenotypes like abnormality of dentin, pterygoid process, thyroid cartilage, Reichert's cartilage morphology; absent ultimobranchial arch, abnormal development of thymus, palate shelf elevation and molar growth retardation; and polysyndactyly (Figure 55) [74, 75].

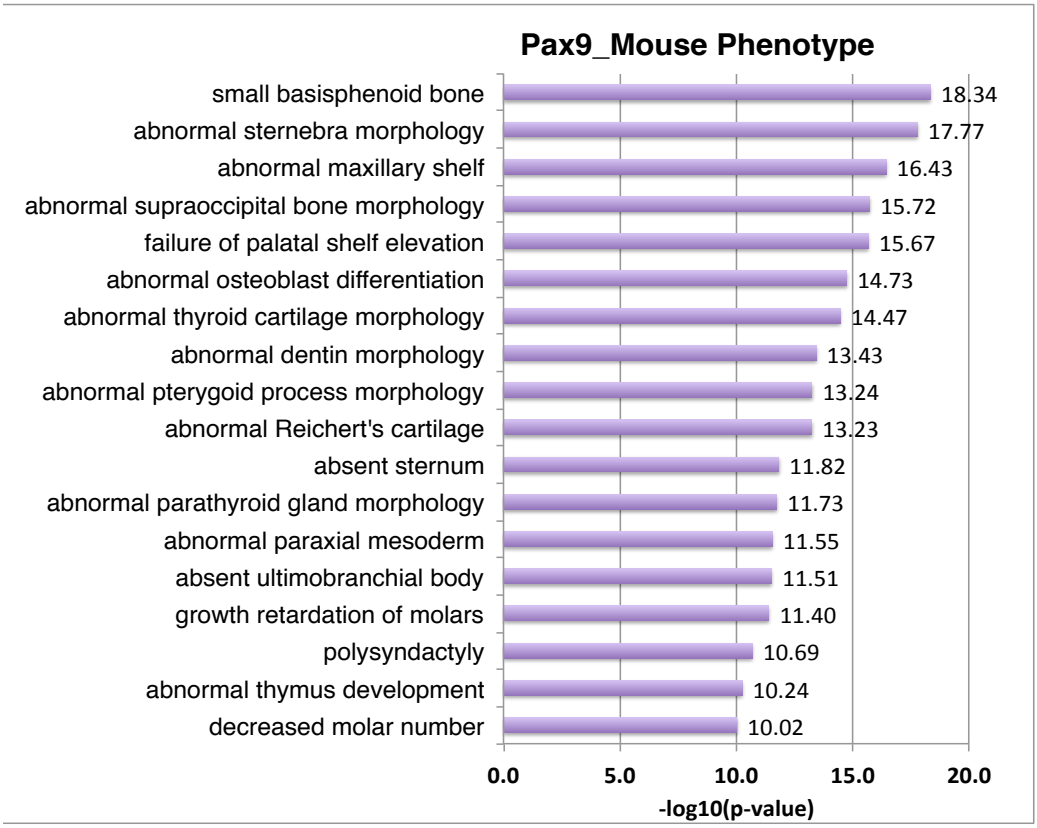


Figure 55: GO enrichment for Mouse phenotype for Pax9 binding sites. Pax9 binding sites were enriched for genes associated with mouse phenotypes like abnormal craniofacial bones (Reichert's & thyroid cartilage), dentin morphology, thymus development, absent ultimobranchial body, polysyndactyly etc, which are all hallmarks of *Pax9*$^{-/-}$ mutant mice. The -log10 (binomial p value) are shown at the end of the bars; a value of more than 1.3 is deemed significant (-log10 transformation of a p value of 0.05 is equal to 1.3).

In conclusion, both Pax1 and Pax9 binding sites were associated with the relevant genes based on the MGI expression and mouse phenotype GO analysis, thus providing greater confidence in the dataset for further analyses.

3.6.4 Pax1 and Pax9 direct targets

The Pax1 and Pax9 binding sites were further analyzed to understand their biological significance. A total of 8,263 genes (10,203 peaks) and 8,221 (11,333 peaks) genes were associated with Pax1 and Pax9 binding sites respectively based on GREAT analysis.

The Pax1 TF mapping data was overlapped with the differentially expressed genes from E12.5 and E13.5 *Pax1*$^{-/-}$ and double-null mutants. This would reveal the direct targets of Pax1 and Pax9. Of these, 50 genes (38.5% of E12.5 gene expression profiling targets) and 62 genes (50.8% of E13.5 gene expression profiling targets) from E12.5 and E13.5 stages respectively, were direct targets of *Pax1* (Figure 56). Notably, *Sox5, Col2a1, Acan and Ctgf* were some of the targets with direct binding site for Pax1 (Table 13 for entire list). Gene ontology of the direct targets of Pax1 at E12.5 an E13.5 showed enrichment of the same set of processes that were observed earlier (Figures 40 and 41) with the list of differentially expressed genes: transcriptional regulation, cell adhesion, cell proliferation, apoptosis, cartilage and skeletal system development, ECM proteins, cell migration and growth factor binding (Figure 56).

When overlapped with the double-null differentially expressed genes, a total of 27.2% (162 genes) of the differentially expressed genes in the double-null overlapped with both Pax1 and Pax9 ChIP-Seq TF mapping targets, 37.3% (222 genes) overlapped with Pax1 ChIP-Seq TF mapping targets and 39.8% (237 genes) with Pax9 ChIP-Seq TF mapping targets. The remaining ChIP-Seq TF mapping targets which did not overlap with the differentially expressed gene lists could represent TF binding regions where Pax1 or Pax9 were poised, but not functional. They could be waiting for the appropriate partners/ co-factors which are probably available only at specific time-points or tissues. The entire 296 genes that overlapped with Pax1 and Pax9 ChIP-Seq were therefore the direct targets of Pax1 and/ or Pax9 and were further analyzed for GO term enrichment. Besides the terms that were seen enriched in the direct targets of Pax1 (single-null, Figure 56), new terms like TGF-beta binding, PDGF binding, thyroid metabolic process, Notch signalling and enzyme-linked receptor signalling pathway, somitogenesis, disease mutation and

chromatin organization were enriched (Figure 57). The genes in “disease mutation” were *Gnas, Col2a1, Col11a1, Pax1, Papss2* and *Lpin1*. Not all of the genes actually known to have disease causing mutations were picked up in the DAVID GO analysis. Some of the interesting targets with their Pax1 and Pax9 binding sites are shown in Table 13, and the list of genes with their associated skeletal defects are shown in Table 15.

The targets which had shown opposite directionality (13 genes) in E12.5 double-null and E13.5 *Pax1*$^{-/-}$ were checked for the presence of Pax1 and Pax9 binding sites (Table 13). 9 out of the 13 genes had a direct binding site for Pax9, indicating a potential regulation by Pax9 at E13.5 *Pax1*$^{-/-}$ (Table 13). Interestingly, *Sox5, Col2a1, Wwp2, Ctgf* and *Acan* were among those targets. Several other targets known to be essential for collagen fibril organization happened to be direct targets of *Pax1* or *Pax9*, such as *Hspg2, Acan* and *Col11a1*, all of which were also positively regulated by these TFs. Some of the binding sites (mapped to the mouse genome reference sequence using the UCSC genome browser) are shown as examples in Figures 59 – 61. Besides, both Pax1 and Pax9 had binding site for themselves and for each other. Pax1 had a distal binding site from itself and Pax9 had binding sites in the intragenic, distal and beyond 100 kb regions from its TSS. Moreover, Pax9 was seen bound to the TSS, intragenic and beyond 100kb region of Pax1, while Pax1 had a binding site >300 kb away from the TSS of Pax9.

Having identified the direct and indirect targets of Pax1 and Pax9, the data from the double-null vs WT gene expression profiling and the individual TF binding sites for Pax1 and Pax9 were combined, and some of the selected targets are shown as a network in Figure 58. Selected genes include TFs, genes involved in cell proliferation, cell motion, ECM and cartilage development.

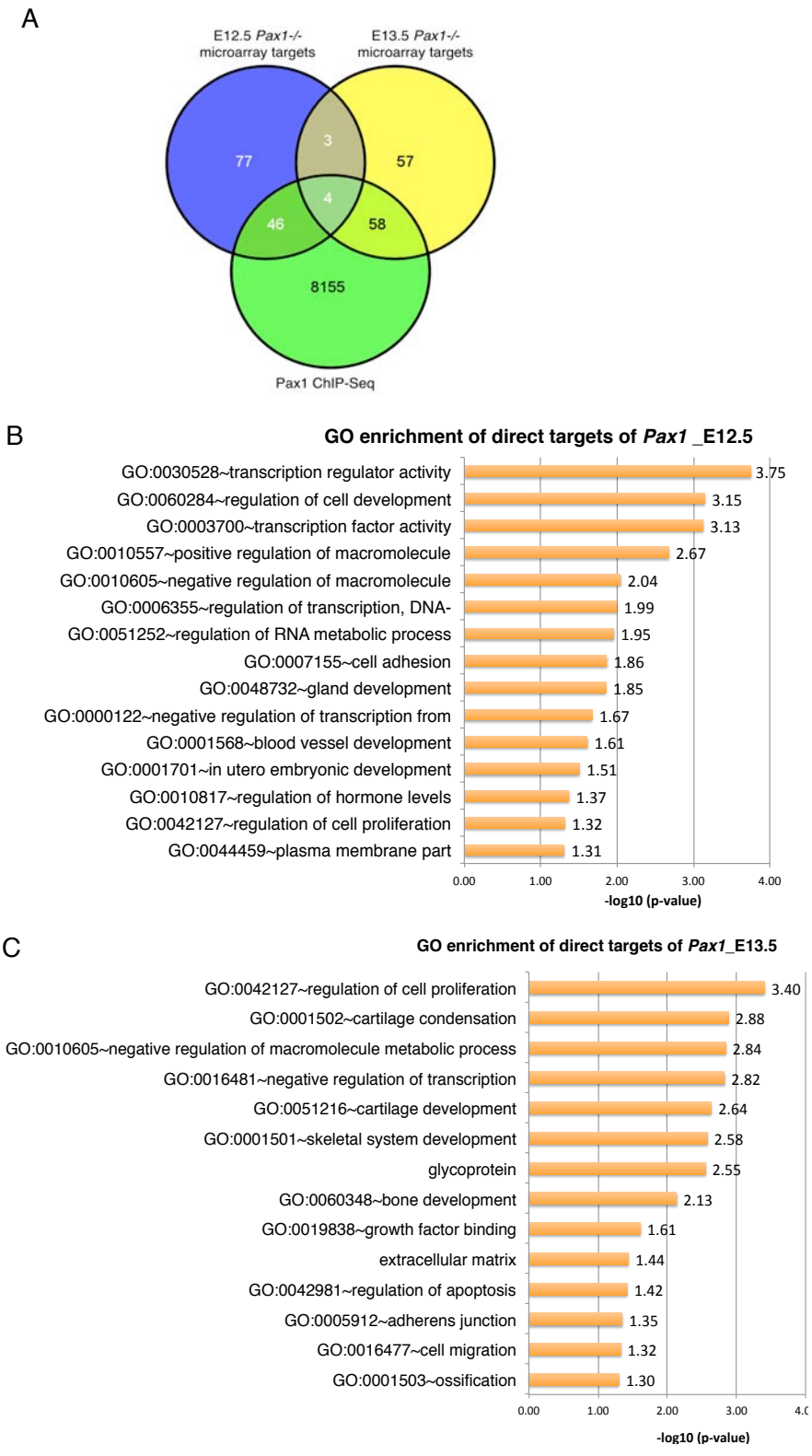


Figure 56: GO enrichment of Pax1 direct targets. (A) Venn diagram of genes overlapping between Pax1 binding sites and E12.5 and E13.5 Pax1 microarray differential gene expression targets. (B & C) GO terms enriched for Pax1 direct targets at E12.5 and E13.5. The -log10 (p-value) are shown at the end of the bars; a value of more than 1.3 is deemed significant (-log10 transformation of a p-value of 0.05 is equal to 1.3).

