Eutherian-specific homeobox genes in the preimplantation embryo; function, evolution, and applications.



Amy H Royall

Merton College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy.

Michaelmas 2019

Declaration and Statement of Authorship.

I declare that this thesis is entirely my own work carried out under the supervision of Professor Peter WH Holland, unless otherwise stated. It has not been submitted, either partially or in full, for any other qualification at this or any institution. Any work quoted, paraphrased or referenced has been appropriately indicated and the original work cited.

I have been approved to submit this thesis in integrated format by the Mathematical, Physical and Life Sciences division. The work contained within chapters 2 and 3 have been published as original scientific research manuscripts with all work and writing taking place whilst holding D.Phil student status. Chapter 2 is a manuscript published in *EvoDevo* in 2018 with myself as first author, chapter 3 is a manuscript published in 2019 in *Proceedings of the Royal Society B*, also with myself as first author. Where publications have multiple authors, the contributions of each author is stated in the respective "Introduction to Manuscript" section and signed authorship forms are inserted following the main text of the chapter. All co-authors on each manuscript evaluated the work and approved the final manuscripts. The majority of the work in each research article is attributed to me.

I acknowledge the Wellcome Trust Centre for Human Genomics, the Zoology sequencing service, and the BGI for transcriptome sequencing data. My funding was provided through the BBSRC Doctoral Training Partnership (DTP) in Interdisciplinary Biosciences [BB/M011224/1] and the Zoology Department, University of Oxford awarded to myself, with some initial experiments in chapter 2 funded by the EPA Cephalosporin Fund, and a Royal Society International Exchanges Grant.

Amy H Royall,

Merton College.

Acknowledgements

The only way to begin this thesis is with a huge thank you to my supervisor, Peter Holland. For his continuous support, dedication and mentorship throughout, for help directing my work, enthusiasm for my results, and for introducing me to both the wild Wallingford beer festival and Nobel laureates. I have been encouraged to publish my findings and to press on when things go wrong, as they so often do. Peter's commitment to his graduate students' work and wellbeing has enabled my time as a DPhil student to be one of my most rewarding experiences so far, even in the face of an asbestos-related disaster, and will continue to inspire me as I move onto the next stage.

I would like to thank the people who have directly helped me progress as a DPhil student. Tim Davies and Paul Fairchild for support and advice with iPSCs, and for rescuing us during the Tinbergen disaster, Thomas Dunwell for initial training, Ignacio Maeso for always going above and beyond to give advice and to help with the complexities of this research project, Maria-José, Sounak, and Anish for patiently walking me through complicated methodologies.

Thank you to the wider Oxford Evo-devo group past and present (listed here alphabetically); Emily Baker, Sharon Cornwell, Serena Dai, Harrie Drage, Adam Hargreaves, Carlos Herrera, Tom Lewin, Silvia Mercurio, Vasilis Papadogiannis, Steve Pates, Alessandro Pennati, Guillaume Poncelet, Roddy Pracana, Shan Quah, and Sonia Trigueros. You have been on this rollercoaster with me. The insight and advice (emotional and academic) from each of you played no small part in the making of this thesis.

Thank you to my Oxford family. Thank you Alex B for holding my hand from the moment we met (and I banged my head on a bus) and for joining me for last minute pints when the going got tough. Thank you Liisa, for being a beam of positive energy and continuous support (and for helping me with maths on day 1). Thank you both for filling my life with acapella Kelly Clarkson (+ a trillion friend points). Thank you Alex P, Eleonora, Harrie (again), Ivo, Izzy, Jo, and Kaytee for your continuing friendship and love through my time as a D.Phil student in Merton College and Oxford.

Thank you to my other family, legal and not. To my parents, to my brother, to my fairy godparents (Gill & David), to Jessica and the very much extended list of my support network. I am very lucky to have friends who ride the highs and the lows with me (even when the highs and lows are about evolutionary genetics and cell culture and are not very relatable).

Thank you, Paul. You have supported and encouraged me absolutely, you are the ultimate cheerleader.

Table of Contents

List of Figures IX

List of Tables XI

Acronyms XII

Thesis Abstract 1

Chapter 1: General Introduction

1. Gene Evolution 3

1.1 Gene duplication and evolution 3
1.2 Duplication, Degeneration, Complementation (DDC) 6

2. Homeobox genes; evolution, duplication, and loss 8

2.1 The Homeobox superclass 8
2.2 Gain and loss of homeobox genes 9
2.3 Evolution of the homeodomain sequence 10
2.4 The PRD-like class and Otx family 11

3. ETCHbox genes and preimplantation embryogenesis 13

3.1 Preimplantation development and milestones 13
3.2 Pluripotency and induced Pluripotent Stem Cells (iPSCs) 15
3.3 Comparative development of the early mammalian embryo 16
3.4 ETCHbox genes 19
3.4.1 Current understanding of human ETCHbox genes 21
3.4.2 Current understanding of mouse ETCHbox genes 22

4. Project aims and objectives 25

5. References 27

Chapter 2: Mouse Obox and *Crxos* modulate preimplantation transcriptional profiles revealing similarity between paralogous mouse and human homeobox genes

Introduction to manuscript 38
Abstract 40

1. Introduction 42

2. Materials and Methods 47

2.1 Embryonic expression and phylogenetic analyses 47
2.2 Ectopic gene expression 47
2.3 Differential gene expression analysis 48
2.4 Profile enrichment 49

3. Results 50

3.1 Sequence and expression diversity within expanded mouse ETCHbox clusters 50
3.2 Ectopic expression of *Crxos* and *Obox* genes induces transcriptomic changes 55
3.2.1 *Crxos* up- and down-regulated profiles 56
3.2.2 Obox up- and down-regulated profiles 56
3.3 Comparison between human and mouse ETCHbox gene functions 60
3.4 Targets with strong response are associated with embryonic milestones 61

