Development of a Resource for the Exploration of Gene Expression in the Mouse Foetus

Elizabeth Ann Campbell CBiol. MIBiol.

A dissertation submitted to the Open University
for the degree of Master of Philosophy

January 2004

Sponsoring and collaborating establishments:

The Sanger Institute,
Wellcome Trust Genome Campus,
Hinxton,
Cambridgeshire,
CB10 1SA

Anatomy Department,
Cambridge University,
Downing Site,
Cambridge,
CB2 3DY

Funded by Parke Davis and The Wellcome Trust

CONTENTS

Page Number

CHAPTER 1 Construction of the Panel

CHAPTER 2 Profiling the Family of Sox Genes by PCR

CHAPTER 3 Validation of the Sox Genes PCR data by *in-situ* hybridisation

CHAPTER 4 Discussion

Abbreviations

European Molecular Biology Laboratory	EMBL
Expressed sequence tag	EST
sequence tagged site	sts
bacterial artificial chromosome	bac
days post conception	dpc
final concentration	*fc*
catalogue number	#
molar	M
litre	L
millilitre	ml
microlitre	μl
milligram	mg
microgram	μg
nanogram	ng
hour	hr
minutes	mins
seconds	secs
revolutions per minute	rpm
hydrochloric acid	HCl
diethylpyrocarbonate	depc
ethylene diamine tetraacetic acid	EDTA

trizma base	Tris
1 M Tris 0.1 M EDTA	T0.1E
tris borate EDTA buffer	TBE
ethidium bromide	EtBr
deoxyribonuclease 1	Dnase1
moloney murine leukemia virus reverse transcriptase	M-MLV
terminal deoxynucleotidyl transferase	TdT
2’-deoxyribonucleoside 5’-triphosphates	dNTPs
dithiothreitol	DTT
reverse transcription polymerase chain reaction	RTPCR
reverse transcription	RT
complementary deoxynucleic acid	cDNA
ribonucleic acid	RNA
sodium dodecyl sulphate	SDS
sodium chloride	NaCl
bovine serum albumin	BSA
disodium hydrogen orthophosphate	Na_2HPO_4
sodium dihydrogen orthophosphate	NaH_2PO_4
poly acrylamide gel electrophoresis	PAGE

Acknowledgements

I am indebted to Tom Freeman for giving me the impetus to embark on this project. From the Sanger Institute, I would like to thank the many scientists for sharing the fun of science, in particular Kate Rice, Adam Butler and Sarah Hunt for helping to me to understand bioinformatics and Dave Vetrie for his meticulous proof reading.

Of the scientists at Parke Davis, I would like to thank Rob Pinnock, Peter Cox, Kevin Lee and Alistair Dixon for all their support whilst working in their laboratory.

And for the continued and persistent encouragement throughout the course of this work, I owe Jane Rodgers a huge thank you.

I would like to thank Martin Johnson, Anne Ferguson-Smith, Christy Starr and Martin George for supplying the materials and expertise to pursue this study and Peter Wooding for assisting me with the technique of *in-situ* hybridisation. I would also like to thank John Bashford, Adrian Newman and Ian Bolton of the AVMG, for assistance with computing, microscopy and printing of this thesis. And I especially would like to thank Josie McConnell for the varied and valuable discussions.

I am particularly grateful to Martin Johnson for his gentle but untiring encouragement and guidance during this project and for helping me to order my thoughts sufficiently for writing.

And finally I would like to thank two of the most important men on my life, my husband, John and my son, Simon for their patience, humour and support throughout the duration of this work.

Abstract

Seventy five different foetal and placental tissues/organs, from naturally mated C57BL/6J mice, were dissected at 8.5d, 9.5d, 10.5d, 11.5d, 12.5d, 13.5d, 15.5d and17.5 days post-coitum. Tissues were stored at –80^{o}C and ribonucleic acid (RNA) extracted. Good quality RNA, as assessed by OD and agarose gel electrophoresis, was treated with Dnase1 and 50µg was reverse transcribed. The resulting complementary desoxyribonucliec acid (cDNA) was checked for the presence of genomic DNA, and amounts of cDNA adjusted so that equivalent amounts of PCR product from a set of two presumptive housekeeping genes were found in each of the 89 samples. The final panel consisted of cDNA from 75 mouse foetal tissues, 11 adult mouse tissues, 2 glycogen samples and rat, human and mouse genomic DNA, stored in a deep well microtitre plate.

Primers designed to 7 genes with a known expression pattern in adult tissues were used in the polymerase chain reaction to validate the integrity of the panel contents. The usefulness of the panel was tested by running PCR reactions with primers designed to the newly emerging family of *Sox* genes. Expression profiles for 17 members of the *Sox* family were obtained. PCR expression patterns of two of the *Sox* genes (*Sox2* and *6*) were compared with profiles obtained by *in-situ* hybridisation on staged sections of the mouse foetus. A reflection on the process of the panel formation and testing has highlighted a number of refinements to the process of exploring gene expression in fetal and placental tissues.

INTRODUCTION

From gene sequence to gene expression

The physical map of the mouse genomic sequence has been published [3], the basic descriptive human genome project is now completed [4], and the analysis of variation in both genomes is underway [5]. The next big challenge is to locate regions of gene expression and to understand how their expression is controlled and what their expression means mechanistically for development. During foetal developmental, certain genes and pathways are activated in a defined pattern and sequence for a fully functioning organism to develop. Many genes are known to be important during embryogenesis, because, when they are inactivated or only present on one chromosome, developmental pathology can result [6]. Mapping genes that are expressed in a particular tissue at a given time point will help in our understanding of gene hierarchy and function. Moreover, the precise form of gene expression observed is important, as evidenced by gene isoforms *Sox17* [7], *Plectin* [8] and *WT1* [9] performing different functions. Thus, observing either the expression pattern or the sequence information in isolation can be misleading. Genes switched on during development are often developmentally specific, as seen with the Cystic Fibrosis Transmembrane conductance Regulator (*CFTR*) gene [10], but the same gene can also be expressed later in development, often in a variant form through splicing, initiation or post-translational differences. Having an expression map will provide a useful database for intelligently manipulating development, so that cause and effect can be explored.

A model organism

The very early stages of development are comparable in all eutherian mammalian species and events that subsequently occur have many similarities between the species [11]. Beginning with the fusion of the male and female gametes, through pre-implantation and post-implantation stages of embryogenesis to the formation of the embryo and then the foetus, the construction of organs and tissues seems to be broadly conserved at molecular and organisational levels. The mouse is an ideal organism to study the expression of genes during foetal development, as it has been the organism of choice for many genetic, teratogenic, manipulative and developmental studies. Thus, information on expression patterns in the mouse provides a useful guide to patterns in other mammalian species.