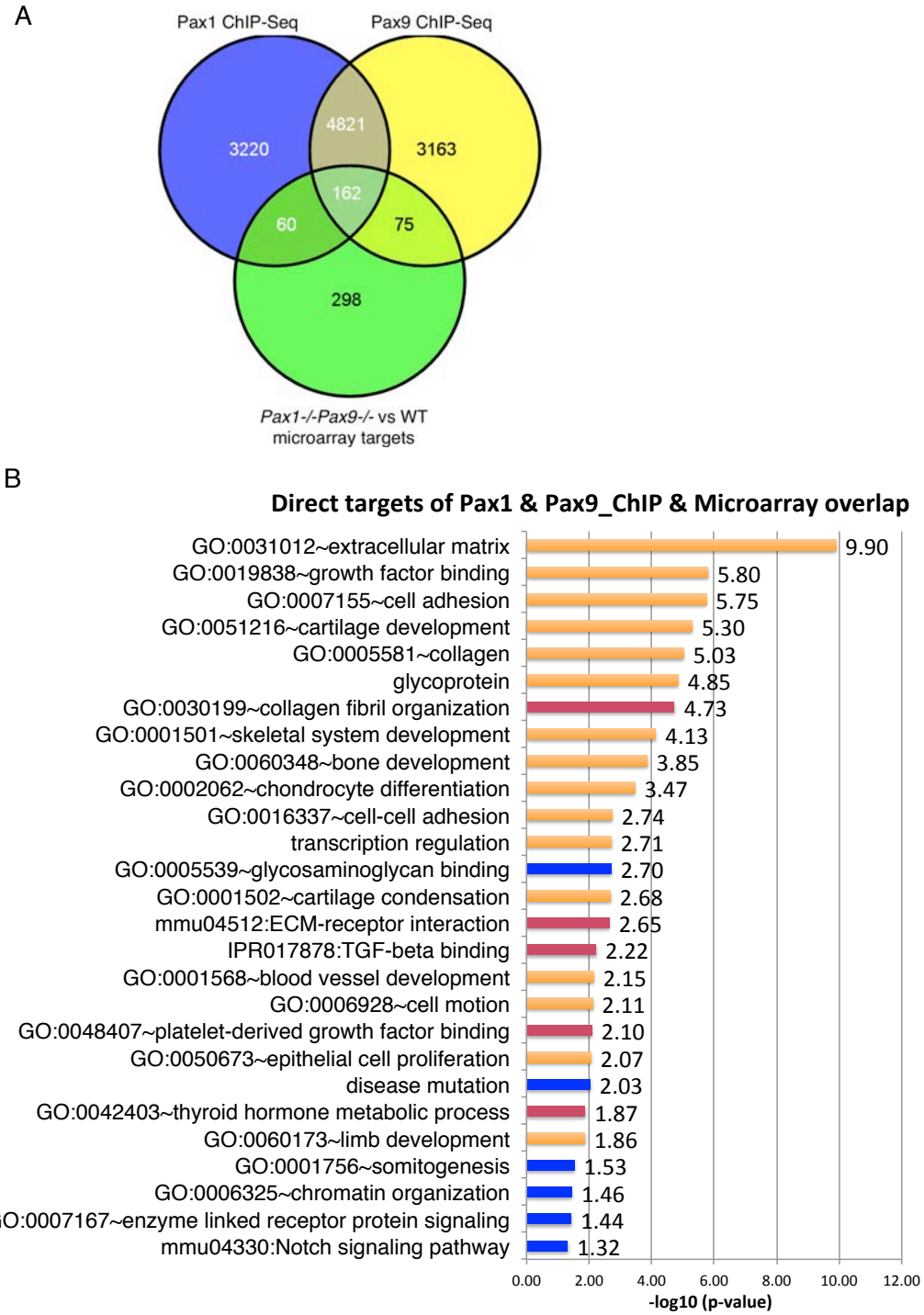


Figure 57: GO enrichment of Pax1 and Pax9 direct binding targets. (A) Venn diagram of genes overlapping between genes differentially expressed in the double-null and the Pax1 and Pax9 binding targets. (B) GO terms enriched for Pax1 and Pax9 direct targets. Only relevant terms are shown. Terms highlighted in red were not found enriched in Pax1 direct targets and were found enriched in the double-null gene expression profiling. Terms in blue were not enriched in the double-null gene expression profiling targets (direct and indirect targets). The -log10 (p-value) are shown at the end of the bars; a value of more than 1.3 is deemed significant (-log10 transformation of a p-value of 0.05 is equal to 1.3.

Table 13: Direct and indirect targets of *Pax1* and/or *Pax9*

Gene symbol	E12.5 *Pax1*$^{-/-}$*Pax9*$^{-/-}$ vs. WT	E13.5 *Pax1*$^{-/-}$ vs. WT	Strand	Pax1 binding site & association	Pax9 binding site & association
Col2a1	Down (3.12)	Up (2.21)	(-)	Distal (-37420)	TSS (+142) Intragenic (-11709) Distal (+22076)
Wwp2	Down (2.46)	Up (1.60)	(+)	Others (+119082)	TSS (-42) Intragenic (+36571, +47089)
Sox5 / A730017 D01Rik	Down (2.23)	Up (1.63)	(-)	Proximal (+6681) Intragenic (-221871, -42187) Others (+166368,+176598, +213791, -447581)	Intragenic (-228324, -136902) Others (+188612)
Extl1	Down (1.78)	Up (3.08)	NIL	None	None
Ctgf	Down (1.57)	Up (1.74)	(+)	Distal (-28987, +45617)	Distal (-25577, +45681) Others (-112141)
Nnat	Down (2.62)	Up (1.78)	NIL	None	Others (-102387, -97245, -96403, -95841, -85698)
Greb1	Down (1.85)	Up (1.54)	(-)	Intragenic (-27723)	None
Cdc25c	Down (1.79)	Up (1.95)	(-)	None	TSS (+160)
Csrp2	Down (1.75)	Up (1.52)	(+)	Others (-132496,-103554)	Others(-125448, -114997)
Tle1	Down (1.68	Up (1.63)	(-)	Others (-168337)	Intragenic (-1706) Others (-771651, -361646, -263771)
Sorl1	Down (1.54)	Up (1.57)	(-)	Distal (+39660)	Others (-263565, -263361,-173813, 173447, -172853)
Mgst2	Down (2.39)	Down (2.39)	(+)	Promoter (-1969)	Distal (+65657, +88790) Others (+211192)
Hspg2	Down (2.37)	Down (1.66)	(+)	None	Intragenic (+17479, +20390)
Crh	Up (6.04)	E12.5 (Up, 2.51) E13.5 (Up, 2.85)	(-)	Others (+45513)	Distal (+44422)
Trpc4ap	Down (1.72)	E12.5 (Down, 1.54) E13.5 (Down, 1.59)	NIL	None	None
Gene symbol	**E12.5 *Pax1*$^{-/-}$*Pax9*$^{-/-}$ vs. WT**	**E12.5 *Pax1*$^{-/-}$ vs. WT**	**Strand**	**Pax1 binding site**	**Pax9 binding site**
Col11a1	Down (1.85)	Down (2.07)	(+)	None	Others (-170205)
Sorbs2/ 9430041 O17Rik	Down (2.09)	Down (1.87)	(+)	None	Intragenic (+180350)
Mmp24	Down (1.53)	Down (1.72)	(+)	Distal (+32671,+37478)	None

Collagen fibril organization				
Gene symbol	**E12.5 *Pax1*$^{-/-}$*Pax9*$^{-/-}$ vs. WT**	**Strand**	**Pax1 binding site**	**Pax9 binding site**
Acan	Down (2.60)	(+)	Intragenic (+20455, +37855)	Others (-126874)
Col3a1	Up (1.53)	(+)	None	Others (-175921)
Col5a2	Up (1.61)	(-)	Distal (+84613)	Others (-262665,-172549)
Dpt	Down (2.42	NIL	None	None
Atp7a	Down (1.52)	NIL	None	None
Cell motion genes				
Psen1	Down (1.58)	(+)	Intragenic (+35387)	None
Etv1	Down (1.61)	(+)	None	Others (-225544)
Sema6a	Down (1.62)	(-)	Others (+324074)	Intragenic (-2499) Others (-189029)
Rpl24	Down (1.58	NIL	None	None
Growth factor binding				
Fbn1	Up (1.53)	(-)	Intragenic (-6096, -119119)	Others (+150155, +174241,+223932)
Col4a1	Down (1.50)	(-)	Intragenic (-89590, -91485) Others (-130183, -150727)	Intragenic (-29532)
Other interesting targets with associated skeletal defects				
BMP4	Down (2.00)	(-)	None	Others (-328284, -214745)
Cspg2/ Versican	Down (1.62)	NIL	None	None
Papss2/ 1810018 P12Rik	Down (1.81)	(+)	Intragenic (+10172)	Intragenic (+10133)
Esrrb	Down (3.10)	(+)	Intragenic (+63181) Distal (-11467)	Promoter (-1578) Intragenic (+13345)
Col19a1	Down (2.57)	(-)	Intragenic (-180946)	Promoter (+3866) Intragenic (-116063, -126269)
Sim2	Down (2.31)	(+)	Others (+339750, +408467)	Others (-128077, -127751,-86933, -54220, +347636, +363871, +389907)
Gnas/ A930027 G11Rik	Up (1.97)	(+)	None	TSS (-88, -137, -328) Intragenic (+9644) Distal (+17113, -95156)
Pax1	Down (2.35)	(+)	Distal (+66304)	TSS (+126) Intragenic (+7442, +8789) Others (+206782, +325684)
Pax9	NIL	(+)	Others (+300930)	Intragenic (+12602, +13122, +13665) Distal (+24632) Others (+210172)

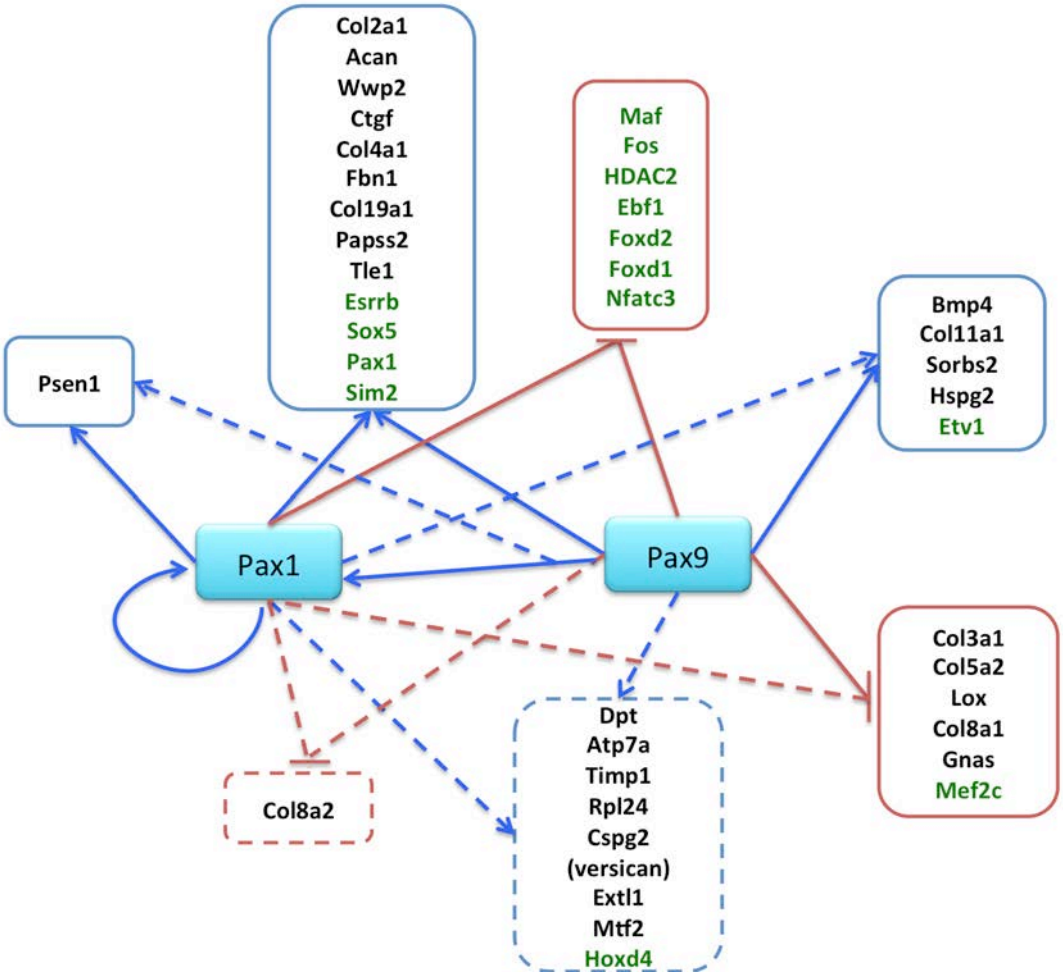


Figure 58: Network representation of selected *Pax1* and *Pax9* targets. Only selected targets like TFs, key cell proliferation, cell motion and ECM genes were used to construct this network for easier visualization. Directionality was based on the results from double-null vs WT gene expression profiling. TFs are shown in green. Blue solid arrows indicate direct and positive regulation; red solid bars indicate direct and negative regulation; dotted lines indicate indirect regulation.

Table 14: Fold-change of selected genes that show gene-dosage dependency

Gene symbol	Double-null vs. WT	Double-null vs. *Pax1*$^{-/-}$	Double-null vs *Pax1*$^{-/-}$ *Pax9*$^{+/-}$	E13.5 *Pax1*$^{-/-}$ vs. WT
Col2a1	Down (3.12)	Down (2.80)	Down (2.38)	Up (2.21)
Wwp2	Down (2.46)	Down (1.95)	Down (1.81)	Up (1.60)
Sox5 / A730017D01Rik	Down (2.23)	Down (2.35)	Down (1.92)	Up (1.63)

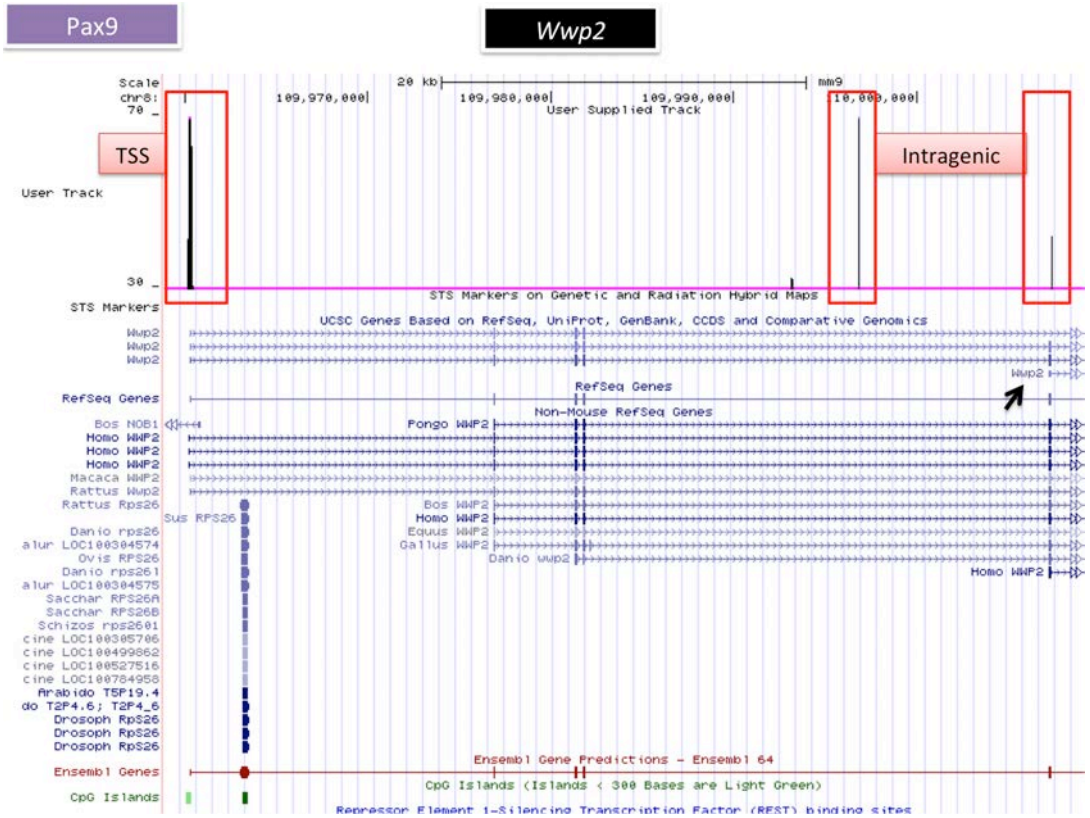


Figure 59: UCSC track for Pax9 binding site at *Wwp2*. The true peaks are boxed in red. The binding peak at the TSS corresponds to CpG islands shown below. Black arrow indicates the start of another Wwp2 transcript.