4. Discussion 65

5. References 74

6. Supplementary files 78

7. Authorship form 86

Chapter 3: Of Eyes and Embryos; Subfunctionalisation of the *CRX* Homeobox Gene in Mammalian Evolution

Introduction to manuscript 88
Abstract 90

1. Introduction 91

2. Materials and Methods 95

2.1 Protein sequence analysis 95
2.2 Gene expression analysis 95
2.3 *Sminthopsis crassicaudata* conceptus expression 96
2.4 Gene expression in adult *Sminthopsis crassicaudata* tissues 96
2.5 Ectopic expression and transcriptomic responses 97

3. Results 100

3.1 Eutherian-specific substitutions in CRX homeodomain 100
3.2 *CRX* expression in metatherian embryos and eyes 102

3.3 Radical restructuring of transcriptomes by the metatherian CRX protein 105
3.4 Metatherian CRX protein regulated genes expressed in preimplantation development 107
3.5 Common downstream effects of metatherian *CRX* and eutherian Obox genes 109

4. Discussion 114

5. References 119

6. Supplementary files 123

7. Authorship form 128

Chapter 4: ETCHbox variability and evolution in Rodentia

Author contributions 130
Abstract 131

1. Introduction 132

2. Materials and Methods 137

2.1 Genome Analyses 137
2.2 ETCHbox sequence identification 137

3. Results 139

3.1 *Tprx1/Crxos* organisation in *Mus musculus* is conserved with *Rattus norvegicus* 139
3.2 *Crxos* sequence evolution 141
3.3 Obox organisation in rat sheds light on the Obox content of a murine common ancestor 143
3.4 ETCHbox distribution in Rodentia 146
3.5 The arrival of the Obox-like genes - the *Tprx2* transition 150

4. Discussion 154

4.1 ETCHbox gene evolution in murid rodents 154
4.2 Small scale ETCHbox gene duplication and loss is common in Rodentia 155
4.3 *Tprx2* became Obox-like during Myomorpha evolution 157
4.4 Implications 158

5. References 160

6. Supplementary files 164

Chapter 5: ETCHbox genes and Pluripotency

	Author contributions	171
	Abstract	172
1.	Introduction	173
2.	Materials and Methods	178
	2.1 Bioinformatic analysis of pluripotent cells	178
	2.2 Induced pluripotent stem cell (iPSC) culture	178
	2.3 RNA-extraction and Reverse Transcriptase (RT)-PCR	179
	2.4 Fibroblast culture, electroporation, and reprogramming	180
	2.5 Analysis of marker gene expression in ETCHbox+ iPSCs	181
3.	Results	182
	3.1 Mouse ETCHbox genes are expressed in pluripotent stem cells	182
	3.2 *Crxos* may directly interact with the pluripotency network	185
	3.3 Effect of ETCHbox ectopic expression on reprogramming efficiency	185
	3.4 Regulation of two-cell embryo markers in iPSCs exposed to *Crxos* before reprogramming	187
	3.5 Regulation of *Zscan4c* in colonies exposed to human ETCHbox genes	190
4.	Discussion	191
	4.1 ETCHbox genes and the pluripotency network	191
	4.2 Exposure to ETCHbox genes might increase reprogramming efficiency	192
	4.3 Influence of ETCHbox exposure on totipotent markers in iPSCs	193
	4.4 Regulation of two-cell embryo markers by paralogous ETCHbox genes supports “Distributed Robustness” theory of ETCHbox evolution	195
	4.5 Regulation of *Zscan4c* in colonies exposed to *ARGFX* and *TPRX1* recapitulates ETCHbox functional antagonism in human cells	196
	4.6 Conclusions	198
6.	References	199
7.	Supplementary files	205

Chapter 6: General Discussion and Perspectives

1.	Synopsis	214
2.	Protein coding sequence changes and evolution	215
3.	ETCHbox genes and functional overlaps; robustness and evolution	216
	3.1 ETCHbox genes walk hand-in-hand with gaining functions	219
4.	The pluripotency network	224
5.	Perspectives and future work	226
6.	References	228

Appendices 231

List of figures

Identifier	Title	Page
1.1	The possible outcomes following gene duplication as described by the DDC model.	7
1.2	The progression of preimplantation mouse development from fertilisation to early blastocyst.	14
1.3	Differences in the timing of Embryonic Genome Activation (EGA) between mouse and human preimplantation embryo.	17
1.4	ETCHbox gene retention across mammals is highly variable.	20
1.5	Genomic organisation of *TPRX* genes of human and mouse.	23
2.1	ETCHbox genes in human and mouse	44
2.2	ETCHbox organisation and expression in preimplantation development	54
2.3	Enriched profiles following *Crxos* or *Obox* ectopic expression.	58
2.4	Overlap between genes downstream of Obox genes.	60
2.5	Similarity between genes downstream of human *ARGFX* and mouse *Crxos*.	62
2.6	Cellular and embryonic processes potentially regulated by mouse ETCHbox genes.	68
2.7	Simplified scenario for the evolution of functional redundancy.	71
3.1	Organisational and amino acid changes in mammalian CRX	101
3.2	Comparison of eutherian (mouse, top) and metatherian (fat-tailed dunnart, bottom) *CRX* gene expression	103
3.3	Results of *de novo* transcriptome search for *CRX* transcript	104
3.4	Effect on transcriptome of *CRX* ectopic gene expression.	106
3.5	Mouse preimplantation clusters significantly regulated by metatherian *CRX* ectopic expression	108