Ways to map gene expression patterns – strengths and weaknesses

There are many different ways to study gene expression, including the techniques of northern blotting, nuclease S1 mapping, the ribonuclease protection assay, RTPCR (Reverse Transcription Polymerase Chain Reaction), *in-situ* hybridisation, computational studies, and more recently microarray technologies. Each of these will be described in turn, together with their strengths and weaknesses.

The northern blot has traditionally been applied to measure the total size/mobility of a transcript. It is thus useful for detecting variants due to differential splicing and initiation sites. It can also be used to screen for family gene members from related organisms, by varying the temperature of hybridisation. The northern blot detects relatively high abundance messages like □ actin, which make up as much as 0.1% (300pg from 5mg tissue) of the mass of total RNA whereas specific rarer messages exist at levels below 0.001% (10fg – 10pg from 5mg tissue) of the total mass and can be more difficult to distinguish. With care, the northern blot can detect levels as low as 10 pg (picogram) amounts. RNA samples to be studied are first separated on an agarose gel under denaturing conditions, and then transferred to a filter and immobilized. The filter is then hybridized with a labelled probe (usually cDNA - complementary Deoxyribo Nucleci Acid - or RNA) and specific targets are identified by autoradiography or nonradioactive methods such as digoxigenin. The main advantage is that the filter can be screened repeatedly under a variety of stringencies with different probes. Sequences with only partial homology (cDNA from different species or genomic DNA fragments that contain an intron) can be used as probes to identify transcript size and recognize alternatively spliced transcripts. However, the technique requires very high quality, full length RNA at the high concentration of 1mg/ml. Poor quality RNA will result in loss of signal, whilst contributing to the background 'noise', for example, a single nick in 20% of a 4 Kb transcript will reduce the resulting signal by a full 20% (http://www.ambion.com/techlib/basics/northerns/index.html).

Nuclease S1 mapping can be used to identify particular sequences of RNA and can also be used to position cap sites or splice junctions. In this procedure, single stranded DNA probes are generated that are labelled at either the 5' or 3' end. Probes are hybridized to the RNA and single strands digested with single-strand specific nucleases. The residual double strands are separated by PAGE (polyacrylamide gel electrophoresis) and visualised by autoradiography. The major drawback of the nuclease assay is the difficulty in controlling non-specific digestion of AT rich regions, which often transiently become single stranded at the operating temperature of the nucleases at 16^{o}C and so become available for digestion resulting in inaccurate data [12].

The ribonuclease protection assay, which follows on from nuclease mapping, uses a radiolabelled antisense probe generated to the RNA under investigation. This probe is hybridized to the RNA sample, and the resulting hybrid is treated with single-stranded ribonucleases. The surviving duplex is visualised by PAGE and autoradiography. High specific activity probes are used for rarer messages and low specific activity probes for abundant messages. This method is very sensitive, shows fewer degradation problems and a lower background than the S1 mapping assay. However the northern blot and RTPCR are both better at detecting multi-gene families and inter-species variants.

The technique of RTPCR has been extensively studied and found able to provide a reliable and reproducible method for the global study of gene expression. This technique relies on the efficient and faithful copying of an mRNA strand into a complementary DNA (cDNA) strand and the amplification of specific regions as directed by primers designed to particular regions of the sequence. When conditions are optimal, a single copy of cDNA can be detected in a complex mixture with this approach and it is the method of choice when looking for rare transcripts. Through carefully optimising the reaction conditions, a level of relativity can be achieved whereby within a series of reactions it is possible to illustrate which tissue source has more mRNA than others in the same series for a given sequence. The main disadvantage is that PCR will identify sequences in genomic DNA as readily as cDNA. Therefore, often it is necessary to DNase 1 treat the RNA prior to reverse transcription, which may damage RNA species. The only effective way to distinguish between the genomic and cDNA contributions is the careful design of primers such that the resulting amplicons differ in size. This can only be done for sequences that contain at least one intron. Quantitative RTPCR is a development of the PCR technique, which makes use of a fluorescent reporter molecule that is cleaved during the extension stage of the PCR cycle. An extension of this technique uses the fluorescent marker sybr green: this marker binds to the double stranded product of PCR, fluorescing significantly more when bound to this double stranded product than in the unbound state. The technique of real time PCR is a more sensitive technique in the search for rare transcripts. Some detection systems can distinguish between different reporter molecules at these low quantities, making it possible to multiplex for different genes in a single sample. However, this is a relatively new technology, and to conduct real-time RTPCR requires expensive instrumentation and reagents.

In-situ hybridisation is a technique that detects specific nucleic acids in morphologically preserved sections. It is a technique that has evolved from immuno-histochemistry, which highlights the morphology of a section with specific stains and antibodies to proteins. Traditionally, tissue sections are used, but whole cells or embryos and chromosome spreads can also be used. The method requires the fixing of samples either to a slide or as a whole mount in a chamber. The specifically designed nucleic acid is labelled and hybridised to the fixed sample. A wide variety of non-radioactive labels are available, but radioactivity is still regarded as the most sensitive method for the detection of rare sequences. The sequence for labelling is either ordered from oligo synthesising companies and then enzymatically labelled, or PCRed, cloned and transcribed as a radioactive riboprobe. In tissue sections, this technique microscopically locates the region of interest to the internal composition of individual cells. Pathbase (www.pathbase.net) is a database repository of histopathology photomicrographs and macroscopic images of the mouse throughout development, covering all strains, mutant, chimaeric, transgenic and knockout specimens. A similar database to map the protein expression in mouse sections – Atlas project - is in the planning stages at the Sanger Institute. In Edinburgh, a database is being built website (http://genex.hgu.mrc.ac.uk/ (EMAGE)) describing ISH expression patterns in the mouse embryo, which is publicly available. A controlled vocabulary is used during data/image entry to facilitate the linking with related databases. Databases of this nature are likely to become invaluable sources of information to the mouse scientific community.