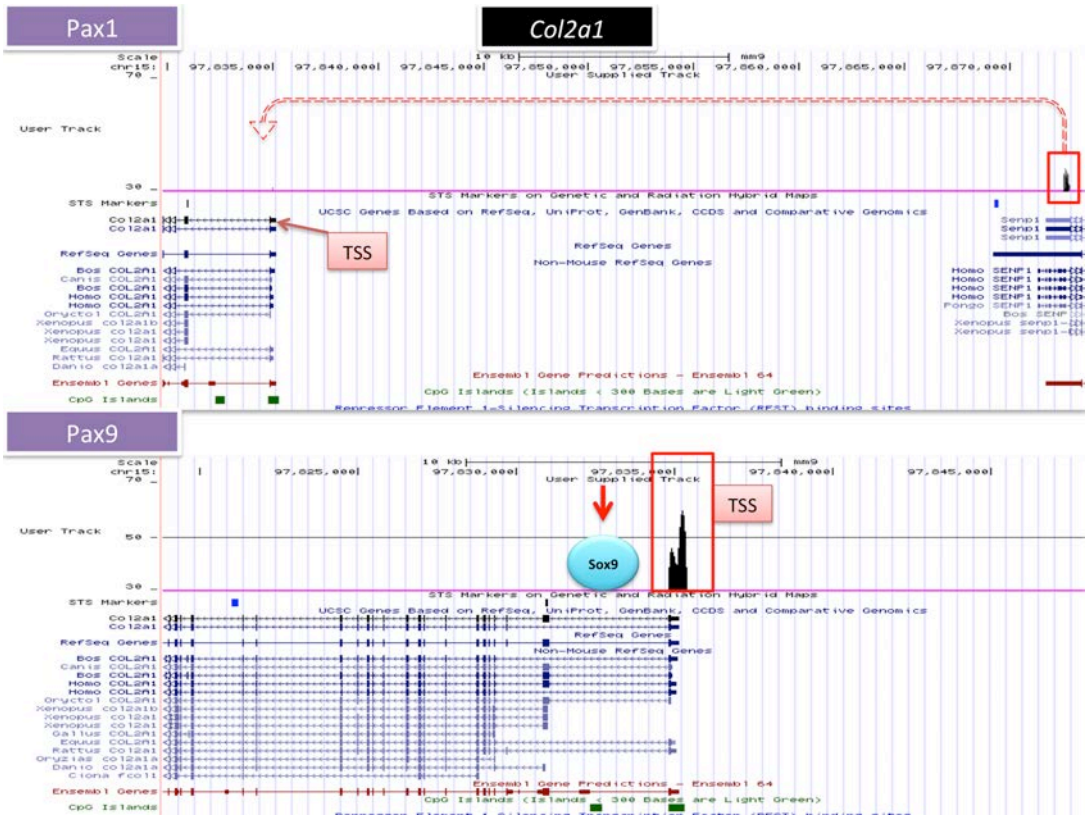


Figure 60: UCSC track for Pax1 and Pax9 binding sites for *Col2a1*. The true peaks are boxed in red. Top: Pax1 binds at a distal region (-37420 bp from the TSS) and its potential looping mechanism to contact the TSS region is shown with a dotted red arrow. Bottom: the peak for Pax9 binding at the TSS (+142 bp from TSS) is shown. The location of known Sox9 binding site in the intronic enhancer region is shown with a red arrow. This region corresponds to the CpG island shown below.

Figure 61: UCSC tracks of Pax1 and/or Pax9 binding sites for selected targets. The true peaks are boxed in red. (A) Pax1 binding site for *Acan*. (B) Pax1 binding site for itself. The potential looping mechanism for Pax1 to contact the TSS region is shown with a dotted red arrow. (C) Pax1 and Pax9 binding sites for *Papss2*. Red arrows indicate the start of alternative transcript for reference.

Table 15: Genes with associated skeletal defects

Gene	Function	Mutant phenotype/ associated defects	Ref
Col2a1 **(Collagen, type II, alpha 1)**	- Collagen fibrillogenesis - Cartilage development - TGF-beta tethering in ECM to modulate its signaling	- *Col2a1*-null mice show alterations in collagen fibril formation. Mutant cartilage was replaced with collagen III and collagen I. This gives rise to abnormal collagen fibres, which are structurally unstable. - show defective IVD formation & abnormal endochondral ossification - Heterozygous mutations in humans give rise to type II collagenopathies, and rare dysplasia like dysspondyloenchondromatosis (DSC). - associated with IVD degeneration	**[127, 129, 162]**
Col11a1 ***(Collagen, type XI, alpha 1)***	- Collagen fibrillogenesis - co-polymerizes with collagen II molecules - IVD metabolism	- *Col11a1* null mice exhibit short limbs, cleft palate, short snout, flared ribs, shortened vertebral column - Mutations in *Col11a1*- polymorphisms are associated with lumbar disc herniation (LDH) - Mutations in *Col11a1* are also associated with connective tissue diseases like Stickler syndrome, characterized by eye/vision (vitreous phenotype) and hearing defects.	**[130, 133, 134]**
Acan / Agc1 ***(Aggrecan)***	- major ECM component; water retention & maintain osmotic pressure in cartilage ECM -cartilage development	- Mutation in mice leads to vertebral misalignments, cleft palate, dwarfism, short tail and limbs, progressive disc degeneration. - Mutations are associated with spondyloepiphyseal dysplasia, premature osteoarthritis and IVD degeneration	**[128, 131, 132, 135, 163]**
Sox5 ***(SRY-box containing gene 5)***	- Cartilage development - ECM synthesis - Chondrocyte differentiation	-*Sox5*-null mice exhibit cleft secondary palate, defects of sternum, narrow thoracic cage, and delayed bone mineralization. - in conjunction with Sox6-null, shows abnormal IVD (inner annulus and nucleus pulposus) and impaired notochordal sheath formation	**[119, 140]**
Ctgf ***(Connective tissue growth factor)***	- Adhesion - Proliferation - Migration - ECM synthesis and organization	- *Ctgf*-null mice show axial skeletal defects, shortened sternum, kinked ribs, cleft palate, impaired osteogenesis, chondrogenesis (delayed endochondral ossification) and growth plate angiogenesis. Mice die from pulmonary hypoplasia.	**[150, 152]**
Fbn1 ***(Fibrillin 1)***	- tissue homeostasis -maintenance of microfibril	- Null mutants exhibit kyphosis, malformed/ overgrowth of ribs - associated with Marfan syndrome (a connective	**[164, 165]**

	integrity -ECM constituent -Protein and calcium ion binding	tissue disorder) - Mutations in *FBN1* associated with kyphoscoliosis	
Psen1 ***(Presenilin 1)***	-Cell adhesion -Anti-apoptosis -Proliferation -Peptidase activity	- null mice show defects of axial skeletons, short tail, impaired brain growth, cerebral hemorrhages, inhibited cleavage of amyloid precursor protein, reduction in Notch signaling, and perinatal death. - associated with Alzheimer's disease	**[154, 166]**
Wwp2 ***(WW domain containing E3 ubiquitin protein ligase 2)***	-ubiquitylation of proteins	- Mutant mice are runted and exhibit cleft palate, domed skull, and short snout.	**[167]**
Hspg2 ***(Perlecan)***	-ECM organization -ECM component -chondrocyte proliferation & differentiation	- Null mice show numerous skeletal dysplasias: dwarfism, cleft palate, short snout, defective vertebral bodies with multiple ossification centres, narrow thorax, and shortened limbs.	**[144, 145]**
Papss2 ***(3'-phosphoadenosine 5'-phosphosulfate synthase 2)***	- sulfation of proteoglycans	- Null mice display dome-shaped skull, short thick tail, overall shortening of axial length, shortened limbs, delayed growth and defective bone formation. - Mutations in human PAPSS2 are associated with Brachyolmia, a type of skeletal dysplasia mainly affecting the spine.	**[168, 169]**
Sim2 ***(Single-minded homolog 2 (Drosophila))***	- transcription factor activity	- Null mutants possess cleft palate defect and malformed pterygoid processes. - postnatal death	**[170]**
Rpl24 ***(Ribosomal protein L24)***	-cell cycle/ cell proliferation -ribosome component -ribosome biogenesis -protein synthesis	- Heterozygotes show short kinked tail, defects of the spine, eyes and possess ventral white spots, and exhibit malocclusion - homozygotes show perinatal lethality	**[171]**
Hoxd4	-anterior-posterior	-Mutants show homeotic transformations of 2nd cervical vertebrae (C2 → C1); malformed neural	**[172]**

(Homeobox D4)	specification of axial skeleton (confer positional information)	arches (C1 → C3); defective basioccipital bone	
Hoxc9 ***(Homeobox C9)***	-anterior-posterior specification of axial skeleton	- Homozygotes show slow growth; transformation of lumbar to thoracic vertebrae (L1 → T10); defects of sternum and ribs and mutants have hunched backs.	**[173, 174]**
Mtf2 (a.k.a Pcl2) ***(Metal response element binding transcription factor 2)***	-polycomb-like gene -binds to *Hox* genes -mediates repression of *Hox* genes	- Null mutants show vertebral transformation defects.	**[175]**
Ankrd11 ***(Ankyrin repeat domain 11)***	-bone development, skeletal system morphogenesis	- ENU-induced heterozygous mutants show craniofacial defects, develop osteoporosis and osteopenia.	**[176]**
Col5a2 ***(Collagen, type V, alpha 2)***	-collagen fibril organization	- Homozygous mutants exhibit kyphosis and lordosis to varying extent, defects of skin and cornea, slowed bone growth and decrease body weight.	**[177]**

3.6.5 Discussion

The identification of the Pax1 and Pax9 direct binding targets has helped to distinguish the direct and indirect targets of *Pax1* and *Pax9*. Almost one-third of the targets (27.2%) differentially expressed in the *Pax1/Pax9* double-null mutants are direct targets of both Pax1 and Pax9. Certainly not all of the distal/ intragenic binding sites identified via TF mapping represent enhancers. They could also be silencers, assisting in the repression of the gene associated with the binding site. By comparing both differential gene expression trend and the TF mapping (ChIP-Seq) data, we can postulate if they ought to be enhancers or silencers. Moreover, not all of the binding sites overlap with the gene expression profiling data. These binding sites, as mentioned earlier, could be non-functional at the particular developmental time-point,

or in that particular tissue, and are probably simply poised, waiting for the co-factors or partners to be available.

Moreover, coupling the gene expression profiling data with the TF mapping (ChIP-Seq) data reiterates that Pax1 and Pax9 are indeed true regulators of cartilage development genes, some of which are discussed in further detail below.

(I) *Pax1* and *Pax9* are regulators of key cartilage development genes

- **Regulation of *Wwp2, Col2a1* and *Sox5* by Pax1 and Pax9**

Interestingly, Pax1 and Pax9 had binding sites near *Col2a1*, *Wwp2* and *Sox5*. As mentioned in section 3.5.4.2, owing to the opposite directionality that was observed in the E13.5 *Pax1*$^{-/-}$ and the double-null, one of the hypothesis was that *Pax1* might be repressing while *Pax9* might be activating it. The other hypothesis was that *Pax9* was also regulating these targets to up-regulate them, in response to the loss of *Pax1*. Considering that these targets all possess a binding site for *Pax9*, some of them at the TSS itself (*Wwp2* and *Col2a1*), the compensation hypothesis appears more plausible. Thus, *Pax9* most likely did play a role in the up-regulation of these targets (*Col2a1*, *Wwp2* and *Sox5*) as seen in E13.5 *Pax1*$^{-/-}$, possibly in conjunction with *Sox9*.

- **Sox9-Wwp2-Med25-Sox5-Sox6 complex in *Col2a1* transcription**

Sox9, a master regulator of chondrogenesis, is well-known to function by up-regulating key genes of the cartilage ECM component such as *Col2a1* [120], *Col11a1* [124], *Ctgf* [124], *Acan* [138] and *Wwp2* [124]. One of the well-studied targets of *Sox9* is *Col2a1*, whereby it binds to an intronic enhancer region of *Col2a1* to regulate its transcription [120, 178]. Additionally, this regulation involves the formation of a

complex of *Sox9* with co-factors like *Wwp2* and *Med25* [124]. *Wwp2* is WW domain containing E3 ubiquitin protein ligase 2, required for ubiquitylation, while *Med25* (mediator of RNA polymerase II transcription, subunit 25 homolog (yeast) is a cofactor protein. It was shown *in vitro* that *Sox9* also regulates *Wwp2*, by binding to its promoter region (luciferase assay with *Wwp2* promoter which inherently contains the Sox9 motif). Wwp2 then mono-ubiquitylates Sox9 and enhances its transcriptional activity, while Sox9 mediates the translocation of Wwp2 into the nucleus, where the Sox9-Wwp2-Med25 complex drives *Col2a1* expression. Unlike poly-ubiquitylation which results in proteasomal degradation of the protein, mono-ubiquitylation is a form of post-translational modification [167]. Such mono-ubiquitylation mediated enhancement of transcriptional activity of TFs was shown for *Goosecoid* (*Gsc*) as well [167]. Thus, Wwp2 was shown to be associated with Sox9 in a complex with the co-factor protein Med25, which then bind to the intronic enhancer of *Col2a1* to drive its expression [124]. Sox5 and Sox6 which are known to enhance the transcriptional activity of Sox9 were postulated to be bound to Sox9 in this complex as well [124, 139].