3.6	Gene sets downstream of mouse ETCHbox genes are more similar to gene sets downstream of metatherian *CRX* than eutherian *CRX*.	111
3.7	Genes with known roles in early eutherian development are similarly regulated following metatherian *CRX* and ETCHbox ectopic expression.	113
3.8	Duplication of *CRX* in eutherian lineage lead to specialisation of duplicates into embryonic or retina-specific functions	118
4.1	Published ETCHbox repertoire in Euarchontoglires shows that there has been numerous gene losses and additional duplications on the murid lineage.	134
4.2	Cladogram of the rodent evolutionary history showing representative families with available genome assemblies.	136
4.3	Comparison of the *Crx*/*Tprx* locus in rat to human and mouse.	140
4.4	Amino acid sequence evolution of eutherian TPRX1	142
4.5	Maximum-likelihood tree of Rat and Mouse Obox amino acid sequences (no OboxD) with superimposed groupings indicated by various colours.	145
4.6	Presence of homeodomains of ETCHbox genes in the genomes of selected rodent families suggests timings of ETCHbox losses.	148
4.7	Phylogenetic tree of homeodomain amino acid sequences of rodent TPRX.	153
4.8	Summary of ETCHbox gene loss in *Rodentia* suggests that the ETCHbox repertoire in mouse is not due to exceptional divergence	156
5.1	Development of the preimplantation embryo	176
5.2	Expression of mouse ETCHbox genes is almost exclusively restricted to that observed in the preimplantation embryo	183
5.3	Expression of ETCHbox genes in iPSCs by RT-PCR	184
5.4	The effect of ETCHbox pre-exposure on the efficiency of reprogramming.	186

5.5	The effect of *Crxos* exposure before reprogramming alters the expression of "totipotency" markers.	188
5.6	Exposure to human ETCHbox genes *ARGFX* and *TPRX1* prior to reprogramming alters the expression of *Zscan4c* in resulting stem cell colonies.	190
6.1	Partially overlapping functions of ETCHbox genes could allow for fast evolutionary change.	219
6.2	The "DDC+" hypothesis and ETCHbox genes.	220
6.3	Proposed extended mechanism for the evolution of the ETCHbox gene variability and overlapping functions	223

List of Tables

Identifier	Title	Page
2.1	Groupings of mouse Obox genes with corresponding expression patterns.	52
2.2	Genes with highest up-regulated expression fold-change following ectopic expression of mouse ETCHbox genes.	63
4.1	Family, species, and common name of genomes used to analyse rodent ETCHbox content.	147
4.2	Novel findings of ETCHbox search in rodents.	149
4.3	The transition from *Tprx2*-like to Obox-like in genes rodents.	151
5.1	Primers used for RT-PCR analysis of ETCHbox expression	179
5.2	PCR protocol used for RT-PCR.	179
5.3	Primers used for qPCR of potency marker genes	181
5.4	Statistical output following exposure of *Crxos* to fibroblasts before reprogramming	189

Acronyms

A/T/C/G – Adenine/Thymine/Cytosine/Guanine nucleotides

Aa – amino acids

ATAC – Assay for Transposase-Accessible Chromatin

Argfx – Arginine-50 Homeobox

CDS – Coding sequence

Crx – Cone-rod homeobox

Crxos – Crx-Opposite-strand

DDC – Duplication, Degeneration, Complementation model

DNA – deoxyribonucleic acid

Dprx – Divergent paired-related homeobox

EGA – Embryonic Genome Activation (see MZT)

ESC – Embryonic Stem Cell

ETCHbox – Eutherian Totipotent Cell Homeobox

GRN – Gene regulatory network

Hox – Homeotic genes (name of family)

ICM – Inner Cell Mass

iPSC – induced Pluripotent Stem Cell

Leutx – leucine-20 homeobox

LIF – Leukaemia Inhibitory Factor

MEF – Mouse Embryonic Fibroblast

mRNA – messenger RNA

MZT – Maternal Zygotic Transition (see EGA)

Obox – Oocyte-specific homeobox

Pargfx – Parent of arginine-50 homeobox

PRD – Paired class of homeodomain transcription factors

RNA – ribonucleic acid

TE - Trophectoderm

Tprx1 – Tetrapeptide repeat homeobox 1

Tprx2 – Tetrapeptide repeat homeobox 2

WGD – Whole Genome Duplication

Thesis Abstract

This thesis explores the evolution, function, and possible applications of a group of lineage-specific genes known as the Eutherian Totipotent Cell Homeobox (ETCHbox) genes. The ETCHbox genes duplicated from a highly conserved vertebrate eye-specific homeobox gene called Cone-rod homeobox (*CRX*) in the lineage leading to eutherian mammals. The ETCHbox genes are unusual in many respects. Firstly, they are eutherian-specific, meaning that they are only present in the genomes of one group of mammals and not in marsupials or monotremes, which only have *CRX*. Secondly, they are highly variable in copy number and sequence, both compared to the "parental" gene from which they diverged, and between different species of eutherian mammal. Thirdly, their expression is limited to a few stages in the very early preimplantation embryo, after which they are not known to be expressed in any tissue. How can duplicates of a gene conserved in sequence and in function become so divergent so quickly, and what can their role be?

I explore how the extreme divergence has evolved in eutherians, what their new functions could be, and whether their specificity for the very early embryo can be applied to the field of regenerative biology. I initially focus on the mouse genome, using ectopic expression to reveal unexpected functional overlaps between human and mouse paralogues, not orthologues, and hypothesise inferred ETCHbox function. I use sequence analysis and ectopic expression to explore the evolution of ETCHbox genes from *CRX* using comparisons of eutherian and marsupial homeobox genes and embryology, finding evidence for a nuanced subfunctionalisation occurring on the eutherian lineage between *CRX* and ETCHbox. I then continue the evolutionary approach specifically within the order Rodentia, where I find that stepwise gene losses and sequence divergences occurred. Finally, I examine the effect of ETCHbox genes on pluripotency before discussing a model for how ETCHbox complexities could evolve. Overall, I have used a blend of computational and experimental techniques to increase our understanding of ETCHbox gene evolution and function, before exploring potential applications.