Computational analysis of the genome has provided the biologist with a wealth of information relating to structure and composition for a number of organisms. In particular it has highlighted the similarities between species and the usefulness of the mouse as a model organism [13]. Of the 30,000 or so genes in the mouse sequence, 99% have direct counterparts in the human sequence [14]. There is 40% direct alignment between mouse and human, with 80% of human genes having one corresponding gene in the mouse. The mouse genome is, however, 14% smaller at 2.5Gb, compared to 2.9Gb for the human genome. Computer analyses have identified blocks of synteny, between the genomes, mapping the relevant regions with corresponding linking threads to illustrate those areas of synteny. An example of conserved synteny can be found in human chromosome 20: consisting of only three segments, which are identical to regions of mouse chromosome 2, with only one small segment altered in its order on the chromosome [5]. Through linking databases, information regarding sections of the mouse sequence can now be quickly compiled on screen for a thorough evaluation of genes of interest. With the complete annotation of the sequence, that reveals sequences flanking genes to identify promoter elements, non-protein coding transcripts, intron-exon structure, splice variants, alternative polyadenylation sites, sense and antisense pairings, and related genes, the task of the scientific researcher will be made much simpler. From a recent evaluation of the mouse gene Sox 8, using Compugen's Gencarta software, 9 possible transcripts were identified. On closer examination, only 4 of these were thought to be transcribed. From studies of this nature, experiments will in future be more precisely designed to study the biological condition under investigation.

Microarray is a widely available technology that is being employed to address increasingly complex scientific questions. The basic concept behind all microarray experimentation is the precise positioning of probes at high density on a solid support. The probes then act as molecular detectors [15]. From minuscule amounts of starting RNA, from as low as 30ng [16], the expression patterns of thousands of genes can be profiled in a single experiment. The aim is to describe quantitatively the gene expression profile characteristics associated with, for example, particular physiological processes or behaviours. By comparing patterns under different conditions, the association of particular gene expression profiles with particular types of activity can be quantified. Microarray probes are traditionally nucleic acid sequences as either synthesised oligonucleotides or PCR fragments. Protein arrays, which are composed of, either antibodies, aptamers, small-molecule drugs or phage particles, are currently under development. The target sample is labelled with radioactivity or a range of fluorescent molecules and hybridized to the probe array under stringent conditions. The resulting image is scanned and the location of successful hybridizations recorded together with the intensity of signal. In essence, by using different sources of target molecules, one can map changing temporal profiles or responses of tissues to signals or disease pathogenesis. Thus, this technique has scope in the fields of gene screening, target identification, pathway mapping, developmental biology and disease progression and diagnostic characteristics. This is a powerful technique as illustrated by Miki et al. [17], who profiled cDNA from 49 different mouse tissues and revealed related patterns of expression in related tissues.

For the pharmaceutical industries the microarray approach is beginning to be applied to RNA and protein alterations in early drug screening and non-clinical toxicology studies. For medicine it may become useful in the field of diagnostic biomarkers and patient tailored therapies (pharmacogenetics) [18]. However, initially, this technology requires more stringent validation to be useful to the medical profession. The entire process requires accurate selection, amplification and location of probes, accurate reference sequence information, identification of unique probe oligonucleotides, accurate distinction among multiple products of a single gene, accurate reconstruction of expressed sample nucleotide sequences, precise image scanning, and reproducible and accurate transformation of image files to numerical data. To be reliable, the probes must hybridize with high sensitivity and specificity, reproducibly between experiments and between laboratories, with a biologically meaningful outcome. One concern is that statistically a single microarray of 10,000 elements, with 99% accuracy may generate as many as 100 false positives, which can seriously affect results. Another source of error is that 50% of expressed eukaryotic genes are estimated to be expressed as splice variants, therefore the precise biological outcome may not be uniquely identifiable.

The enormous amount of data points generated by a single microarray experiment necessitates the application of bioinformatics to analyse results. The Gene Ontology consortium is creating standard contextual terms that are recognisable by computers to aid the analysis of these large data sets. With multiple experiments, data storage becomes an issue. There are a number of database repositories for microarray data, where data sets can be up- and down-loaded for comparative analysis [19]. This technique also requires some expensive instrumentation and significant computational skills and capabilities. A limited number of results from analysed microarray data can be validated with other technologies, such as real time RTPCR or in situ hybridisation. However, it would be unrealistic to attempt the validation in this way of the thousands of data points this technology generates.

Summary of Strengths and Weaknesses of these Gene Expression Methods

Method	**Total RNA amounts**	**Mode of detection**	**Weakness**	**Strength**
Northern Blot	10 pg – 30 µg	Radiography or nonradiography	Requires very high quality RNA.	Measures relative RNA levels within a single blot.
Nuclease S1 mapping	10pg – 100 µg	Radiography or nonradiography	Non specific digestion of AT rich regions	More sensitive than Northerns, used to map mRNA termini and intron/exon junctions.
Ribonuclease protection	5 femtograms – 100 µg	Radiography or nonradiography	Must know complete sequence of mRNA for probe design.	Up to 10 probes per sample for comparisons of muti-gene families.
Reverse Transcription Polymerase Chain Reaction	1 attog – 40 ng	Ethidium Bromide or Fluorescence	Primer design	Most sensitive method for mRNA detection and quantitation. Tolerates slightly degraded mRNA.
In-Situ hybridization	10-20 copies	Radiography or nonradiography	Long procedure	Precisely localises position of expression.
Computational Studies	N/A	World wide web	Relies on other people's data	Requires validation with biological samples.
Microarray technologies	5 – 100 µg	Radiography or fluorescence	Methods vary between labs. Optimisations ongoing.	Thousands of targets per sample.

<u>My approach</u>

To capitalise on the technological advances in high throughput screening (HTS) for gene expression profiling in as broad a number of tissues as possible, the most effective and sensitive method is the RTPCR with confirmative evidence from *in-situ* hybridization. This approach minimises costs and radioactivity usage. Computational analysis is essential for collecting and analysing sequence information, for primer/probe design and collating, presenting and evaluating laboratory results.

An earlier study, systematically characterising genes in the adult mouse over a range of tissues, used the technique of RTPCR. This database was submitted electronically to the Jackson Laboratories [20]. Our group had amassed over a thousand PCR primer pairs and had begun to create similar cDNA resources for mouse mammary gland differentiation during pregnancy, transgenic mouse gut [21], cancerous disease states in the human, and drug treatment in rats (much of this work is unpublished due to commercially orientated funding). Finely dissected tissues were supplied by collaborators and processed into cDNA resources to probe for genes of interest in a high through put fashion. With this background, and using the same rationale, I decided to create and test a similar resource of foetal tissues during progression to full term to help identify developmentally significant genes. Prior to joining the Sanger Centre, I had been working on the pre-implantation mouse embryo at the Anatomy Department, Cambridge University and understood the difficulties associated with obtaining reliable data from such small amounts of material. Through advances in HTS and molecular biology, it is becoming increasingly possible to obtain reproducible gene information from these small quantities of starting material.