Pax1 and Pax9 may have a role in this complex. Firstly, Pax9 has a binding site at the TSS and intronic regions of *Wwp2*. An intronic region between the exon4 and exon5 of *Wwp2* has been reported to be bound by Sox9 (ChIP-PCR on ATDC5 cells), which lies at around +38 kb from the TSS [167]. The intragenic regions identified in this Pax9 in *in vivo* TF mapping are +36,571 bp and +47,089 bp from the TSS. The first intronic peak is in close proximity to the known Sox9 intronic binding site, while the second intragenic peak corresponds to the start site of another *Wwp2* transcript. Hence, Pax9 and Sox9 may be brought in close proximity to each other via chromatin looping, which brings Sox9 from the intronic region to the promoter. This also shows that the multiple binding regions identified for each gene in the TF

mapping probably correspond to the multiple points of contact involved in the chromatin looping. Similarly, Pax1 has a binding site > 100 kb away from *Wwp2* (+119,082 bp), and might also be involved in the regulation of *Wwp2*.

A simplified diagram of this postulated looping mechanism is shown in Figure 62 for Pax9 and Sox9. Also, *Pax9* positively regulates *Wwp2* owing to the decrease in transcript levels by 2.46-fold in the double-null. There is also a gradual decrease seen with increasing loss of *Pax9* (1.95-fold down with 2 copies lost; 1.81-fold down with 1 copy of *Pax9* in the absence of *Pax1;* Table 14). Furthermore, *Wwp2*-null mice were runted and had craniofacial defects [167]. Similarly, knock-down of *wwp2* in zebrafish gave rise to a palatal phenotype. This phenotype in zebrafish was only partially rescued by the co-injection of *sox9a* and *sox9b* transcripts [124], indicating that other factors are involved in the control of *Wwp2*, which could be *Pax9* and/or *Pax1*. Indeed, *Pax9*$^{-/-}$ mice are known to possess cleft palate too, similar to *Sox9* mutants [179], proving that *Pax9*, *Sox9* and *Wwp2* are all interacting in a common pathway.

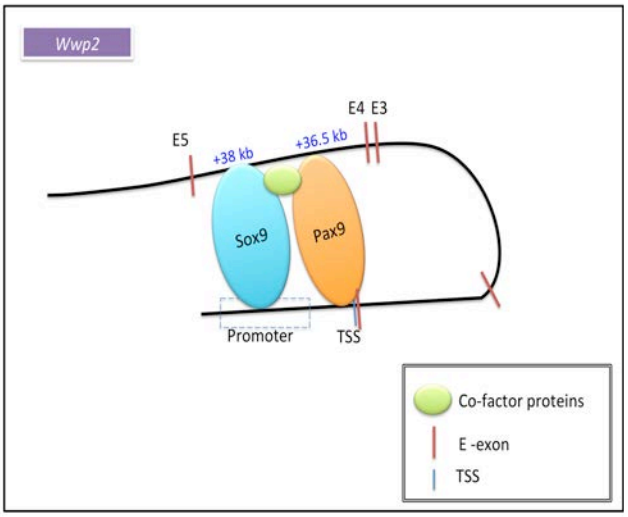


Figure 62: Postulated model of *Wwp2* regulation by Pax9 in co-operation with Sox9. Sox9 potentially contacts the promoter region which possesses the Sox9 motif, and also the intronic region (+38 kb from TSS) identified through *in vitro* studies by others. Binding site for Pax9 has been identified in the *in vivo* TF mapping at the TSS (-42 bp) and intragenic region (+36571 bp).

Secondly, Pax9 and Pax1 have binding sites for *Col2a1*; Pax9 contacts the TSS (+142), intragenic (-11,709) and distal region (+22,076) and Pax1 binds at a single distal region (-37,420). As mentioned earlier, Sox9, Sox5 and Sox6 complex contact the intronic enhancer region (at around +2,113 to +2,343 bp from TSS, in the intron1) to drive *Col2a1* expression [120, 139, 178]. In addition, the loss of *Pax9* (in the absence of *Pax1)* leads to a decrease in *Col2a1* transcript levels in a gene-dosage dependent fashion, indicating an activating effect on *Col2a1*. In the comparison of double-null vs $Pax1^{-/-}$ and double-null vs $Pax1^{-/-}Pax9^{+/-}$, *Col2a1* was decreased by 2.80-fold and 2.38-fold respectively. In the double-null vs WT, the decrease was 3.12-fold (Table 14). Thus, the dramatic decrease in *Col2a1* levels in the double-null, coupled with presence of binding sites by both *Pax1* and *Pax9*, reflect a potential mechanism of *Col2a1* regulation by the *Pax*es with the *Sox*es. It is not yet known if Pax1 and Pax9 form a complex with the Sox9-Wwp2-Med25 proteins as well, and what other intermediary proteins may come into play. Similar to *Pax1*/*Pax9* double-null mutants, loss of *Col2a1* also gives rise to cleft palate and axial skeleton defect [127]. As mentioned earlier, *COL2A1* defects in humans is known to give rise to kyphoscoliosis – a defect associated with *PAX1* and *PAX9* as well [83, 86, 129].

On top of these, *Sox5* was also down-regulated in the double-null and showed a similar decrease in transcript with progressive loss of *Pax9* (Table 14). Pax1 and Pax9 also have binding sites in proximal, intragenic and > 100 kb from the TSS of *Sox5*.

All of these observations indicate a high possibility for the *Pax* genes being directly involved in the regulation of *Col2a1, Wwp2* and *Sox5*. While the shared craniofacial and axial defect phenotype of *Sox5/Sox6* [140], *Sox9* [179], *Wwp2* [167], *Col2a1* [127] and *Pax1/Pax9* [62] mutants all clearly reflect the interplay of these

factors *in vivo*, further *in vitro* validations via luciferase assays, EMSA, co-IP and pull-down assays could help to affirm these regulatory mechanisms. Moreover, through such assays, we can confirm if *Pax1* and *Pax9* are involved in these regulations at the same time and how these two TFs can compensate for each other. Most importantly, the regulation of all of these factors by Pax1 and Pax9 is novel and their association with the *Sox* genes has never before been shown.

- ***Acan* is regulated by *Pax1* and *Pax9*, possibly directly**

Aggrecan, an important ECM component was also positively regulated by *Pax1* and *Pax9* based on the double-null gene expression profiling results. While Pax1 had binding sites in the intragenic regions of *Acan*, Pax9 binding site was >100 kb upstream of the TSS of *Acan*.

Notably, mouse *Pax1* was shown to induce aggrecan expression in the explants of chick presomitic mesoderm, independent of *Shh* [57]. This corroborates with the results of this study, whereby *Pax1* positively regulates *Acan* at E12.5-E13.5, possibly directly via its intragenic binding sites.

(II) *Bapx1* was not regulated by *Pax1* or *Pax9* at E12.5 and E13.5 in the IVD anlagen

As mentioned in the introduction, *Bapx1* is an important TF for the transition of pre-chondrogenic cells to chondroblasts [53]. *Bapx1* was identified as one of the direct binding targets of Pax1 and Pax9 through *in vitro* assays (**-880 to -844** from the TSS) (Figure 63) [57]. It is indeed the only known direct binding target of Pax1 and Pax9 in literature. Also, the authors had claimed a reduction in *Bapx1* levels only in the sclerotome in the double-null embryos based on their E10.5 whole-mount in situ hybridization and E11.5 sectioned in situ hybridization data [57].

However, in our gene expression profiling analysis which was performed on highly pure population of *Pax1/ Pax9*-specific cells, *Bapx1* was not differentially expressed in any of the single-null, *Pax1*$^{-/-}$*Pax9*$^{+/-}$ or double-null comparisons. The loss of *Bapx1* expression seen in the double-null embryos by the authors could indeed be due to the prevalent loss and mis-localization of sclerotomal cells in the double-null (refer: Figure 30). Moreover, in our *in vivo* TF mapping data, *Bapx1* did not have a Pax1 binding site, but had a Pax9 binding site in the intragenic region (-2445 bp from TSS, (-) strand).

These observations were not too surprising as it was postulated by the authors [57] that Pax1 and Pax9 may only be required to initiate the expression of *Bapx1* at E9.5 and not for its maintenance. In the wild-type, chondrocytes continue to express *Bapx1* even when *Pax1/Pax9* expression is reduced in the chondrocytes from E12.5 onwards [57]. The regulatory loop between *Sox9* (a master-regulator of chondrogenesis) and *Bapx1* is believed to maintain *Bapx1* expression in the chondrocytes during chondrogenesis [55]. In addition, a Sox9 binding site was identified at **-868 to -852** from the TSS of *Bapx1*, which largely overlaps with the Pax1/ Pax9 binding site identified in the *in vitro* study [54]. Similarly, Meox1/ Meox2 (TFs also important in sclerotome differentiation) also possess binding sites within the *Bapx1* promoter, immediately adjacent to the Pax1/ Pax9 binding site (**-840 to -810)** [54-57]. Therefore it is highly likely that by E12.5 and E13.5, the stages at which the TF mapping was done, *Bapx1* promoter is actually bound by a different set of proteins (Sox9/ Meox1/ Meox2) and hence not bound by Pax1 and/or Pax9 (Figure 62 for illustration). Furthermore, *Sox9* or *Meox1/ Meox2* may compensate for the loss of *Pax1/ Pax9* and thus help to initiate the expression of *Bapx1* at E9.5. These hypotheses also correlate with the lack of differential expression of *Bapx1* in the absence of *Pax1* and/or *Pax9*. Importantly, they all the more reflect the need for *in*

vivo studies to complement the *in vitro* results, and together they greatly assist to comprehend such dynamic changes in binding sites as well as reveal the complex interplay of multiple factors in gene regulation.

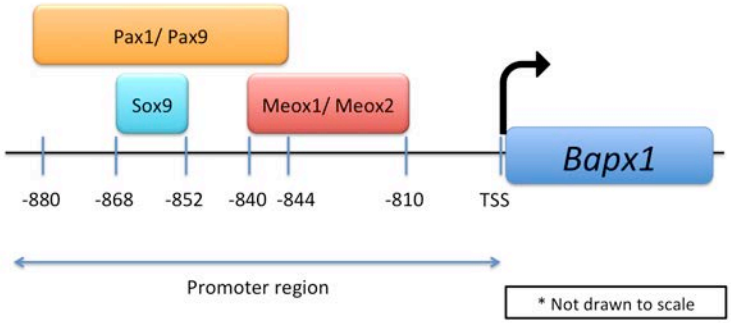


Figure 63: Illustration of binding sites identified for *Bapx1* promoter *in vitro* by other studies. The positions of binding sites shown are in base pairs from the TSS of *Bapx1*. Refer to main text above for explanation.

(III) Regulation of *Pax1* and *Pax9*

The *Pax1* transcript levels decreased in the *Pax1/Pax9* double-null and also when *Pax9* was lost in the absence of *Pax1* (double-null vs WT - 2.35-fold down, $Pax1^{-/-}Pax9^{+/-}$ vs WT – 1.63-fold down and double-null vs $Pax1^{-/-}$ - 1.50-fold down). On top of that, Pax1 had a distal binding site associated with itself, indicating a potential self regulation. While the initiators of *Pax1* are *Shh* and *Nog*, Pax1 might be partly required to maintain itself under normal circumstances. Considering that Pax1 itself was not differentially expressed in the $Pax1^{-/-}$, it seems likely that there may be other regulators of *Pax1* (eg. *Shh* or *Nog*) involved in the maintenance.

Furthermore, since Pax9 also has binding sites for *Pax1*, it might also have a role in regulating *Pax1*. When *Pax9* is also lost, the decrease in *Pax1* levels becomes more prominent. Then again, validation of these binding sites are required to

ascertain if there is a true direct regulation of *Pax1* by Pax1 and Pax9, or if the observed decline in *Pax1* levels is an indirect consequence.

On the other hand, Pax9 had binding sites in its own introns and several distal and >100 kb sites associated with itself, and a Pax1 binding site >300 kb from the TSS of *Pax9*. However, there was no differential expression of its transcripts in the double-null or any of the 3 allele KO or *Pax1*$^{-/-}$ mutants. A similar observation was made by Peters *et al* (1999), whereby Pax9 transcript expression did not decrease in the double-null embryos [62]. The authors proposed that neither Pax1 nor Pax9 are required for the transcription of *Pax9* in the vertebral column [62]. Hence, it is likely that *Pax9* does not self-regulate, nor regulated by *Pax1* in the vertebral column cells. The observed binding sites maybe functional in some other tissue or regulates some other neighbouring gene since they are not found in the TSS or promoter region of *Pax9*.

(IV) Connection to the chondro-osteogenic pathway

Based on the genome-wide gene expression profiling and binding site of Pax1 and Pax9, *Sox5* appears to be the key point of connection to the *Pax* genes in the chondro-osteogenic pathway.