Chapter 1: General Introduction

1. Gene Evolution

1.1 Gene duplication and evolution

A central question in the field of evolutionary-developmental biology is how genetic change can influence phenotypic change. One of the most influential mechanisms of changing gene content of a genome is through duplication. The duplication of a gene can occur through different mechanisms; a single locus can be duplicated through unequal crossing over events during cell division, or through retrotransposition of transcribed genes (Marques *et al.*, 2005). There are noted differences in the resulting duplicate following each mechanism, with genes duplicated through unequal crossing over often being in tandem and containing relevant regulatory elements such as promoters (Zhang, 2003). In contrast, retrotransposition occurs through transcribed messenger RNA (mRNA) being processed into double stranded copy DNA (cDNA) and being randomly inserted into the genome using transposable element machinery (Ostertag and Kazazian, 2001). This then leads to duplicates located distantly to each other in a genome, and the new ('daughter') gene can be identified often by the lack of introns and associated regulatory landscape as well as by the presence of a poly-A tail, a feature of processed mRNA. Importantly, only retrotransposed duplicates of genes expressed in the germ line are heritable (Booth and Holland, 2004).

Genes can also be duplicated as parts of larger regional duplications. Entire chromosomes can be duplicated through incomplete anaphase and sister chromatid separation in meiosis, large segments of chromosomes can be duplicated also by unequal crossing over, and whole genomes can be duplicated (Sankoff, 2001). Whole genome duplication (WGD) has been important in the shaping of vertebrate

genomes which have undergone two rounds of WGD, the most recent of which 500 million years ago on the stem lineage of vertebrates (Dehal and Boore, 2005; Kuraku, Meyer and Kuratani, 2009; Holland and Ocampo Daza, 2018). Teleost fish have also independently experienced a third WGD which has been important in their genome evolution, with salmonid fishes experiencing a fourth (Allendorf and Thorgaard, 1984; Glasauer and Neuhauss, 2014). All of these mechanisms of duplication provide additional genetic material that can be retained or lost through subsequent evolutionary mechanisms.

The importance of gene duplication in the production of new genetic material on which evolutionary processes can act has been discussed for decades. The publication of Susumu Ohno's book, *Evolution by Gene Duplication* (Ohno, 1970) highlighted the potential of gene duplication in the evolution of species to evolutionary biologists. While theories of how genomes might adapt to duplicate genes have been developing since the mid-1990s, the advancement of genome sequencing has revealed how commonplace gene duplication is in extant species (Zhang, 2003). Duplication of a gene creates paralogs which can further be duplicated to give larger gene families, genes of varied functions including transcription factors are known to exist in this familial structure. Generally, once a gene has duplicated both copies are expected to be under relaxed evolutionary constraint and mutation can occur randomly in both the coding sequence and regulatory elements (Conrad and Antonarakis, 2007). These changes can cause subtle changes in gene function and be purged or fixed in a population. If duplicates do develop distinct characteristics they then become under separate evolutionary constraints and can be maintained by independent selective pressures (Zhang, 2003). As well as duplication leading to distinct functions in genes, it can also lead

to the development of complex gene-regulatory networks (GRNs). One example of this has been shown in *Drosophila melanogaster* where the paralog of *eyeless*, *twin-of-eyeless*, regulates *eyeless'* expression (Zhang, 2003). If a duplication event occurs in a species-specific manner it could lead to species-specific gene functions which may allow further adaptation (Zhang, Zhang and Rosenberg, 2002). For example, following the duplication of an opsin gene in papilionid butterflies one duplicate underwent accelerated evolution on one lineage and changed photoreceptor light sensitivity from green-sensitive to red-sensitive, giving this group additional ranges of colour vision (Briscoe, 2001). Furthermore, it has been postulated that whilst many domains of transcription factors including Homeobox, Sox and T-box were present in single-celled ancestors, gradual expansions of these into transcription factor families may have influenced the evolution of multicellularity (Sebé-Pedrós and de Mendoza, 2015). These families then further grew in number through duplication events prior to an expansion of bilaterians and in extant metazoans now exist in large gene families some with more than 100 members (Larroux *et al.*, 2008).

Perhaps the best-known example of gene duplication in the evolutionary developmental biology field is that of the Hox genes. The Hox genes exist in a cluster in metazoans and specify the positions along the head-to-tail axis in the developing embryo. In invertebrates there is a single cluster but the gene number within this cluster varies between species, whereas vertebrates have several Hox clusters. Advances in genomic sequencing have demonstrated that the Hox cluster originated from duplication of a single ancestral Hox gene and that subsequent duplication of the cluster gave rise to the multiple clusters seen in vertebrates (Bailey *et al.*, 1997; Prince and Pickett, 2002).

1.2 Duplication, Degeneration, Complementation (DDC)

Following the duplication of a gene or a collection of genes, there are multiple mechanisms for resolving the additional redundant genetic material (Rastogi and Liberles, 2005). The most common of which is nonfunctionalisation (pseudogenisation) where one duplicate gradually degrades due to the lack of evolutionary constraint. In some cases however, the duplicates may diverge in function through one duplicate adopting a completely new function (neofunctionalisation) or through sharing of the ancestral function between the duplicates (subfunctionalisation). This retention of gene duplicates is classically described under the Duplication-Degeneration-Complementation (DDC) model of gene evolution (Force *et al.*, 1999, figure 1.1). Subfunctionalisation occurs when the ancestral gene carried out more than one function, following the relaxed evolutionary pressure after duplication, both copies are free to mutate resulting in each copy fully adopting one of the ancestral functions. This could be a temporal or spatial separation where the ancestor was expressed in multiple contexts, or where the ancestor carried out two distinct roles, in each scenario the duplicates split the functionality between them and are subsequently maintained under independent evolutionary processes. Subfunctionalisation can relieve evolutionary constraint as each duplicate is now under evolutionary pressure to fine-tune its function after the removal of any trade-off experienced by the ancestral dual-function gene, this is known as "Escape from Adaptive Conflict" (Hittinger and Carroll, 2007). Neofunctionalisation however, is when one duplicate retains ancestral function leaving the other to mutate as in nonfunctionalisation. The difference here is that mutations are purified rather than purged likely due to a fitness advantage which