Profiles of gene expression from this resource will illustrate when and to what degree (relative to other stages/tissues), genes are being expressed during foetal development, using a single method across staged tissues. This knowledge can be useful when studying individual tissues/organs during development, as it is possible to establish correlations, which illustrate which other tissues and organs are being affected by the same or similar genes, during the time of interest. With the careful design of primers, the approach will demonstrate the presence of mRNA in differing forms (splice variants). In a single experiment, genes expressed exclusively in the foetus (or adult) can be illustrated by using this collection of cDNAs (as shown by [22]). The resource can then be made available to the mouse community with more specific interests.

Having established the resource, we chose to investigate the usefulness of this resource by profiling the family of Sox genes.

The Family of Sox Genes

The Sox (Sry box) gene family, derive their name from the founding gene member Sry (Sex-determining Region of the Y chromosome). These genes are thought to encode transcription factors. However, there are many members of this expanding gene family and their functional role has yet to be fully elucidated. Sox genes are involved in a number of diverse functions, governing cell fate and organ development at different times and in different tissue locations during embryology. They thus provide a good target for trialling this cDNA resource.

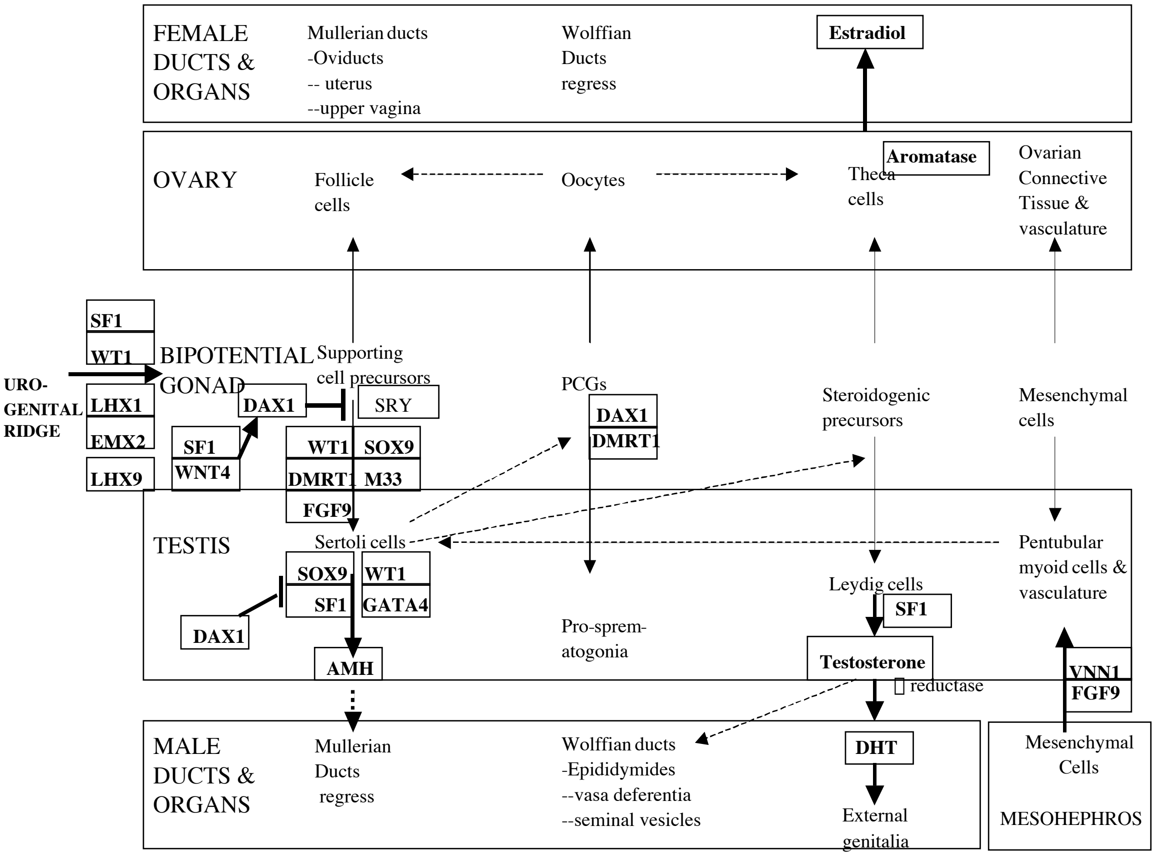
Sry Gene

The regulation of sex is controlled genetically in eutherians. The presence of a Y chromosome directs construction of a male gonad from the gonadal rudiment (Ford et al 1958 Nature **181**, p.1565). The later hormonal output of the male gonad then secondarily directs the developing embryo to form a male. The search for the testis-determining factor (*Tdy*) led researchers to a 35kb region of the Y chromosome now known to be the minimum necessary for male development. An intense research study culminated in 1990 in the identification of the sex-determining region (*Sry*) of the Y chromosome, found to be expressed in the genital ridge cells of the male mouse during the period in which testes begin to form [23]. The identity of *Sry* was confirmed by transgenic studies [24], in which female mice with two X chromosomes were made transgenic for the *Sry* gene, and developed as normal but sterile males with testes. Sry-like genes have been isolated from the males of a number of mammals [25] including rat, human, cattle [26] and voles [27]. However, in some lower vertebrates (birds, fish and amphibians) different mechanisms exist for sex determination. The avian females are heterogametic (ZW) and males homogametic (ZZ), which is the reverse of the situation found in mammals. And in alligators, temperature plays a role in sex determination [28].

Studies on the mouse XY gonad have found a marked increase in cell proliferation following the expression of the *Sry* gene at 11.25dpc, which is not mirrored in the XX gonad [29]. A reduced level of SRY in the developing mouse gonad, due to either a reduced number of SRY producing cells or a reduced level of SRY per cell, results in an ovarian pathway of gonad development [30]. The SRY protein is thought to act within the context of other gene products required for gonad development, and may act on or through one or more genes to ensure the differentiation and maintenance of Sertoli cells [31].

The human Y chromosome has been found to have approximately 50 genes (compared to approximately 1500 genes on the X chromosome), half of which are associated with male sex development and spermatogenesis [32]. Incidences of human sex reversal (some familial) have lead investigators to try to understand the nature of the genetic abnormalities involved. Many of these sex reversal cases show mutations within the HMG box of the *Sry* gene (XY females) or the translocation of *Sry* to an X chromosome (XX males). However, 10% [33] of XX males do not have the *Sry* gene, and 75% of XY females have no detected *Sry* mutation. Many of these cases have been explained by the observation of mutations in other genes, thereby implicating them in sex reversal. For example, mutations in a gene thought to act downstream of *Sry, Sox9*, causes a condition known as Campomelic dysplasia, a skeletal malformation syndrome. Two thirds of XY individuals with sex reversal show Campomelic dysplasia [34]. Conversely, using a knock-in approach, the gene *Sox9* has been shown to substitute for *Sry* at the stage of gonad development [35]. From these and other studies, a picture is emerging of the possible pathways surrounding the action of the *Sry* gene.