In our own lab's analysis of *Sox9* and *Sox5/Sox6* double-null mutants, *Pax1* was negatively regulated by these *Sox* genes at E12.5 (2.71-fold up in *Sox9*-null) and E13.5 (1.71-fold up in *Sox5/Sox6* double-null) respectively. In fact, *Sox9* had a direct binding site for *Pax1* in a distal region (+331,470 bp). This indicates a potential feed-back loop, which could explain the initial co-expression of the *Sox* and *Pax* genes in the sclerotome and IVD anlagen and the eventual down-regulation of *Pax1* in the differentiating chondrocytes of the cartilage (Figure 64 for illustration). Such negative feed-back loops in a network are known to diminish expression levels of the target gene [5, 180]. Figure 64 shows the expression of *Pax1* in the IVD anlagen at E13.5.

Sox9 and *Sox5* are also expressed in the *Pax1*-expressing cells (Figure 36 and Table 7) at this stage. Subsequently by E15.5, the IVD anlagen have differentiated into two morphologically distinct components, inner cartilaginous and outer fibrous annulus fibrosus. At this stage, Pax1 and Pax9 expression becomes restricted to the outer annulus fibrosus. The negative feed-back loop potentially explains how the *Pax1* expression becomes down-regulated in the mature cartilage. Interestingly, at E12.5 and E13.5, none of these *Sox* genes regulated *Pax9* (our lab's analysis of the *Sox* mutants). What factors regulate *Pax9* in the IVD is still unknown.

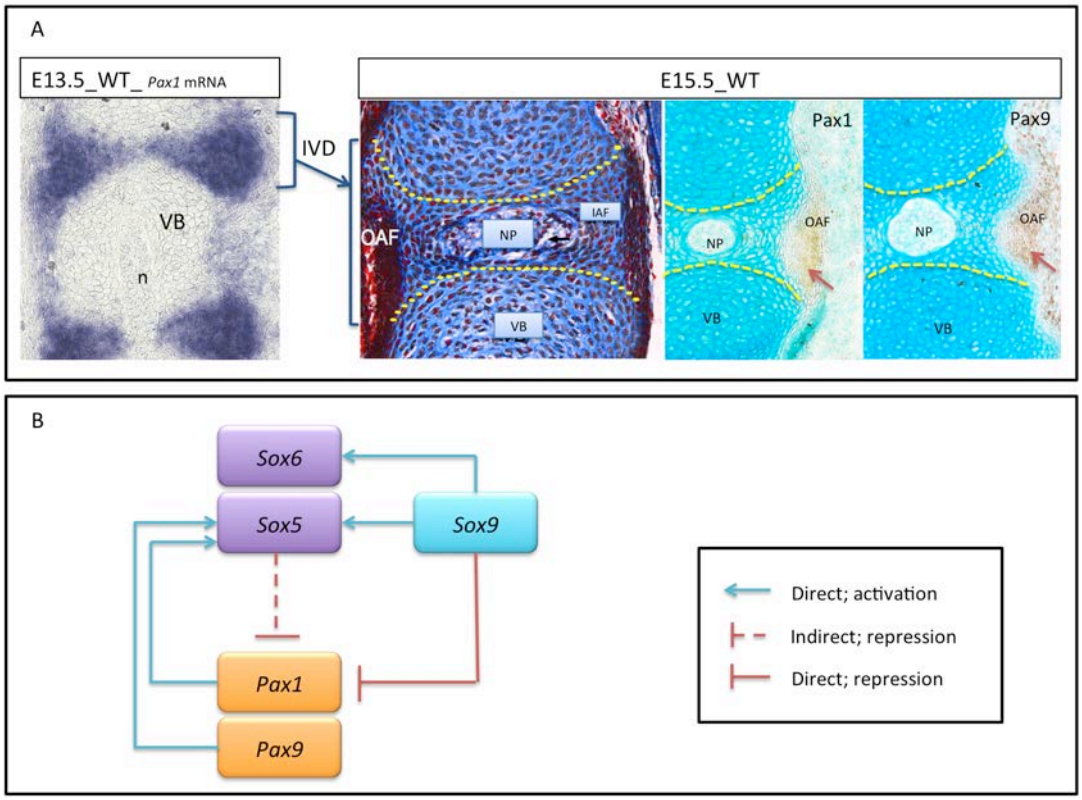


Figure 64: Expression of Pax1 and Pax9 and morphology of IVD during E13.5 and E15.5. (A) Left: *Pax1* mRNA expression is in the IVD anlagen at E13.5. Right: the IVD anlagen develop into cartilaginous inner annulus and fibrous outer annulus fibrosus by E15.5. Pax1 and Pax9 protein expression could be seen restricted to the outer annulus cells at E15.5. Yellow dotted lines demarcate prospective boundaries between the vertebral body and the IVD. Sections were counter-stained with Mallory's tetrachrome staining (left) or alcian blue (right). (B) Regulatory connections between *Sox5*, *Sox6*, *Sox9*, *Pax1* and *Pax9* in the IVD anlagen based on the results from this study and our lab's analyses of *Sox5/Sox6* double-null and *Sox9* mutants in chondrogenic cells. It is assumed that *Pax1* and *Pax9* are acting synergistically and not antagonistically in this model (refer to section 3.5.4.2 for explanation). VB – vertebral body; IVD – intervertebral disc; n – notochord; NP – nucleus pulposus; IAF – inner annulus fibrosus; OAF – outer annulus fibrosus; WT – wild-type.

3.7 Conclusion

In conclusion, *Pax1* and *Pax9* clearly appear to have a role in regulating the early functions of IVD tissue morphogenesis, i.e. cell proliferation, cell adhesion, mesenchymal condensation, ECM organization and cartilage development. The annulus fibrosus of the IVD is derived from *Pax1* and *Pax9* expressing sclerotomal cells. Subsequently, by E15.5, the annulus fibrosus differentiates into inner cartilaginous and outer fibrous tissues. Since *Pax1* and *Pax9* are positively regulating cartilage development genes at E12.5, when the IVD anlagen is formed, *Pax1* and *Pax9* may be assisting the *Sox* genes to activate the early chondrogenic genes. As seen from the genome-wide data, some of the key cartilage development genes regulated by the *Pax* are well-known to be regulated by the *Sox* trio as well. Once the chondroblasts mature into chondrocytes in the inner annulus fibrosus, these *Pax* genes are down-regulated in the chondrocytes. *Pax1*- and *Pax9*-expressing cells then become restricted to the outer annulus fibrous cells where they may be involved in more patterning functions. While the lack of sufficient double-null embryos precluded further histological analysis of their vertebral column, studies in spontaneous *Pax* mutants provide support for this hypothesis. In the $Un^{s/+}$ adult mice, the annulus fibrosus structures were malformed and had a fibrous appearance instead of forming a normal hyaline cartilage. Also, the normally bony acromion process of the scapulae was ligamentous in the mutants [71, 76].

Notably, the presence of dorsal cartilage structures in the *Pax1/Pax9* double-null mutants indicate that *Pax1* and *Pax9* genes are not needed for overt chondrogenic differentiation [62]. They are not master regulators of chondrogenesis but probably perform essential accessorial functions in chondrogenesis. Their primary roles may still be in ensuring proper formation of mesenchymal

condensations in the right places by controlling cell proliferation, cell adhesion and migration.

Thus, based on the literature and the regulatory connections that have been deciphered in this study, I propose the following model to depict the involvement of the various TFs in the IVD development (Figure 65). As mentioned earlier, in terms of sclerotome-derived cells, the focus has largely been on the regulatory mechanisms involved in vertebral bodies. The regulatory pathways involved in IVD morphogenesis require equal attention. Shown here is just a small circuit of the larger network. In our lab, we have generated similar genome-wide data sets for the *Sox5*, *Sox6*, *Sox9, Bapx1, Runx2 and Runx3*. By connecting this *Pax* data with the *Sox* data, we can generate the bigger network and from there dissect out such smaller regulatory connections and possibly identify network motifs. That is indeed the essence of GRNs – to be able to represent the interplay of various TFs in a global scenario, as well as allow the interrogation of the network at the individual gene level.

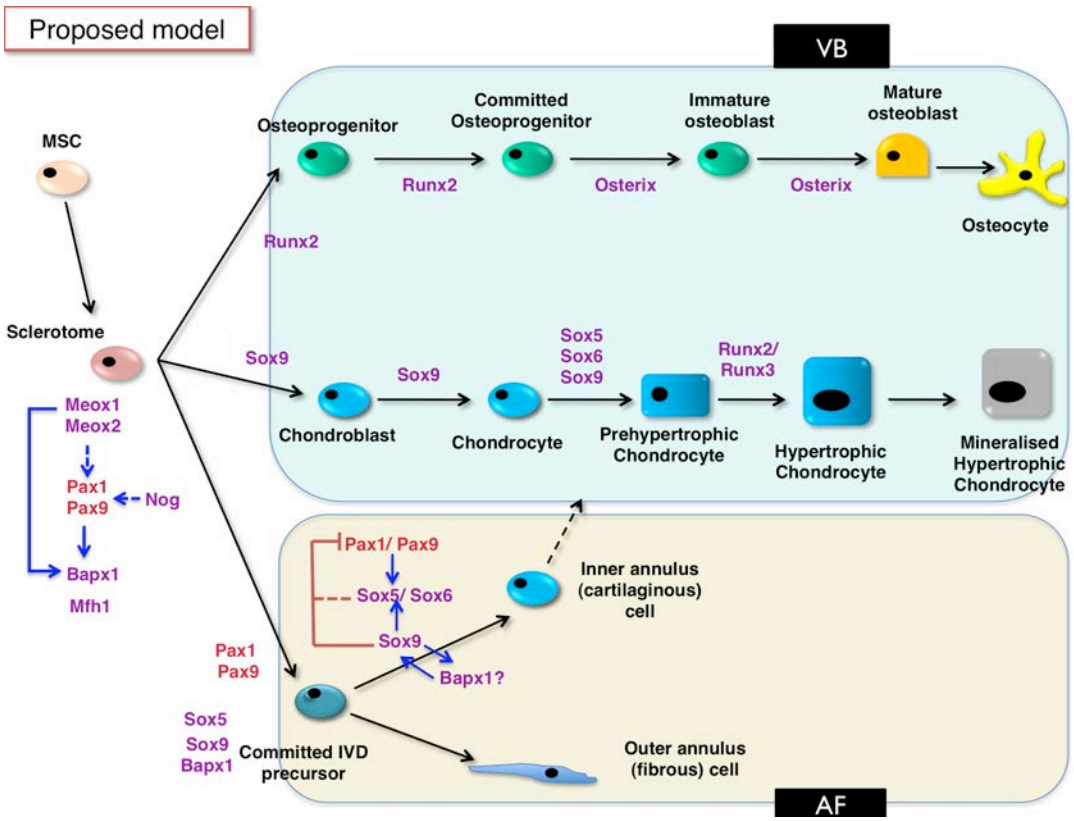


Figure 65: Proposed model of regulatory connections between TFs involved in the sclerotome-derived components of the IVD development. The connections shown are based on literature search (see introduction) and this study. Blue arrows represent positive regulation; red bars represent negative regulation; dotted lines indicate indirect regulation and solid lines indicate direct regulation. Note: It is not yet known if *Meox1/Meox2* regulation of *Pax1/Pax9* is direct or indirect. For simplification purposes, self-regulatory connections are not shown. VB – vertebral body; AF – annulus fibrosus; MSC – mesenchymal stem cell; IVD – intervertebral disc.

3.7.1 Future work

The direct and indirect targets of *Pax1* and *Pax9* have thus been identified. The immediate future work would involve *in vitro* validations of the key regulatory connections that have been deciphered.

This includes luciferase assays and EMSA (electrophoretic mobility shift assay) to confirm if *Pax1* is an activator or repressor of *Sox5* and *Col2a1* genes which showed opposite directionality and had a binding site for these genes. Considering that *Pax1* and *Pax9* are paralogous genes and are known to function

synergistically, the current hypothesis is that *Pax1* is an activator of *Sox5* and *Col2a1*. Activation of these targets by *Pax9* would also be validated *in vitro* similarly. The *Col2a1* regulation by *Pax1/Pax9* might potentially occur in conjunction with *Sox9* and its associated complex of proteins (Sox5, Sox6, Med25 and Wwp2). Therefore, it would be assessed if these proteins are associated with *Pax9* and/or *Pax1*. For example, HA-tagged *Pax9* or *Pax1* can be over-expressed *in vitro* in chondrogenic cell lines (ATDC5 or C3H10T1/2 cells) and myc-tagged Med25 can be transfected into the same cells. After 24 – 48 hrs, total protein lysates can be immunoprecipitated using anti-HA antibodies and probed on Western blot with anti-myc antibody. Such co-immunoprecipitation assays can be performed for Pax1 or Pax9 with each of the proteins separately to assess their interactions at protein level as a protein complex.

Moreover, the chromosome looping mechanism for Pax1 or Pax9 to contact the promoter regions of particular genes (eg. in the regulation of *Wwp2*) can be affirmed by Chromosome Conformation Capture (3C) or Chromosome Conformation Capture Carbon Copy (5C) assays.

While the target genes of *Pax1* and *Pax9* in the IVD have been identified, the question of how *Pax1* and *Pax9* compensate for each other remains unanswered. Considering that the paired domains of Pax1 and Pax9 are highly similar (differing only at 3 amino acids within the paired box) [50], and the fact that in *in vitro* assays they could interact with similar sequences (e5 sequence [50], *Bapx1* promoter [57]), it is possible that they have the ability to bind to each other's binding sites in the absence of the other protein. That is, in the absence of *Pax1*, Pax9 might be able to bind to Pax1 binding sites and *vice versa*. Whether such dynamic changes in binding sites occur can be assessed *in vivo*. For instance, TF mapping can be performed using anti-Pax9 antibodies in the $Pax1^{-/-}$ embryos. Similarly, TF mapping on $Pax9^{-/-}$ embryos using anti-Pax1 antibodies can be performed. The results of such

experiments would provide unique perspectives on the compensation mechanisms at a molecular level. It would also be an evidence of the existence of alternate gene regulatory pathways in the event of abnormal conditions (in this case the loss of *Pax1* or *Pax9*) or diseased states.