occurred by random chance (figure 1.1). Neofunctionalisation can be a subtle change in function, or it can result in large sequence divergence through accelerated evolution as described above in butterflies. There are now many studies that give evidence for the DDC model of gene evolution in many species including invertebrates, vertebrates, and yeast (Force *et al.*, 1999; Li, Yang and Gu, 2005; Van Hoof, 2005; Suvorov *et al.*, 2017; Barton-Owen, Ferrier and Somorjai, 2018; Leite *et al.*, 2018).

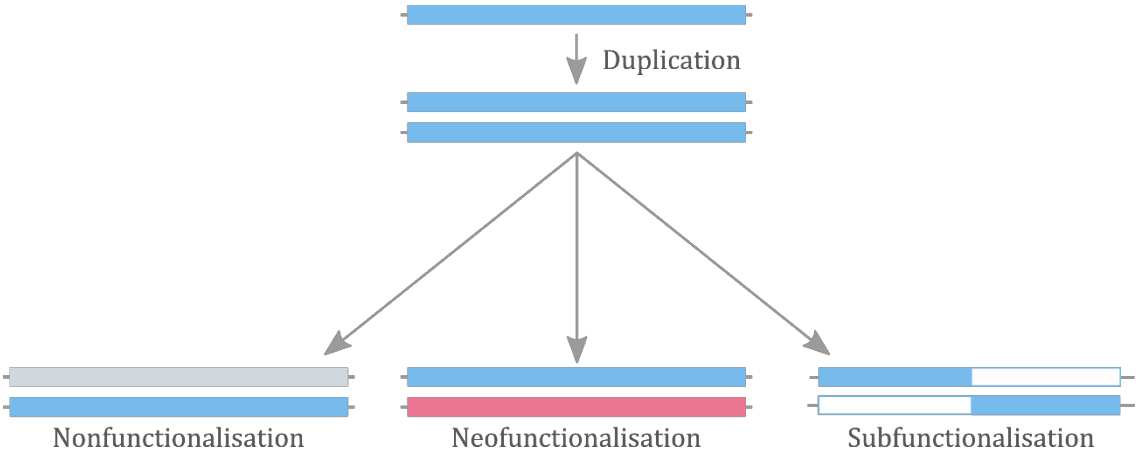


Figure 1.1 The possible outcomes following gene duplication as described by the DDC model. Colours reflect alterations in function rather than of genetic sequence specifically, where blue colour denotes ancestral function and grey is non-functional. When a gene duplicates, one copy can become non-functional (nonfunctionalisation), one copy can obtain an entirely new function (pink, neofunctionalisation), or the two duplicates can split ancestral function/s between them (subfunctionalisation). Image simplified and inspired by Force *et al.*, 1999.

In addition to nonfunctionalisation, neofunctionalisation, and subfunctionalisation described above, under some circumstances both duplicates can be retained with little change in sequence or regulation. In the case where two genes persist with comparable gene function and expression a species can become genetically robust. Distributed robustness is a model described by Andreas Wagner

(Wagner, 2005) and suggests that in a system where a network of genes are partially functionally redundant, mutation or loss of function in one copy can be compensated for and progression through the network can continue. This could be important in cases where a gene has a function vital for survival of the individual, such as in embryo development where failure of progression can lead to death of the embryo.

2. Homeobox genes; evolution, duplication, and loss

2.1 The Homeobox superclass

Members of the homeobox superclass are widespread throughout the eukaryotic world and their encoded proteins function as transcription factors (Holland, 2013). In the 1980s the first homeodomain proteins to be discovered were identified through homeotic transformations in the fruit fly *Drosophila melanogaster* and subsequently named "HOX". Misexpression of these genes led to major developmental abnormalities, with tissues taking on new identities, which exemplifies their vital role in patterning in embryonic development (Bürglin, 2011). Homeobox gene-encoded transcription factors all contain a variable DNA-binding domain known as the "homeodomain". At 60 amino-acids in length, homeodomains form a helix-loop-helix-turn-helix structure in which the third helix is responsible for DNA recognition and specificity (Bürglin, 2011). Aside from this characteristic structure, homeodomains (and hence homeobox-containing genes) can be highly variable and fall into 11 gene classes with 100 gene families in humans. Due to their vital functions, some homeobox proteins have been highly conserved throughout eukaryotes. However, this is not a steadfast rule; some homeobox genes such as the

X-linked *Rhox* cluster recently evolved (MacLean *et al.*, 2005) and some families have been lost from lineages completely (Zhong and Holland, 2011; Tsai *et al.*, 2013).

Homeobox genes are classified into classes characterised by additional domains present in all members, these classes are then divided into families. The Zinc-Finger (ZF) class for example have a Zinc-Finger domain in addition to the homeodomain (Holland, 2013). This method for grouping homeobox genes aims to reflect their evolutionary history - genes in the same family have a more similar ancestry than those in the same class but different family. Similarly, genes in a class are evolutionarily more closely related that those in another. Some families have evolved clusters, such as the Hox or Rhox clusters, with multiple genes in one chromosomal region most likely through tandem gene duplication (Zhong and Holland, 2011). This can lead to high numbers of similar genes which can result in gene evolution due to functional redundancy as described above. In the case of large recently evolved clusters, it is unclear whether natural selection has had adequate time to fully refine the duplicated genes.