Figure 1: The Jigsaw Puzzle of Vertebrate Sex Determination



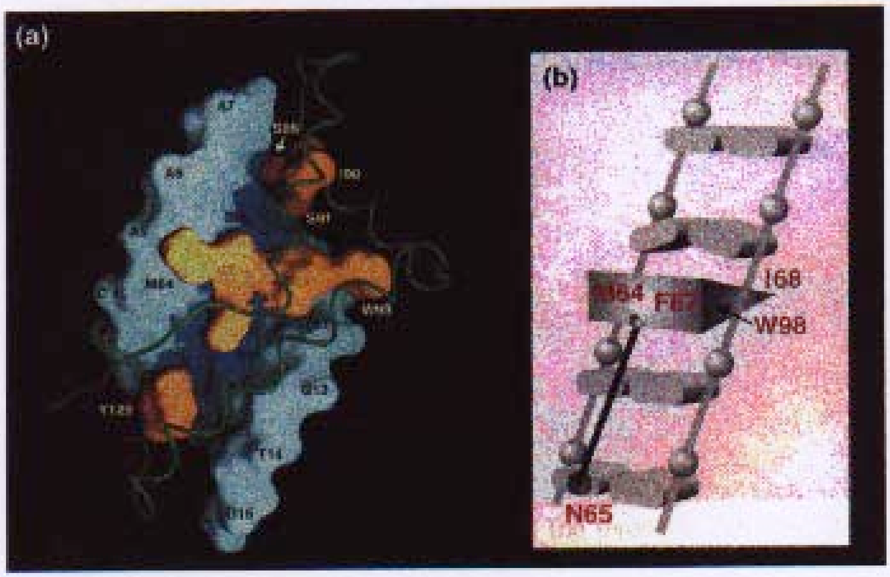
Vertebrate sex determination results from a complex network of regulatory interactions. The same basic set of genes appears to operate during early gonadal development in all vertebrate classes, despite the difference in mechanisms; and vertebrate sex determination results from a complex network of regulatory interactions. [1]

***Sox* Family of Genes**

Other members of the *Sox* gene family were identified through homology to a 79 amino acid motif known as the High Mobility Group (HMG) present in *Sry* [23] [36]. An HMG box of the *Sox* genes encodes an HMG domain with at least 50% amino acid identity with that of SRY. The SOX family are a subgroup of the HMG box proteins that show a highly specific tissue distribution and bind to identifiable DNA sequences with high affinities [37] [38]. However, this specificity appears to be context-dependent, that is, they appear to act in conjunction with other proteins [39] [40].

Figure 5 at the end of this section illustrates the homology of 19 of the mouse Sox gene HMG domains at the amino acid level. Classification of the HMG box proteins breaks them into two main families: (i) the MATA/TCF/SOX family and (ii) the UBF/HMG family. This partition of the HMG superfamily is directly related to the number of HMG boxes. Where there is more than one DNA binding domain, the binding occurs preferentially to bent DNA. This group include the HMG1 and HMG2 proteins. In the case of the *Sox* family, only one HMG domain is found. They may bind to pre-structured DNA with little or no specificity, but they also bind with high affinities in a sequence-specific manner to linear DNA [41]. The SOX HMG box binds specifically with the heptamer motif A/T A/T CAAAG of DNA, contacting the adenosines on both strands in the minor groove, causing the DNA helix to bend and opening up the helix [42]. As illustrated in Figure 2.

Figure 2: Protein-DNA interactions in the SRY-HMG/DNA



(a) Residues that directly interact with DNA bases are depicted as a yellow molecular surface attached to a green protein backbone. The intercalative wedge is formed by Met64, Asn65, Phe67, Ile68 and Try98, and intercalates between A-T basepairs 5 and 6. (b) Schematic representation of the wedge indicating how it inserts into the DNA. The wedge is anchored by Asn65, which hydrogen-bonds to the C4-G13 basepair below the site of insertion.

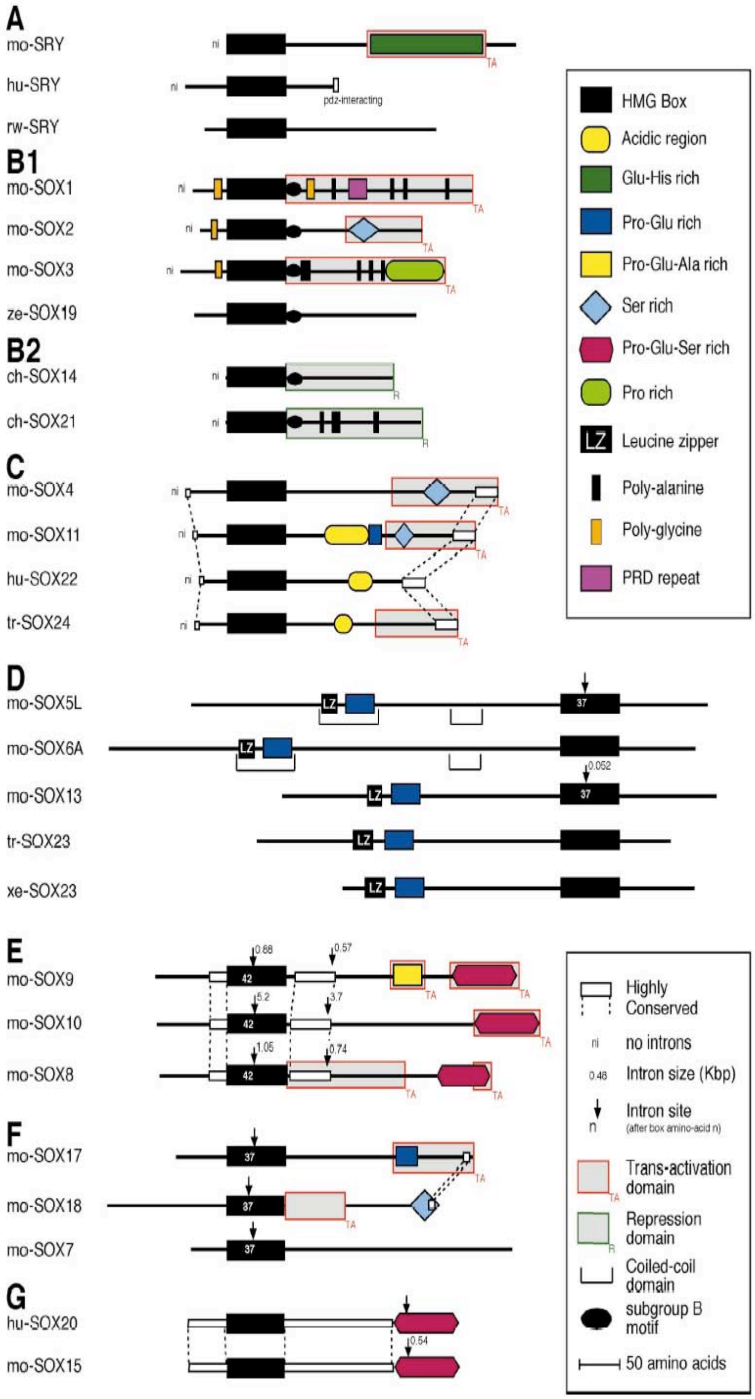
Ref:[43]