3.7.2 Challenges & Improvements

The main challenge in this study was to work with small numbers of cells, which posed limitations on the amount of RNA that was available for gene expression profiling analysis. This also hampered the identification of *Pax1* targets at E9.5 which would reveal the earliest functions of *Pax1* in the sclerotome. Moreover, since *Pax9* transcripts are only beginning to be expressed at that stage, the targets identified at E9.5 could possibly be unique targets of *Pax1*.

Furthermore, the regional differences in the severity of defects in the *Pax1*$^{-/-}$ mutants highlights the inherent differences in the regulatory mechanisms involved in the development of different vertebral segments. To capture such mechanisms, one has to look at individual vertebral segments or even at a single cell level. This was not possible before owing to the limitations of the technology. However, emerging technologies like RNA-sequencing and the availability of Illumina's Clontech SMARTer^{TM} Ultra Low RNA Kit for Illumina Sequencing (for RNA-Seq library construction) are promising and such intricate studies could be possible in the future.

Besides these, since this study was focused on the E12.5 - E13.5 embryos, a few days after the initiation of *Pax1* and *Pax9* expression, one may argue that we are looking at gene expression changes caused by the phenotype rather than the actual functions of the gene. Here, this issue has been circumvented by comparing the differentially expressed genes with the TF mapping performed on WT tissues. If the genes do have a binding site for Pax1 and Pax9 in the WT at the same stage and also are differentially expressed, then they are most likely not an effect of the

phenotype, but true targets of *Pax1* and/or *Pax9*. Nevertheless, an alternative approach would be to make inducible conditional knock-out mice whereby the gene-of-interest can be flanked by *loxP* sites and then mated to mice with inducible-*Cre* alleles. The construct can be designed in such a way that only upon floxing the gene-of-interest, the EGFP reporter will be expressed. This way, Cre expression can be induced at specific time-points, which will result in the deletion of the gene only at those specific time-points, and the cells with the deleted gene can be isolated using FACS for downstream analyses.

CHAPTER 4 - CONCLUSION

The main aim of this study was to identify the direct and indirect targets of *Pax1* and *Pax9*, in a cell-type specific manner, during IVD development. This has been achieved using a traditional transgenic approach and two of the currently widely used genome-wide technologies - microarray and ChIP-Seq. Importantly, this is the very first study in which the target genes of *Pax1* and *Pax9* have been identified, *in vivo*, in a specific cell type.

Using the strategies mentioned in section 1.6.2, 130 genes and 122 genes were differentially expressed in *Pax1*$^{-/-}$ at E12.5 and E13.5. In general, these genes were enriched for cell adhesion, transcriptional regulation, macromolecule biosynthesis, skeletal system development and regulation of apoptosis, all of which are relevant to mesenchymal condensation process. Only a small number of genes were differentially expressed at these early stages in the *Pax1*$^{-/-}$, exposing the influence of compensation by *Pax9*, which obscured the true targets of *Pax1* and *Pax9*. Through the use of 3-allele (*Pax1*$^{-/-}$*Pax9*$^{+/-}$) and 4-allele KO (*Pax1*$^{-/-}$*Pax9*$^{-/-}$) embryos, targets genes that had been masked by the redundant roles of *Pax9* have been uncovered in this study. By extension, these would also be the common set of target genes of *Pax1* and *Pax9*. There were six times more genes differentially expressed (599 genes) in the double-null embryos compared to *Pax1*$^{-/-}$. Moreover, through different combinations of comparisons of the multiple allele KOs, genes regulated by 2 copies of *Pax9* and 1 copy of *Pax9* in the absence of *Pax1* were identified, which in turn represent the target genes regulated by *Pax9* to compensate for the loss of *Pax1*. Thus, the various hypotheses proposed by prior groups regarding the functions of *Pax1* and *Pax9* were validated by utliziing the differential gene expression data – i.e. *Pax1* and *Pax9* regulate genes involved in cell proliferation, cell motion, cell adhesion and ECM genes involved in cartilage

development or mesenchymal condensation process. Furthermore, processes such as collagen fibril organization and blood vessel development, which were not anticipated by other reseachers in prior publications, were revealed in this study. Notably, 17 of the genes regulated by *Pax1/Pax9* are also associated with skeletal developmental abnormalities, some of which phenocopy *Pax1/Pax9*-deficient mutants.

In addition, it is through the temporal analysis of *Pax1* regulated targets in *Pax1*$^{-/-}$ mutants that interesting trends could be observed – the genes with opposite directionality, which could potentially be a consequence of compensation. That is, genes identified to be down-regulated in the E12.5 double-null were unchanged at E12.5 *Pax1*$^{-/-}$ but up-regulated in the E13.5 *Pax1*$^{-/-}$. Such observations reveal that gene regulation is dynamic and temporal analyses can unveil such intricate mechanisms of regulation.

Importantly, by the means of ChIP-Seq, the direct binding sites of Pax1 and Pax9 were identified in the WT vertebral column tissues. Overlapping of the TF mapping data with the differential gene expression data distinguished the direct and indirect targets of *Pax1* and *Pax9*. The direct targets of both *Pax1* and *Pax9* were also enriched for ECM, cartilage development, cell adhesion, cell motion, proliferation, ECM-receptor interaction and blood vessel development, indicating that these are the true molecular functions of *Pax1/Pax9*.

Interestingly, *Col2a1*, *Wwp2*, *Acan* and *Sox5* were among the direct targets of Pax1 and Pax9. Mutations/ knock-out in any one of these four key ECM genes result in vertebral column and facial abnormalities similar to the *Pax1/Pax9*-deficient mice. Notably, Sox9 is also known to regulate *Wwp2* and *Col2a1* directly, whereby the regulation of the latter involves the binding of the Sox9-Wwp2-Med25 complex at the intronic enhancer of *Col2a1*. Therefore, it is hypothesized from this study that Pax1

and Pax9 could be regulating *Col2a1* together with Sox9, potentially interacting with the Sox9-Wwp2-Med25 complex and several other intermediary proteins may be involved in this complex formation.

Also, *Sox5* is one of the crucial genes involved in the chondro-osteogenic pathway and serves as a key link between these *Pax* genes and osteo-chondrogenic pathway. Mining of the *Sox5/Sox6* and *Sox9* differential gene expression and TF mapping data generated by other researchers in the lab (unpublished data) showed that these *Sox* genes also regulate *Pax1* in return, but negatively, thus forming a negative feedback loop. It is known that *Pax1* and *Pax9* expression becomes down-regulated in the mature cartilage (i.e. upon chondrogenesis) while the *Sox* trio are essential for and are up-regulated during chondrogenesis. Based on these observations, it could be hypothesized that this negative feedback loop mechanism is how the *Pax1/Pax9*, which are initially uniformly expressed in the IVD analgen, subsequently become down-regulated in the cartilaginous inner annulus and are restricted to the fibrous outer annulus of the IVD. While only a hypothesis currently, it is an important point that warrants further investigation in the future.

Thus, the four specific aims that were put forth at the beginning of this study have been addressed. The enormous amount of data that has been generated in this study is a valuable resource that can be used to build the GRN of embryonic skeletal development. Constructing the bigger network from the various datasets is a time-consuming and a bioinformatics-intensive task. Hence, in this study, the focus has only been on the connections within a small circuit (*Sox* and the *Pax* genes). More importantly, the connection between the *Sox* genes and *Pax* genes identified in this study is novel. Indeed, this reiterates that there is still much to be learnt about the regulatory mechanisms involved in chondrogenesis and IVD development, and by

combining and mining such genome-wide data sets, more of such surprising connections could be delineated.

Moreover, the numerous mouse lines and the identification of genes enriched in *Pax1*- and *Pax9*- specific cells are all important resources for the scientific community. The endogenously tagged $Pax1^{HA3}$ and $Pax9^{HA3}$ mouse lines are invaluable for *in vivo* protein-protein interaction studies, TF mapping etc. These various mouse lines can also be used to study the other functions of the *Pax* genes such as in odontogenesis or thymus development.

It is imperative to understand though, that the identification of *Pax1* and *Pax9* targets in the IVD is only the beginning. How they compensate for each other at DNA binding level, who are their interacting partners, how they execute their pleiotrophic roles in different tissues are all important questions that deserve answers. Resolving those questions also requires a variety of other approaches to be undertaken. This study nevertheless serves as the starting point and certainly brings us closer to achieving our ultimate goal of constructing the gene regulatory network of embryonic skeletal development.

REFERENCES

1. *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
2. Metzker, M.L., *Sequencing technologies - the next generation.* Nature reviews. Genetics, 2010. **11**(1): p. 31-46.
3. Hobert, O., *Common logic of transcription factor and microRNA action.* Trends in biochemical sciences, 2004. **29**(9): p. 462-8.
4. Martinez, N.J. and A.J. Walhout, *The interplay between transcription factors and microRNAs in genome-scale regulatory networks.* BioEssays : news and reviews in molecular, cellular and developmental biology, 2009. **31**(4): p. 435-45.
5. Walhout, A.J., *Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping.* Genome research, 2006. **16**(12): p. 1445-54.
6. Taft, R.J., M. Pheasant, and J.S. Mattick, *The relationship between non-protein-coding DNA and eukaryotic complexity.* BioEssays : news and reviews in molecular, cellular and developmental biology, 2007. **29**(3): p. 288-99.
7. Bourque, G., et al., *Evolution of the mammalian transcription factor binding repertoire via transposable elements.* Genome research, 2008. **18**(11): p. 1752-62.
8. Polak, P. and E. Domany, *Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes.* BMC genomics, 2006. **7**: p. 133.
9. Polavarapu, N., et al., *Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA.* BMC genomics, 2008. **9**: p. 226.
10. Arnone, M.I. and E.H. Davidson, *The hardwiring of development: organization and function of genomic regulatory systems.* Development, 1997. **124**(10): p. 1851-64.
11. Ravasi, T., et al., *An atlas of combinatorial transcriptional regulation in mouse and man.* Cell, 2010. **140**(5): p. 744-52.
12. del Sol, A., et al., *Diseases as network perturbations.* Current opinion in biotechnology, 2010. **21**(4): p. 566-71.
13. Wilczynski, B. and E.E. Furlong, *Challenges for modeling global gene regulatory networks during development: insights from Drosophila.* Developmental biology, 2010. **340**(2): p. 161-9.
14. Davidson, E.H., et al., *A genomic regulatory network for development.* Science, 2002. **295**(5560): p. 1669-78.
15. Oliveri, P., Q. Tu, and E.H. Davidson, *Global regulatory logic for specification of an embryonic cell lineage.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(16): p. 5955-62.
16. Stathopoulos, A., et al., *Whole-genome analysis of dorsal-ventral patterning in the Drosophila embryo.* Cell, 2002. **111**(5): p. 687-701.
17. Stathopoulos, A. and M. Levine, *Genomic regulatory networks and animal development.* Developmental cell, 2005. **9**(4): p. 449-62.
18. Inoue, T., et al., *Transcriptional network underlying Caenorhabditis elegans vulval development.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(14): p. 4972-7.
19. Hobert, O., *Regulatory logic of neuronal diversity: terminal selector genes and selector motifs.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(51): p. 20067-71.
20. Koide, T., T. Hayata, and K.W. Cho, *Xenopus as a model system to study transcriptional regulatory networks.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(14): p. 4943-8.

21. Loose, M. and R. Patient, *A genetic regulatory network for Xenopus mesendoderm formation.* Developmental biology, 2004. **271**(2): p. 467-78.
22. Lefebvre, V. and P. Bhattaram, *Vertebrate skeletogenesis.* Current topics in developmental biology, 2010. **90**: p. 291-317.
23. Karsenty, G., H.M. Kronenberg, and C. Settembre, *Genetic control of bone formation.* Annual review of cell and developmental biology, 2009. **25**: p. 629-48.
24. Mackie, E.J., et al., *Endochondral ossification: how cartilage is converted into bone in the developing skeleton.* The international journal of biochemistry & cell biology, 2008. **40**(1): p. 46-62.
25. Hartmann, C., *Transcriptional networks controlling skeletal development.* Current opinion in genetics & development, 2009. **19**(5): p. 437-43.
26. Goyal, J.P., A. Gupta, and V.B. Shah, *Campomelic dysplasia.* Indian journal of human genetics, 2011. **17**(3): p. 247-8.
27. Mansour, S., et al., *A clinical and genetic study of campomelic dysplasia.* Journal of medical genetics, 1995. **32**(6): p. 415-20.
28. Karsenty, G., *Transcriptional control of skeletogenesis.* Annual review of genomics and human genetics, 2008. **9**: p. 183-96.
29. P., P.R., *Intervertebral Disc: Anatomy-Physiology-Patthophysiology-Treatment.* Pain practice : the official journal of World Institute of Pain, 2009. **8**(1): p. 18-44.
30. Risbud, M.V. and I.M. Shapiro, *Notochordal cells in the adult intervertebral disc: new perspective on an old question.* Critical reviews in eukaryotic gene expression, 2011. **21**(1): p. 29-41.
31. Gilbert, S.F., *Paraxial and Intermediate Mesoderm*, in *Developmental Biology*, C. Wigg, Editor 2006, Andrew D. Sinauer: Massachusetts. p. 443-455.
32. Burgess, R., et al., *Requirement of the paraxis gene for somite formation and musculoskeletal patterning.* Nature, 1996. **384**(6609): p. 570-3.
33. Weinmaster, G. and C. Kintner, *Modulation of notch signaling during somitogenesis.* Annual review of cell and developmental biology, 2003. **19**: p. 367-95.
34. Lewis, J., A. Hanisch, and M. Holder, *Notch signaling, the segmentation clock, and the patterning of vertebrate somites.* Journal of biology, 2009. **8**(4): p. 44.
35. Fan, C.M. and M. Tessier-Lavigne, *Patterning of mammalian somites by surface ectoderm and notochord: evidence for sclerotome induction by a hedgehog homolog.* Cell, 1994. **79**(7): p. 1175-86.
36. Cairns, D.M., et al., *A gradient of Shh establishes mutually repressing somitic cell fates induced by Nkx3.2 and Pax3.* Developmental biology, 2008. **323**(2): p. 152-65.
37. Monsoro-Burq, A.H., et al., *The role of bone morphogenetic proteins in vertebral development.* Development, 1996. **122**(11): p. 3607-16.
38. Johnson, R.L., et al., *Ectopic expression of Sonic hedgehog alters dorsal-ventral patterning of somites.* Cell, 1994. **79**(7): p. 1165-73.
39. Koseki, H., et al., *A role for Pax-1 as a mediator of notochordal signals during the dorsoventral specification of vertebrae.* Development, 1993. **119**(3): p. 649-60.
40. Dietrich, S., F.R. Schubert, and P. Gruss, *Altered Pax gene expression in murine notochord mutants: the notochord is required to initiate and maintain ventral identity in the somite.* Mechanisms of development, 1993. **44**(2-3): p. 189-207.
41. Furumoto, T.A., et al., *Notochord-dependent expression of MFH1 and PAX1 cooperates to maintain the proliferation of sclerotome cells during the vertebral column development.* Developmental biology, 1999. **210**(1): p. 15-29.
42. McMahon, J.A., et al., *Noggin-mediated antagonism of BMP signaling is required for growth and patterning of the neural tube and somite.* Genes & development, 1998. **12**(10): p. 1438-52.