2.2 Gain and loss of homeobox genes

In addition to local gene duplications, species' profiles of homeobox genes can vary a large amount due to gene loss and gain in a specific lineage. Research by Zhong and Holland (2011) found that whilst the rate of gaining new, divergent genes is similar in the human and rodent lineages, the rate of homeobox gene loss has been much greater in rodent evolution. This mode of gene evolution has led to the loss of

eight genes in the rodent lineage and just one in primates since their most recent common ancestor (Zhong and Holland, 2011). Conversely, there are examples where homeobox genes have been duplicated and gained throughout species. The *Dux* (double homeobox) gene family for example usually has four intron-containing loci (some lineages have less due to gene loss), but in humans it is thought that *DUX* became captured by a transposable element and there are now in excess of 30 *DUX* loci (Leidenroth and Hewitt, 2010; Holland, 2013). The mouse genome is also home to two gene clusters which arose due to mouse-specific duplications. The Rhox cluster contains 36 loci on the mouse X chromosome (MacLean *et al.*, 2005; Zhong and Holland, 2011) and the Obox cluster contains over 60 loci on mouse chromosome 7(Rajkovic *et al.*, 2002; Maeso *et al.*, 2016).

2.3 Evolution of homeodomain sequence

Where many homeobox genes have a primary sequence that has been highly conserved through their evolutionary history, there are a few examples of extreme divergence following gene duplication. One such example of this is the fat sand-rat (*Psammomys obesus*) *Pdx1* gene. Where the PDX1 homeodomain is very highly conserved across tetrapods, in the fat sand-rat accelerated fixation of mutation rate has led to a very high GC content in the coding-sequence (CDS) of *Pdx1* resulting in a number of non-synonymous mutations (Hargreaves *et al.*, 2017; Dai and Holland, 2019). This massive divergence in sequence has been of interest as the fat sand-rat is incredibly prone to type-2 diabetes. *Pdx1* has a function in the insulin secretion network in addition to the developing pancreas, areas of biology implicated in diabetes pathology, and so this provides some evidence that mutation in the coding

sequence of a homeodomain can correlate with a phenotypic and life-history trait change.

Secondly, the *bicoid* (*Bcd*) gene in fruit flies (*Drosophila melanogaster*) also displays large levels of divergence. The *bicoid* gene is located within the Hox cluster and is a maternally-supplied anterior-specifying gap-gene in the early embryo (Stauber *et al.*, 1999). Soon after the discovery of *bcd* there was an unsuccessful search to find the orthologue in other insects. This culminated in the late 1990s with the use of comparative genetics which defined *bcd* as a divergent Hox3-class gene (Stauber *et al.*, 1999). Ancestrally, *Hox3* is expressed maternally in the anterior embryo although in insects this anterior specification role has been lost. *Hox3* is expressed in the embryo and extra-embryonic tissue outside cyclorrhaphans and not anteriorly, it duplicated in the Cyclorrhaphan lineage to give two paralogs; *bcd* and its paralog *zerknuilt* (*zen*). *Zen* is expressed extra-embryonically and is more similar to *Hox3* in sequence. It has since been widely accepted that following *Hox3* duplication, asymmetric divergence of *zen* and *bcd* in cyclorrhaphans under the DDC model meant *zen* retained extra-embryonic expression and *bcd* became anteriorly expressed in the embryo (McGregor, 2005). This provides an example of rare cases where extensive homeobox gene divergence can make the evolutionary history of a gene elusive.

2.4 The PRD-like class and Otx family

The PRD-like class of homeobox genes were initially named after the Drosophila gene "*paired*" and includes, but is not limited to, the PRD genes with a paired-box "PAX" motif (Holland, 2013). In addition to PAX genes, PRD-like class

encompasses around 30-40 related gene families with large sequence diversity (Holland, Booth and Bruford, 2007; Bürglin, 2011; Holland, 2013). Other families within the PRD-like class include *Dux*, *Otx*, and *Rhox*.

The Otx gene family was initially identified through the Drosophila homolog *Orthodenticle* (*Otd*) which led to the discovery of mouse and human *OTX1*. *Otd* is a gap gene important for anterior segmentation in Drosophila and is homologous to *OTX1* and *OTX2* in vertebrates (Holland, Ingham and Krauss, 1992; Simeone *et al.*, 1992). In mammals *OTX1/2* have comparative roles defining the embryonic brain (Simeone, 1998). In 2000, analysis of the Xenopus genome revealed another Otx gene, *Otx5*, which hadn't been identified in mammals (Vignali *et al.*, 2000). Following this, phylogenetic analysis of other vertebrates identified that the Cone-rod homeobox (*CRX*) gene was a misnomer and *CRX* was actually the mammalian *OTX5* gene (Germot *et al.*, 2001; Sauka-Spengler *et al.*, 2001; Plouhinec *et al.*, 2003) which had encountered high levels of sequence evolution. The provisionally titled *OTX3* and *OTX4* genes were not confirmed as true Otx genes. Most vertebrates appear to have three Otx genes stemming from two vertebrate WGDs; *OTX1*, *OTX2*, and *CRX*. While in mammals *CRX* has lost its anterior-specifying role in the early embryo, it is expressed in the developing photoreceptive tissue of the eye and, in eutherian mammals, has further duplicated giving the ETCHbox genes (Maeso *et al.*, 2016).