This altering of the local chromatin structure at specific sites may act to facilitate the interaction of distant enhancer nucleoprotein complexes with the transcription machinery [44] or the bending may serve to prevent the binding of such factors to adjacent sites in the major groove. This change in the architecture of the DNA is known to affect transcriptional elements vital to sex differentiation [45].

In addition to the HMG box, there are three other defining domains of *Sox* genes: a transactivating (TA) domain at the carboxy terminal, and distinctive C and N terminal domains [46]. A comparison of SOX protein domains and their groupings is illustrated below, together with the proposed phylogenic tree [2]. For some SOX groups, conservation of sequence and structure correlates with similarity of function, notably the B1 grouping which are all involved in the CNS (Central Nervous System). The phylogenic tree demonstrates the relatedness of the family of *Sox* genes between species, but it remains to be seen whether this relatedness is conserved at the molecular level in a mechanistic fashion.

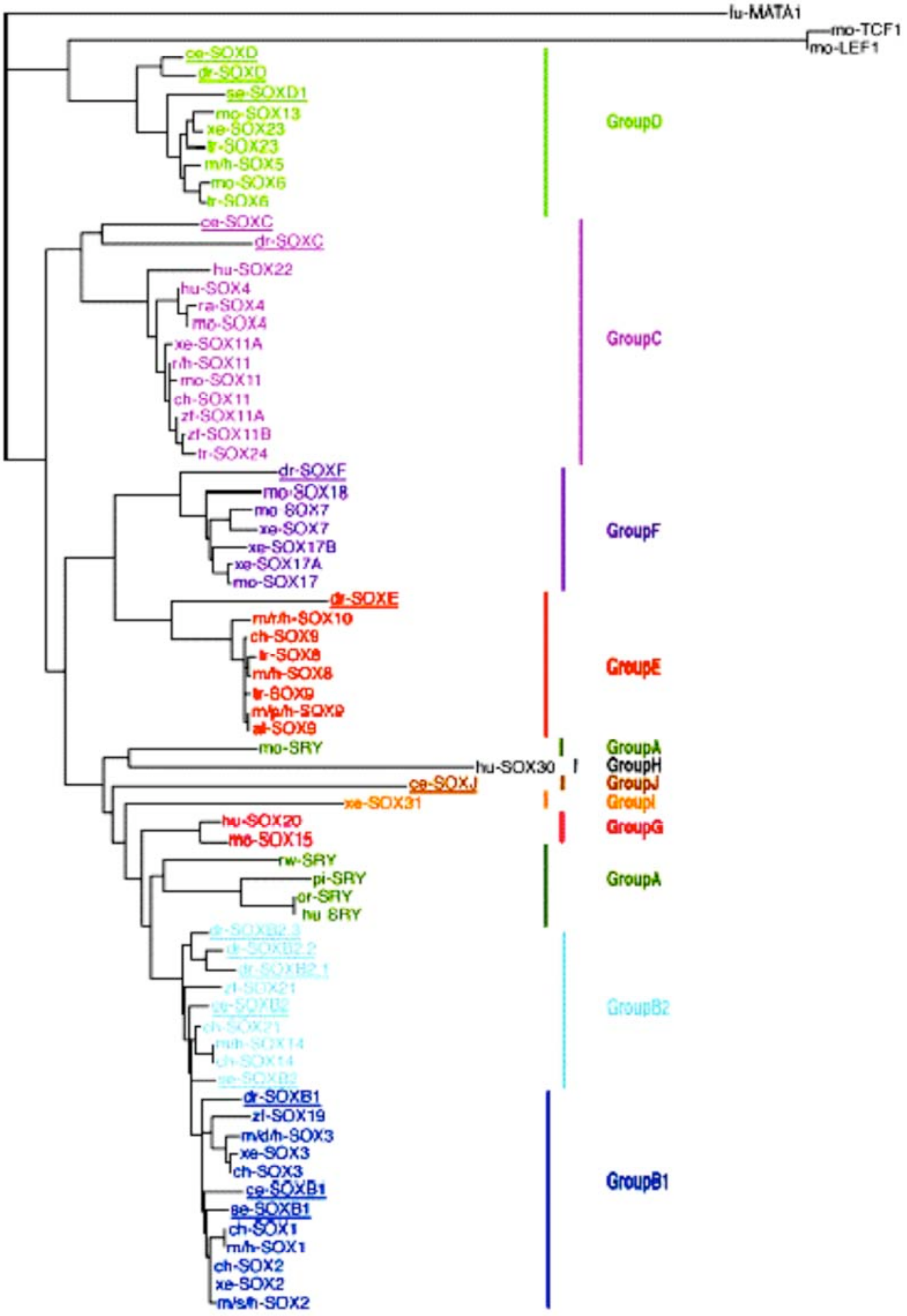
Through profiling the *Sox* genes across the mouse foetal collection of cDNAs in this project, a more detailed knowledge of expression will be recorded to add to the growing data on this interesting family of genes.

Figure 3: Schematic representation of SOX proteins



Diagram, highlighting conservation within SOX family groups. Proteins are arranged in groups as defined by HMG domain sequences. Various structural features, motifs, (demonstrated or putative) are shown along with intron positions and sizes where known. Genomic structures are known in some cases - ‘ni’ (no intron) indicates than an intronless structure has been reported[2].