43. Zhang, X.M., M. Ramalho-Santos, and A.P. McMahon, *Smoothened mutants reveal redundant roles for Shh and Ihh signaling including regulation of L/R symmetry by the mouse node.* Cell, 2001. **106**(2): p. 781-92.
44. Buttitta, L., et al., *Interplays of Gli2 and Gli3 and their requirement in mediating Shh-dependent sclerotome induction.* Development, 2003. **130**(25): p. 6233-43.
45. Stafford, D.A., et al., *Cooperative activity of noggin and gremlin 1 in axial skeleton development.* Development, 2011. **138**(5): p. 1005-14.
46. Murtaugh, L.C., J.H. Chyung, and A.B. Lassar, *Sonic hedgehog promotes somitic chondrogenesis by altering the cellular response to BMP signaling.* Genes & development, 1999. **13**(2): p. 225-37.
47. Chiang, C., et al., *Cyclopia and defective axial patterning in mice lacking Sonic hedgehog gene function.* Nature, 1996. **383**(6599): p. 407-13.
48. Christ, B. and J. Wilting, *From somites to vertebral column.* Annals of anatomy = Anatomischer Anzeiger : official organ of the Anatomische Gesellschaft, 1992. **174**(1): p. 23-32.
49. Brand-Saberi, B. and B. Christ, *Evolution and development of distinct cell lineages derived from somites.* Current topics in developmental biology, 2000. **48**: p. 1-42.
50. Neubuser, A., H. Koseki, and R. Balling, *Characterization and developmental expression of Pax9, a paired-box-containing gene related to Pax1.* Developmental biology, 1995. **170**(2): p. 701-16.
51. Mankoo, B.S., et al., *The concerted action of Meox homeobox genes is required upstream of genetic pathways essential for the formation, patterning and differentiation of somites.* Development, 2003. **130**(19): p. 4655-64.
52. Bi, W., et al., *Sox9 is required for cartilage formation.* Nature genetics, 1999. **22**(1): p. 85-9.
53. Tribioli, C. and T. Lufkin, *The murine Bapx1 homeobox gene plays a critical role in embryonic development of the axial skeleton and spleen.* Development, 1999. **126**(24): p. 5699-711.
54. Yamashita, S., et al., *Sox9 directly promotes Bapx1 gene expression to repress Runx2 in chondrocytes.* Experimental cell research, 2009. **315**(13): p. 2231-40.
55. Zeng, L., et al., *Shh establishes an Nkx3.2/Sox9 autoregulatory loop that is maintained by BMP signals to induce somitic chondrogenesis.* Genes & development, 2002. **16**(15): p. 1990-2005.
56. Rodrigo, I., et al., *Meox homeodomain proteins are required for Bapx1 expression in the sclerotome and activate its transcription by direct binding to its promoter.* Molecular and cellular biology, 2004. **24**(7): p. 2757-66.
57. Rodrigo, I., et al., *Pax1 and Pax9 activate Bapx1 to induce chondrogenic differentiation in the sclerotome.* Development, 2003. **130**(3): p. 473-82.
58. Stamataki, D., et al., *Homeodomain proteins Mox1 and Mox2 associate with Pax1 and Pax3 transcription factors.* FEBS letters, 2001. **499**(3): p. 274-8.
59. Winnier, G.E., L. Hargett, and B.L. Hogan, *The winged helix transcription factor MFH1 is required for proliferation and patterning of paraxial mesoderm in the mouse embryo.* Genes & development, 1997. **11**(7): p. 926-40.
60. Sohn, P., et al., *Molecular profiling of the developing mouse axial skeleton: a role for Tgfbr2 in the development of the intervertebral disc.* BMC developmental biology, 2010. **10**: p. 29.
61. Baffi, M.O., et al., *Conditional deletion of the TGF-beta type II receptor in Col2a expressing cells results in defects in the axial skeleton without alterations in chondrocyte differentiation or embryonic development of long bones.* Developmental biology, 2004. **276**(1): p. 124-42.

62. Peters, H., et al., *Pax1 and Pax9 synergistically regulate vertebral column development.* Development, 1999. **126**(23): p. 5399-408.
63. DiPaola, C.P., et al., *Molecular signaling in intervertebral disk development.* Journal of orthopaedic research : official publication of the Orthopaedic Research Society, 2005. **23**(5): p. 1112-9.
64. Balling, R., et al., *Pax genes and skeletal development.* Annals of the New York Academy of Sciences, 1996. **785**: p. 27-33.
65. Walther, C., et al., *Pax: a murine multigene family of paired box-containing genes.* Genomics, 1991. **11**(2): p. 424-34.
66. Bopp, D., et al., *Conservation of a large protein domain in the segmentation gene paired and in functionally related genes of Drosophila.* Cell, 1986. **47**(6): p. 1033-40.
67. Jun, S. and C. Desplan, *Cooperative interactions between paired domain and homeodomain.* Development, 1996. **122**(9): p. 2639-50.
68. Chalepakis, G., et al., *The molecular basis of the undulated/Pax-1 mutation.* Cell, 1991. **66**(5): p. 873-84.
69. Deutsch, U., G.R. Dressler, and P. Gruss, *Pax 1, a member of a paired box homologous murine gene family, is expressed in segmented structures during development.* Cell, 1988. **53**(4): p. 617-25.
70. Capellini, T.D., et al., *Pbx1/Pbx2 govern axial skeletal development by controlling Polycomb and Hox in mesoderm and Pax1/Pax9 in sclerotome.* Developmental biology, 2008. **321**(2): p. 500-14.
71. Wallin, J., et al., *The role of Pax-1 in axial skeleton development.* Development, 1994. **120**(5): p. 1109-21.
72. Wallin, J., et al., *Pax1 is expressed during development of the thymus epithelium and is required for normal T-cell maturation.* Development, 1996. **122**(1): p. 23-30.
73. Wilm, B., et al., *Targeted disruption of Pax1 defines its null phenotype and proves haploinsufficiency.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(15): p. 8692-7.
74. Peters, H., et al., *Pax9-deficient mice lack pharyngeal pouch derivatives and teeth and exhibit craniofacial and limb abnormalities.* Genes & development, 1998. **12**(17): p. 2735-47.
75. Kist, R., E. Greally, and H. Peters, *Derivation of a mouse model for conditional inactivation of Pax9.* Genesis, 2007. **45**(7): p. 460-4.
76. Timmons, P.M., et al., *Expression and function of Pax 1 during development of the pectoral girdle.* Development, 1994. **120**(10): p. 2773-85.
77. Balling, R., U. Deutsch, and P. Gruss, *undulated, a mutation affecting the development of the mouse skeleton, has a point mutation in the paired box of Pax 1.* Cell, 1988. **55**(3): p. 531-5.
78. Blandova, Y.R. and I.U. Egorov, *Sut allelic with un.* Mouse News Lett., 1975. **52**(43).
79. Wallace, M.E., *An inherited agent of mutation with chromosome damage in wild mice.* The Journal of heredity, 1985. **76**(4): p. 271-8.
80. Adham, I.M., et al., *The scoliosis (sco) mouse: a new allele of Pax1.* Cytogenetic and genome research, 2005. **111**(1): p. 16-26.
81. Wakatsuki, Y., et al., *The B cell-specific transcription factor BSAP regulates B cell proliferation.* The Journal of experimental medicine, 1994. **179**(4): p. 1099-108.
82. Warren, N. and D.J. Price, *Roles of Pax-6 in murine diencephalic development.* Development, 1997. **124**(8): p. 1573-82.
83. Bannykh, S.I., et al., *Aberrant Pax1 and Pax9 expression in Jarcho-Levin syndrome: report of two Caucasian siblings and literature review.* American journal of medical genetics. Part A, 2003. **120A**(2): p. 241-6.

84. McGaughran, J.M., et al., *Mutations in PAX1 may be associated with Klippel-Feil syndrome.* European journal of human genetics : EJHG, 2003. **11**(6): p. 468-74.
85. Giampietro, P.F., et al., *An analysis of PAX1 in the development of vertebral malformations.* Clinical genetics, 2005. **68**(5): p. 448-53.
86. Lopez, B.C., K.M. David, and H.A. Crockard, *Inadequate PAX-1 gene expression as a cause of agenesis of the thoracolumbar spine with failure of segmentation. Case report.* Journal of neurosurgery, 1997. **86**(6): p. 1018-21.
87. Satija, N.K., et al., *Mesenchymal stem cells: molecular targets for tissue engineering.* Stem cells and development, 2007. **16**(1): p. 7-23.
88. Kraus, P., et al., *A more cost effective and rapid high percentage germ-line transmitting chimeric mouse generation procedure via microinjection of 2-cell, 4-cell, and 8-cell embryos with ES and iPS cells.* Genesis, 2010. **48**(6): p. 394-9.
89. Day, R., Davidson MW, *The fluorescent protein palette: tools for cellular imaging.* Chem Soc Rev., 2009. **38**: p. 2887-2921.
90. De Felipe, P., *Polycistronic viral vectors.* Curr Gene Ther, 2002. **2**: p. 355-378.
91. Hellen, C., Sarnow P, *Internal ribosome entry sites in eurkaryotic mRNA molecules.* Genes Dev., 2001. **15**: p. 1593-1612.
92. De Felipe, P., *Skipping the co-expression problem: the new 2A "CHYSEL" technology.* Genetic Vaccines and Therapy, 2004. **2**: p. 13.
93. Donnelly, M., Hughes LE, Luke G, Mendoza H, Dam E, et al, *The 'cleavage' activities of foot-and-mouth disease virus 2A site-directed mutants and naturally occurring '2A-like' sequences.* J Gen Virol., 2001. **82**: p. 1027-1041.
94. Chan, H.Y., et al., *Comparison of IRES and F2A-based locus-specific multicistronic expression in stable mouse lines.* PloS one, 2011. **6**(12): p. e28885.
95. Fang, J., et al., *Stable antibody expression at therapeutic levels using the 2A peptide.* Nature biotechnology, 2005. **23**(5): p. 584-90.
96. Lorens, J.B., et al., *Stable, stoichiometric delivery of diverse protein functions.* Journal of biochemical and biophysical methods, 2004. **58**(2): p. 101-10.
97. Surdej, P. and M. Jacobs-Lorena, *Strategy for epitope tagging the protein-coding region of any gene.* BioTechniques, 1994. **17**(3): p. 560-5.
98. Wilson, I.A., et al., *The structure of an antigenic determinant in a protein.* Cell, 1984. **37**(3): p. 767-78.
99. Xie, Z., et al., *Systematic characterization of protein-DNA interactions.* Cellular and molecular life sciences : CMLS, 2011. **68**(10): p. 1657-68.
100. Kolodziej, P.A. and R.A. Young, *Epitope tagging and protein surveillance.* Methods in enzymology, 1991. **194**: p. 508-19.
101. Wang, Z., *Epitope tagging of endogenous proteins for genome-wide chromatin immunoprecipitation analysis.* Methods in molecular biology, 2009. **567**: p. 87-98.
102. Knop, M., et al., *Epitope tagging of yeast genes using a PCR-based strategy: more tags and improved practical routines.* Yeast, 1999. **15**(10B): p. 963-72.
103. Shevchenko, A., et al., *Chromatin Central: towards the comparative proteome by accurate mapping of the yeast proteomic environment.* Genome biology, 2008. **9**(11): p. R167.
104. Krogan, N.J., et al., *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.* Nature, 2006. **440**(7084): p. 637-43.
105. Hofemeister, H., et al., *Recombineering, transfection, Western, IP and ChIP methods for protein tagging via gene targeting or BAC transgenesis.* Methods, 2011. **53**(4): p. 437-52.
106. Robertson, D., et al., *Ultrastructural localization of ras-related proteins using epitope-tagged plasmids.* The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society, 1995. **43**(5): p. 471-80.