3. ETCHbox genes and preimplantation embryogenesis

3.1 Preimplantation development and milestones

Preimplantation embryogenesis in eutherian mammals consists of many distinct morphological and transcriptomic changes as a single fertilized zygote develops into the blastocyst and prepares for implantation into the maternal endometrium. Whilst the zygote is considered totipotent – it has the potential to differentiate into all cell types of the embryo and the extra-embryonic tissue, the blastocyst has differentiated into distinct regions of cells that will form either the embryo proper and primitive endoderm (inner cell mass, ICM) or the extra embryonic tissue such as the placenta (trophectoderm)(figure 1.2)(Wang and Dey, 2006). The precise definition of "totipotency" is debated, with some definitions describing all embryonic stages prior to the first lineage decision following compaction as totipotent (Ziomek, Johnson and Handyside, 1982). With progression in the field, recent definitions are more stringent with some only defining the zygote and two-cell embryo as truly totipotent (Condic, 2013). This is because of emerging research describing asymmetric potency of individual cells in cleavage stage embryos, as well as evidence of a slow reduction in potency as the embryo develops rather than a single cell fate decision defining lineage separation (Casser *et al.*, 2017, 2019; J. Wang *et al.*, 2018; Boiani *et al.*, 2019). Whichever definition is used, these first stages of embryonic development encompass early cell-fate decisions and complex gene regulatory networks which act to pattern the early embryo and reduce cell differentiation potential.

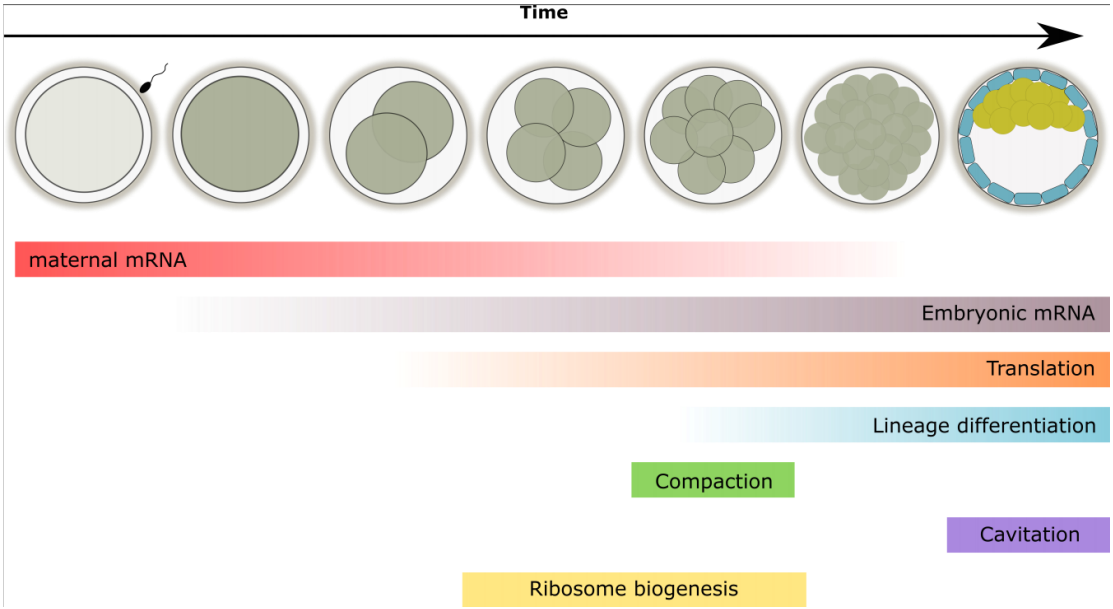


Figure 1.2 The progression of preimplantation mouse development from fertilisation to early blastocyst. This figure displays two cell types; the inner cell mass (yellow) and the trophectoderm (blue). Lengths of processes are approximate for overview understanding. There is a progressive degradation of maternal mRNA following fertilisation and, following EGA and ribosome biogenesis, the embryonic mRNA levels increase, followed by increased translation.

A crucial element of preimplantation development is the transition from dominant expression of maternal genes to zygotic genes, the "Maternal Zygotic Transition" (MZT), also known as "Embryonic Genome Activation" (EGA). Maternal genes are expressed in the embryo due to the presence of maternal mRNA in the oocyte, which is transcriptionally quiescent (Lu *et al.*, 2017; Liu *et al.*, 2018). At the EGA maternal mRNA is degraded and embryonic transcription activated (figure 1.2). In mice, this degradation is initiated by fertilization and is almost complete by the two-cell stage (figure 1.2; Wang and Dey, 2006; Onichtchouk and Voronina, 2015). Following the EGA, the embryonic genome begins to express key genes that push development through the first cell lineage decision. This involves division of blastomeres without growth followed by cell compaction, which occurs at the morula stage (~16/32 cells) and consists of a re-organisation of the cell membranes

and cell-cell interactions so that the cell boundaries become difficult to discern (figure 1.2). It is at this stage that it has been suggested that cell differentiation potential is not uniform and that the cells on the outside of the morula are fated to become trophectoderm (TE) and express markers such as *Cdx2*, whilst those on the inside (with more cell-cell contacts) will express inner cell mass markers such as *Oct4* and *Nanog* (Suwińska *et al.*, 2008; Nishioka *et al.*, 2009; Rayon *et al.*, 2014; Boroviak *et al.*, 2018; Saini and Yamanaka, 2018).

As the blastomeres within an embryo start to become specified to ICM or TE fate, the ICM fated cells are effectively exiting totipotency and becoming pluripotent (Casser *et al.*, 2017) – they are able to differentiate into all cells of the embryo proper but not the extra-embryonic tissue. Cells fated to be TE are effectively multipotent, only differentiating into extra-embryonic cells (Sasaki, 2010; Posfai, Rovic and Jurisicova, 2019). Both the totipotent and pluripotent states of blastomeres are of interest to developmental biologists and biomedical researchers alike as the ability to reinstate pluripotency in differentiated adult cells has potential applications in regenerative medicine (Maoz and Buganim, 2017; Baker and Pera, 2018; Ladstätter and Tachibana, 2018).