Figure 4: Proposed Phylogeny of the SOX Transcription Factors



Tree created using the distance method FITCH (GCG), branch lengths are proportional to evolutionary distance or extent of divergence [2]

Figure 5: Sequence alignment of the HMG domains of the family of *Sox* genes.

```
SX12_MOUSE  .......... .......... .......MVW SQHERRKIMD QWPDMHNAEI SKRLGRRWQL LQDSEKIPFE REAERLRLKH MAD....... ..........
SX19_MOUSE  .......... .......... .......MVW SQIERRKIME QWPDMHNAEI SKRLGKRWKL LPDYEKIPFI KEAERLRLKH MA........ ..........
SOX9_MOUSE  .......... .......... .......MVW AQAARRKLAD QYPHLHNAEL SKTLGKLWRL LNESEKRPFV EEAERLRVQH KKD....... ..........
SX16_MOUSE  .......... .......... .......MVW SSAQRRQMAQ QNPKMHNSEI SKRLGAQWKL LDDEEKRPFV EEAKRLRARH LHDY...... ..........
SOX1_MOUSE  .........G GGTKANQDRV KRPMNAFMVW SRGQRRKMAQ ENPKMHNSEI SKRLGAEWKV MSEAEKRPFI DEAKRLRALH MKEHPDYKYR PRRK......
SOX2_MOUSE  .........G GNQKNSPDRV KRPMNAFMVW SRGQRRKMAQ ENPKMHNSEI SKRLGAEWKL LSETEKRPFI DEAKRLRALH MKEHPDYKYR PRRK......
SOX3_MOUSE  .........G GGGGSDQDRV KRPMNAFMVW SRGQRRKMAL ENPKMHNSEI SKRLGADWKL LTDAEKRPFI DEAKRLRAVH MKEYPDYKYR PRRK......
SX14_MOUSE  .......... ...SKPSDHI KRPMNAFMVW SRGQRRKMAQ ENPKMHNSEI SKRLGAEWKL LSEAEKRPYI DEAKRLRAQH MKEHPDYKYR PRRK......
SX15_MOUSE  .........G ASGGLPLEKV KRPMNAFMVW SSVQRRQMAQ QNPKMHNSEI SKRLGAQWKL LGDEEKRPFV EEAKRLRARH LRDYPDYKYR PRRK......
SOX8_MOUSE  .........G GGTLKAKPHV KRPMNAFMVW AQAARRKLAD QYPHLHNAEL SKTLGKLWRL LSESEKRPFV EEAERLRVQH KKDHPDYKYQ PRRR......
SX10_MOUSE  .........G AS..KSKPHV KRPMNAFMVW AQAARRKLAD QYPHLHNAEL SKTLGKLWRL LNESDKRPFI EEAERLRMQH KKDHPDYKYQ PRRR......
SOX4_MOUSE  .......... ........HI KRPMNAFMVW SQIERRKIME QSPDMHNAEI SKRLGKRWKL LKDSDKIPFI QEAERLRLKH MADYPDYKYR PRKK......
SX11_MOUSE  .......... ........HI KRPMNAFMVW SKIERRKIME QSPDMHNAEI SKRLGKRWKM LKDSEKIPFI REAERLRLKH MADYPDYKYR PRKK......
SOX7_MOUSE  .......... ...KSSESRI RRPMNAFMVW AKDERKRLAV QNPDLHNAEL SKMLGKSWKA LTLSQKRPYV DEAERLRLQH MQDYPNYKYR PRRK......
SX17_MOUSE  .......... ...AKAESRI RRPMNAFMVW AKDERKRLAQ QNPDLHNAEL SKMLGKSWKA LTLAEKRPFV EEAERLRVQH MQDHPNYKYR PRRR......
SX18_MOUSE  .........G ERQTADELRI RRPMNAFMVW AKDERKRLAQ QNPDLHNAVL SKMLGKAWKE LNTAEKRPFV EEAERLRVQH LRDHPNYKYR PRRK......
SOX5_MOUSE  VSESRIYRES RGRGSNEPHI KRPMNAFMVW AKDERRKILQ AFPDMHNSNI SKILGSRWKA MTNLEKQPYY EEQARLSKQH LEKYPDYKYK PRPKRTCLVD
SOX6_MOUSE  VAEARVYRDA RGRASSEPHI KRPMNAFMVW AKDERRKILQ AFPDMHNSNI SKILGSRWKS MSNQEKQPYY EEQARLSKIH LEKYPNYKYK PRPKRTCIVD
SX13_MOUSE  ....RHFSES R....NSSHI KRPMNAFMVW AKDERRKILQ AFPDMHNSSI SKILGSRWKS MTNQEKQPYY EEQARLSRQH LEKYPDYKYK PRPKRTCVVE
```

Sequence alignment of the HMG domain at the amino acid level for 19 members of the Sox family.

Objectives of my Research

To:

- create a panel of cDNAs from pure and intact RNA isolated from mouse foetal and placental tissues, to provide a reliable representation of the mouse tissue/organ transcriptome during development over the second half of gestation up to birth.
- demonstrate the validity and standardise the content of the panel.
- explore the specificity of the panel by PCR with primers designed to genes with known expression patterns in the adult mouse panel [20].
- illustrate the potential usefulness of the panel, using the newly emerging family of Sox genes as a model test system.
- compare the expression of two of the Sox gene members by PCR and *in-situ* hybridisation.
- review the process critically and examine alternative methods to explore gene expression during fetal development

References

http://www.ambion.com/techlib/basics/northerns/index.htm

1. Scherer, G., *The molecular genetic jigsaw puzzle of vertebrate sex determination and its missing pieces.* Novartis Found Symp 2002, 2002. **244**: p. 225- 236.
2. Bowles, J., G. Schepers, and P. Koopman, *Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators.* Developmental Biology, 2000. **227**(2): p. 239-255.
3. Gregory, S.G., et al., *A physical map of the mouse genome.* Nature, 2002. **418**(6899): p. 743-U3.
4. Powell, D., *Human Genome - Wellcome To The Genomic Age*. 2003, The Sanger Institute.
5. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-562.
6. Sudbeck, P., et al., *Sex reversal by loss of the C-terminal transactivation domain of human SOX9.* Nature Genetics, 1996: p. 230-232.
7. Kanai, Y., et al., *Identification of two Sox17 messenger RNA isoforms, with and without the high mobility group box region, and their differential expression in mouse spermatogenesis.* Journal of Cell Biology, 1996. **133**(3): p. 667-681.
8. Fuchs, P., et al., *Unusual 5' transcript complexity of plectin isoforms: novel tissue- specific exons modulate actin binding activity.* Human Molecular Genetics, 1999. **8**(13): p. 2461-2472.
9. Hastie, N., *Life, Sex, and WT1 Isoforms—Three Amino Acids Can Make All the Difference.* Cell, 2001. **106**: p. 391-394.