107. Trogadis, J.E., et al., *Dopamine D1 receptor distribution in Sf9 cells imaged by confocal microscopy: a quantitative evaluation.* The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society, 1995. **43**(5): p. 497-506.
108. Ishii, K., et al., *Possible domains responsible for intracellular targeting and insulin-dependent translocation of glucose transporter type 4.* The Biochemical journal, 1995. **309 (Pt 3)**: p. 813-23.
109. Pines, J. and T. Hunter, *Human cyclins A and B1 are differentially located in the cell and undergo cell cycle-dependent nuclear transport.* The Journal of cell biology, 1991. **115**(1): p. 1-17.
110. Mummery, C.L., *Transforming growth factor beta and mouse development.* Microscopy research and technique, 2001. **52**(4): p. 374-86.
111. Cao, X. and D. Chen, *The BMP signaling and in vivo bone formation.* Gene, 2005. **357**(1): p. 1-8.
112. Chen, G., C. Deng, and Y.P. Li, *TGF-beta and BMP signaling in osteoblast differentiation and bone formation.* International journal of biological sciences, 2012. **8**(2): p. 272-88.
113. Ingham, P.W. and A.P. McMahon, *Hedgehog signaling in animal development: paradigms and principles.* Genes & development, 2001. **15**(23): p. 3059-87.
114. Ling, L., V. Nurcombe, and S.M. Cool, *Wnt signaling controls the fate of mesenchymal stem cells.* Gene, 2009. **433**(1-2): p. 1-7.
115. Hartmann, C., *A Wnt canon orchestrating osteoblastogenesis.* Trends in cell biology, 2006. **16**(3): p. 151-8.
116. Hall, B.K. and T. Miyake, *All for one and one for all: condensations and the initiation of skeletal development.* BioEssays : news and reviews in molecular, cellular and developmental biology, 2000. **22**(2): p. 138-47.
117. DeLise, A.M., L. Fischer, and R.S. Tuan, *Cellular interactions and signaling in cartilage development.* Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society, 2000. **8**(5): p. 309-34.
118. Hall, B.K. and T. Miyake, *Divide, accumulate, differentiate: cell condensation in skeletal development revisited.* The International journal of developmental biology, 1995. **39**(6): p. 881-93.
119. Smits, P. and V. Lefebvre, *Sox5 and Sox6 are required for notochord extracellular matrix sheath formation, notochord cell survival and development of the nucleus pulposus of intervertebral discs.* Development, 2003. **130**(6): p. 1135-48.
120. Lefebvre, V., et al., *SOX9 is a potent activator of the chondrocyte-specific enhancer of the pro alpha1(II) collagen gene.* Molecular and cellular biology, 1997. **17**(4): p. 2336-46.
121. Bell, D.M., et al., *SOX9 directly regulates the type-II collagen gene.* Nature genetics, 1997. **16**(2): p. 174-8.
122. Sekiya, I., et al., *SOX9 enhances aggrecan gene promoter/enhancer activity and is up-regulated by retinoic acid in a cartilage-derived cell line, TC6.* The Journal of biological chemistry, 2000. **275**(15): p. 10738-44.
123. Rentsendorj, O., et al., *Highly conserved proximal promoter element harbouring paired Sox9-binding sites contributes to the tissue- and developmental stage-specific activity of the matrilin-1 gene.* The Biochemical journal, 2005. **389**(Pt 3): p. 705-16.
124. Nakamura, Y., et al., *Wwp2 is essential for palatogenesis mediated by the interaction between Sox9 and mediator subunit 25.* Nature communications, 2011. **2**: p. 251.
125. Akiyama, H., et al., *The transcription factor Sox9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of Sox5 and Sox6.* Genes & development, 2002. **16**(21): p. 2813-28.

126. Hayes, A.J., et al., *Collagen fibrillogenesis in the development of the annulus fibrosus of the intervertebral disc.* European cells & materials, 2011. **22**: p. 226-41.
127. Aszodi, A., et al., *Collagen II is essential for the removal of the notochord and the formation of intervertebral discs.* The Journal of cell biology, 1998. **143**(5): p. 1399-412.
128. Watanabe, H., et al., *Dwarfism and age-associated spinal degeneration of heterozygote cmd mice defective in aggrecan.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(13): p. 6943-7.
129. T.Nakane, et al., *Dysspondyloenchondromatosis: Another COL2A1-Related Skeletal Dysplasia?* Molecular Syndromology, 2011. **1**: p. 21 - 26.
130. Li, Y., et al., *A fibrillar collagen gene, Col11a1, is essential for skeletal morphogenesis.* Cell, 1995. **80**(3): p. 423-30.
131. Gleghorn, L., et al., *A mutation in the variable repeat region of the aggrecan gene (AGC1) causes a form of spondyloepiphyseal dysplasia associated with severe, premature osteoarthritis.* American journal of human genetics, 2005. **77**(3): p. 484-90.
132. Tompson, S.W., et al., *A recessive skeletal dysplasia, SEMD aggrecan type, results from a missense mutation affecting the C-type lectin domain of aggrecan.* American journal of human genetics, 2009. **84**(1): p. 72-9.
133. Mio, F., et al., *A functional polymorphism in COL11A1, which encodes the alpha 1 chain of type XI collagen, is associated with susceptibility to lumbar disc herniation.* American journal of human genetics, 2007. **81**(6): p. 1271-7.
134. Noponen-Hietala, N., et al., *Sequence variations in the collagen IX and XI genes are associated with degenerative lumbar spinal stenosis.* Annals of the rheumatic diseases, 2003. **62**(12): p. 1208-14.
135. Solovieva, S., et al., *Association between the aggrecan gene variable number of tandem repeats polymorphism and intervertebral disc degeneration.* Spine, 2007. **32**(16): p. 1700-5.
136. Shee, E.Y.-S., S.L. Toh, and J.C.-H. Goh, *Effects of radial compression on a novel simulated intervertebral disc-like assembly using bone marrow-derived mesenchymal stem cell cell-sheets for annulus fibrosus regeneration.* . Spine, 2011. **36**(21): p. 1744–1751.
137. Zhang, Y., et al., *Intervertebral disk repair by protein, gene, or cell injection: a framework for rehabilitation-focused biologics in the spine.* PM & R : the journal of injury, function, and rehabilitation, 2011. **3**(6 Suppl 1): p. S88-94.
138. Han, Y. and V. Lefebvre, *L-Sox5 and Sox6 drive expression of the aggrecan gene in cartilage by securing binding of Sox9 to a far-upstream enhancer.* Molecular and cellular biology, 2008. **28**(16): p. 4999-5013.
139. Lefebvre, V., P. Li, and B. de Crombrugghe, *A new long form of Sox5 (L-Sox5), Sox6 and Sox9 are coexpressed in chondrogenesis and cooperatively activate the type II collagen gene.* The EMBO journal, 1998. **17**(19): p. 5718-33.
140. Smits, P., et al., *The transcription factors L-Sox5 and Sox6 are essential for cartilage formation.* Developmental cell, 2001. **1**(2): p. 277-90.
141. Tsumaki, N., et al., *Bone morphogenetic protein signals are required for cartilage formation and differently regulate joint development during skeletogenesis.* Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research, 2002. **17**(5): p. 898-906.
142. Shum, L., et al., *BMP4 promotes chondrocyte proliferation and hypertrophy in the endochondral cranial base.* The International journal of developmental biology, 2003. **47**(6): p. 423-31.

143. Gomes, R.R., Jr., M.C. Farach-Carson, and D.D. Carson, *Perlecan functions in chondrogenesis: insights from in vitro and in vivo models.* Cells, tissues, organs, 2004. **176**(1-3): p. 79-86.
144. Arikawa-Hirasawa, E., et al., *Perlecan is essential for cartilage and cephalic development.* Nature genetics, 1999. **23**(3): p. 354-8.
145. Costell, M., et al., *Perlecan maintains the integrity of cartilage and some basement membranes.* The Journal of cell biology, 1999. **147**(5): p. 1109-22.
146. Choocheep, K., et al., *Versican facilitates chondrocyte differentiation and regulates joint morphogenesis.* The Journal of biological chemistry, 2010. **285**(27): p. 21114-25.
147. Arnott, J.A., et al., *The role of connective tissue growth factor (CTGF/CCN2) in skeletogenesis.* Critical reviews in eukaryotic gene expression, 2011. **21**(1): p. 43-69.
148. Abreu, J.G., et al., *Connective-tissue growth factor (CTGF) modulates cell signalling by BMP and TGF-beta.* Nature cell biology, 2002. **4**(8): p. 599-604.
149. Frazier, K., et al., *Stimulation of fibroblast cell growth, matrix production, and granulation tissue formation by connective tissue growth factor.* The Journal of investigative dermatology, 1996. **107**(3): p. 404-11.
150. Song, J.J., et al., *Connective tissue growth factor (CTGF) acts as a downstream mediator of TGF-beta1 to induce mesenchymal cell condensation.* Journal of cellular physiology, 2007. **210**(2): p. 398-410.
151. Nguyen, T.Q., et al., *CTGF inhibits BMP-7 signaling in diabetic nephropathy.* Journal of the American Society of Nephrology : JASN, 2008. **19**(11): p. 2098-107.
152. Ivkovic, S., et al., *Connective tissue growth factor coordinates chondrogenesis and angiogenesis during skeletal development.* Development, 2003. **130**(12): p. 2779-91.
153. Gilbert, S.F., *Developmental Biology, Eigth Edition* 2006, Sunderland, MA: Sinauer Associates Inc.
154. Shen, J., et al., *Skeletal and CNS defects in Presenilin-1-deficient mice.* Cell, 1997. **89**(4): p. 629-39.
155. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome biology, 2008. **9**(9): p. R137.
156. Lettice, L.A., et al., *A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly.* Human molecular genetics, 2003. **12**(14): p. 1725-35.
157. Wunderle, V.M., et al., *Deletion of long-range regulatory elements upstream of SOX9 causes campomelic dysplasia.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(18): p. 10649-54.
158. Gordon, C.T., et al., *Long-range regulation at the SOX9 locus in development and disease.* Journal of medical genetics, 2009. **46**(10): p. 649-56.
159. Bartkuhn, M. and R. Renkawitz, *Long range chromatin interactions involved in gene regulation.* Biochimica et biophysica acta, 2008. **1783**(11): p. 2161-6.
160. Czerny, T., G. Schaffner, and M. Busslinger, *DNA sequence recognition by Pax proteins: bipartite structure of the paired domain and its binding site.* Genes & development, 1993. **7**(10): p. 2048-61.
161. McLean, C.Y., et al., *GREAT improves functional interpretation of cis-regulatory regions.* Nature biotechnology, 2010. **28**(5): p. 495-501.
162. Sahlman, J., et al., *Premature vertebral endplate ossification and mild disc degeneration in mice after inactivation of one allele belonging to the Col2a1 gene for Type II collagen.* Spine, 2001. **26**(23): p. 2558-65.
163. Watanabe, H., et al., *Mouse cartilage matrix deficiency (cmd) caused by a 7 bp deletion in the aggrecan gene.* Nature genetics, 1994. **7**(2): p. 154-7.

164. Pereira, L., et al., *Pathogenetic sequence for aneurysm revealed in mice underexpressing fibrillin-1.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(7): p. 3819-23.
165. Ades, L.C., et al., *Segregation of a novel FBN1 gene mutation, G1796E, with kyphoscoliosis and radiographic evidence of vertebral dysplasia in three generations.* American journal of medical genetics, 2002. **109**(4): p. 261-70.
166. Davis, J.A., et al., *An Alzheimer's disease-linked PS1 variant rescues the developmental abnormalities of PS1-deficient embryos.* Neuron, 1998. **20**(3): p. 603-9.
167. Zou, W., et al., *The E3 ubiquitin ligase Wwp2 regulates craniofacial development through mono-ubiquitylation of Goosecoid.* Nature cell biology, 2011. **13**(1): p. 59-65.
168. Kurima, K., et al., *A member of a family of sulfate-activating enzymes causes murine brachymorphism.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(15): p. 8681-5.
169. Miyake, N., et al., *PAPSS2 mutations cause autosomal recessive brachyolmia.* Journal of medical genetics, 2012.
170. Shamblott, M.J., et al., *Craniofacial abnormalities resulting from targeted disruption of the murine Sim2 gene.* Developmental dynamics : an official publication of the American Association of Anatomists, 2002. **224**(4): p. 373-80.
171. Oliver, E.R., et al., *Ribosomal protein L24 defect in belly spot and tail (Bst), a mouse Minute.* Development, 2004. **131**(16): p. 3907-20.
172. Horan, G.S., et al., *Mutations in paralogous Hox genes result in overlapping homeotic transformations of the axial skeleton: evidence for unique and redundant function.* Developmental biology, 1995. **169**(1): p. 359-72.
173. Suemori, H., N. Takahashi, and S. Noguchi, *Hoxc-9 mutant mice show anterior transformation of the vertebrae and malformation of the sternum and ribs.* Mechanisms of development, 1995. **51**(2-3): p. 265-73.
174. McIntyre, D.C., et al., *Hox patterning of the vertebrate rib cage.* Development, 2007. **134**(16): p. 2981-9.
175. Li, X., et al., *Mammalian polycomb-like Pcl2/Mtf2 is a novel regulatory component of PRC2 that can differentially modulate polycomb activity both at the Hox gene cluster and at Cdkn2a genes.* Molecular and cellular biology, 2011. **31**(2): p. 351-64.
176. Barbaric, I., et al., *An ENU-induced mutation in the Ankrd11 gene results in an osteopenia-like phenotype in the mouse mutant Yoda.* Physiological genomics, 2008. **32**(3): p. 311-21.
177. Andrikopoulos, K., et al., *Targeted mutation in the col5a2 gene reveals a regulatory role for type V collagen during matrix assembly.* Nature genetics, 1995. **9**(1): p. 31-6.
178. Zhou, G., et al., *A 182 bp fragment of the mouse pro alpha 1(II) collagen gene is sufficient to direct chondrocyte expression in transgenic mice.* Journal of cell science, 1995. **108 (Pt 12)**: p. 3677-84.
179. Bi, W., et al., *Haploinsufficiency of Sox9 results in defective cartilage primordia and premature skeletal mineralization.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(12): p. 6698-703.
180. Macneil, L.T. and A.J. Walhout, *Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression.* Genome research, 2011. **21**(5): p. 645-57.