10. Mouchel, N., F. Broackes-Carter, and A. Harris, *Alternative 5' exons of the CFTR gene show developmental regulation.* Human Molecular Genetics, 2003. **12**(7): p. 759-769.

11. Kaufman, M.H., *The Atlas of Mouse Development*. 1998: Academic Press.

12. Docherty, K., *Gene Expression RNA Analysis*. Essential Techniques, ed. D. Rickwood. 1996: Wiley.

13. Bradley, A., *Mining the mouse genome - We have the draft sequence - but how do we unlock its secrets?* Nature, 2002. **420**(6915): p. 512-514.

14. Boguski, M.S., *Comparative genomics: The mouse that roared.* Nature, 2002. **420**(6915): p. 515-516.

15. Holloway, A.J., et al., *Options available - from start to finish - for obtaining data from DNA microarrays.* Nature Genetics, 2002. **32**: p. 481-489.

16. Smith, L., et al., *Single primer amplification (SPA) of cDNA for microarray expression analysis.* Nucleic Acids Research, 2003. **31**(3): p. art. no.-e9.

17. Miki, R., et al., *Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays.* PNAS, 2001. **98**(5): p. 2199-2204.

18. Gerhold, D.L., R.V. Jensen, and S.R. Gullans, *Better therapeutics through microarrays.* Nature Genetics, 2002. **32**: p. 547-552.

19. Stoeckert, C.J., H.C. Causton, and C.A. Ball, *Microarray databases: standards and ontologies.* Nature Genetics, 2002. **32**: p. 469-473.

20. Freeman, T.C., et al., *Expression Mapping of Mouse Genes.* MGI Direct Data Submission, 1998.

21. Bush, T.G., et al., *Fulminant jejuno-ileitis following ablation of enteric glia in adult transgenic mice.* Cell, 1998. **93**(2): p. 189-201.

22. Takada, S., et al., *Delta-like and Gtl2 are reciprocally expressed, differentially methylated linked imprinted genes on mouse chromosome 12.* Current Biology, 2000. **10**(18): p. 1135-1138.

23. Gubbay, J., et al., *A gene-mapping to the sex-determining region of the mouse y-chromosome is a member of a novel family of embryonically expressed genes.* Nature, 1990: p. 245-250.

24. Koopman, P., et al., *Male Development of Chromosomally Female Mice Transgenic for Sry.* Nature, 1991. **351**(6322): p. 117-121.

25. Margarit, E., et al., *Identification of conserved potentially-regulatory sequences of the SRY gene from 10 different species of mammals.* Biochem. Biophys. Res. Commun., 1998. **245**(2): p. 370 - 377.

26. Liu, W., et al., *A radiation hybrid map for the bovine Y Chromosome.* Mamm Genome, 2002. **13**(6): p. 320-6.

27. Fernandez, R., et al., *Mapping the SRY in Microtus cabrerae: a vole species with multiple SRY copies in males and females.* Genome, 2002. **45**(2): p. 600 - 603.

28. Morrish, B. and A. Sinclair, *Vertebrate sex determination; many means to an end.* Reproduction, 2002. **124**(4): p. 447- 457.

29. Schmahl, J., et al., *Sry induces cell proliferation in the mouse gonad.* Development, 1999. **127**: p. 65 - 73.

30. Washburn, L.L., K.H. Albrecht, and E.M. Eicher, *C57BL/6J-T-Associated Sex Reversal in Mice Is Caused by Reduced Expression of a Mus domesticus Sry Allele.* Genetics, 2001. **158**: p. 1675 - 1681.

31. Lovell-Badge, R., C. Canning, and R. Sekido, *Sex-determining genes in mice: building pathways.* Genetics and Biology of Sex Determination, 2002. **244**: p. 4-22.

32. Graves, J.A.M., *The rise and fall of SRY.* Trends in Genetics, 2002. **18**(5): p. 259 - 264.

33. Boucekkine, C., et al., *Clinical and Anatomical Spectrum in Xx Sex Reversed Patients - Relationship to the Presence of Y-Specific DNA-Sequences.* Clinical Endocrinology, 1994. **40**(6): p. 733-742.

34. Meyer, J., et al., *Mutational analysis of the SOX9 gene in campomelic dysplasia and autosomal sex reversal: Lack of genotype/phenotype correlations.* Human Molecular Genetics, 1997: p. 91-98.

35. Videl, V.P.I., et al., *Sox9 induces testis development in XX transgenic mice.* Nature, 2001. **28**(3): p. 216-217.

36. Denny, P., et al., *A conserved family of genes related to the testis determining gene, SRY.* Nucleic Acids Res, 1992. **20**(11): p. 2887.

37. Harley, V.R., et al., *The HMG box of SRY is a calmodulin binding domain.* Febs Letters, 1996. **391**(1-2): p. 24-28.

38. Pevny, L.H. and R. LovellBadge, *Sox genes find their feet.* Current Opinion in Genetics & Development, 1997. **7**(3): p. 338-344.

39. Kamachi, Y., M. Uchikawa, and H. Kondoh, *Pairing SOX off with partners in the regulation of embryonic development.* TRENDS IN GENETICS Osaka 5650871, Japan Osaka Univ, Inst Mol & Cellular Biol, Suita, Osaka 5650871, Japan, 2000. **16**(4): p. 182-187.

40. Wilson, M. and P. Koopman, *Matching SOX: partner proteins and co-factors of the SOX family of transcriptional regulators.* Current Opinion in Genetics & Development, 2002. **12**(4): p. 441-446.

41. Soullier, S., et al., *Diversification pattern of the HMG and SOX family members during evolution.* Journal of Molecular Evolution, 1999. **48**(5): p. 517-527.

42. van de Wetering, M. and H. Clevers, *Sequence-specific interaction of the HMG box proteins TCF-1 and SRY occurs within the minor groove of a Watson-Crick double helix.* EMBO J, 1992. **11**(8): p. 3039-44.

43. Werner MH, et al., *Molecular determinants of mammmalian sex.* Trends in Biochem Sci, 1996. **8**: p. 302 - 208.

44. Vaccari, T., et al., *Hmg4, a new member of the Hmg1/2 gene family.* Genomics, 1998: p. 247-252.

45. Harley, V.R., et al., *DNA binding activity of recombinant SRY from normal males and XY females.* Science, 1992. **255**(5043): p. 453-6.

46. Kamachi, Y., K.S.E. Cheah, and H. Kondoh, *Mechanism of regulatory target selection by the SOX high-mobility- group domain proteins as revealed by comparison of SOX1/2/3 and SOX9.* Molecular and Cellular Biology, 1999. **19**(1): p. 107-120.