3.2 Pluripotency and induced Pluripotent Stem Cells (iPSCs)

Pluripotency is the capacity of a cell to differentiate into all cell types found in the embryo proper, but not into the cell types in the extra-embryonic tissue (i.e. trophectoderm-derived cells; Rossant, 2016). A totipotent cell has more differentiation potential and can become all cell types, including into those of the extra-embryonic tissue (Baker and Pera, 2018; Ladstätter and Tachibana, 2018). In the eutherian embryo the cells of the inner cell mass are described as pluripotent

whereas, generally, cells from cleavage stage embryos are totipotent. The process of exiting from potency is key in the study of developmental biology and in regenerative medicine. Due to our understanding of the pluripotent network, we can produce pluripotent stem cells from differentiated tissue (Takahashi and Yamanaka, 2006). These cells aid us in our understanding of these very early embryonic events and help advance transplant medicine. The production of these stem cells – induced Pluripotent Stem Cells – occurs through forced expression of genes that mark the inner cell mass of the embryo *in vivo*. The activity of these factors was discovered by Shinya Yamanaka and Kazutoshi Takahashi, and hence are sometimes known as the Yamanaka factors (Takahashi and Yamanaka, 2006). There are four factors: *Oct4*, *Klf4*, *Nanog*, and *C-Myc*. When these are over-expressed in differentiated cell types in unison they revert the genomic organisation to a pluripotent state. This has revolutionised the study of pluripotency, although production is limited by a low success rate (<1% efficiency).

3.3 Comparative development of the early mammalian embryo

The placental mammal (eutherian) preimplantation embryo displays many characteristics that are only found in this lineage and not their closest relatives, the marsupials (metatherians). These characteristics include the presence of a discernible inner cell mass which then develops into the epiblast and primitive endoderm. Marsupial embryos also do not appear to undergo compaction which, as far as we understand, sets apart cells fated to become extra-embryonic tissue from those fated to become inner cell mass (Frankenberg *et al.*, 2016). These differences seem drastic but make more sense when put in context; marsupials and eutherians

have been diverging for between 100 and 200 million years (Dunwell, Paps and Holland, 2017) . This evolution of the early embryo coincides with life history trait differences such as an extended gestation with a persisting invasive placentation and the complete loss of an egg shell coat in eutherians (Renfree, 2010; Frankenberg *et al.*, 2013).

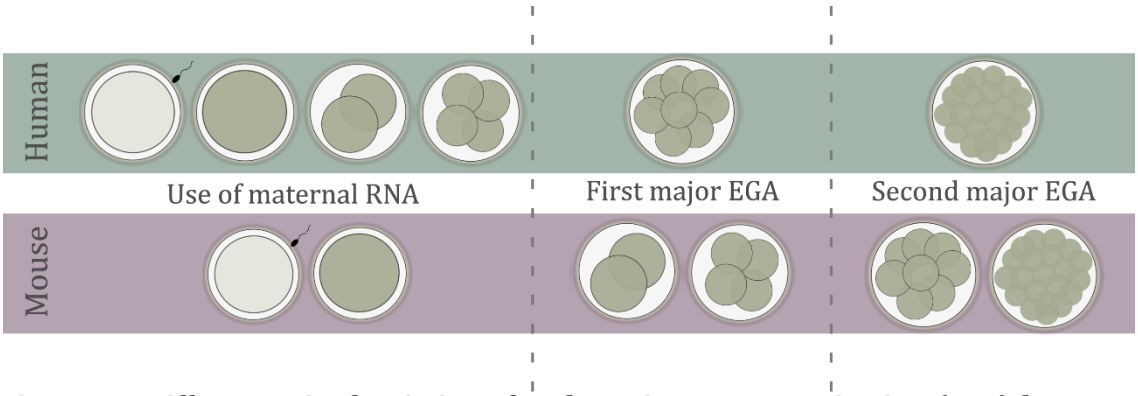


Figure 1.3 Differences in the timing of Embryonic Genome Activation (EGA) between mouse and human preimplantation embryo. The first major EGA in mouse occurs as early as the two-cell stage (pink), whereas in humans this occurs at the eight-cell stage (green). The second EGA where transcription becomes fully activated also occurs a stage earlier in mouse than human. Image inspired by Xue *et al.*, 2013.

While some major characteristics of the eutherian preimplantation embryo are conserved throughout, there is also variation within the eutherian clade. For example, the mouse embryo develops as a cone, whereas the human early embryo like a disk. Given the morphological differences slightly later in development, the cleavage stages of preimplantation embryos look remarkably similar (Rossant, 2015). Although there is a well described difference in the timings of major processes including the EGA (two-cell mouse embryo versus eight-cell human embryo; Xue *et al.*, 2013; figure 1.3).

Our understanding of the preimplantation embryo stems mostly from the study of mouse due to ethical restrictions using human embryos. However, as the

study of human embryogenesis advances, interesting molecular differences between mouse and human are becoming illuminated. Notably, a transcription factor that we classically associate with specification and marking of the TE, *Cdx2*, appears to have a different expression pattern in the mouse and human embryo. In both human and mouse blastocysts, *CDX2* definitively marks the trophectoderm (Nishioka *et al.*, 2009; Rayon *et al.*, 2014), however, embryo immunohistochemistry in humans has shown that *CDX2* is only expressed in the late stage blastocyst in the TE and therefore is not involved in specification of the TE (Niakan and Eggan, 2013). In contrast, in mouse *Cdx2* is expressed from the 16-cell morula stage and is subsequently restricted and uniformly expressed in the TE (Niakan and Eggan, 2013). Experiments such as these emphasize the importance of comparative biology as in this system where a mouse and human blastocyst may appear morphologically identical, the molecular signatures and developmental timelines may differ hugely.

The developmental potential of the ICM of the blastocyst is also variable, demonstrated by the different potencies of embryonic stem cells derived from mouse and human blastocysts (Theunissen *et al.*, 2016). Experimentally, a blastocyst can be dissociated and the inner cell mass is cultured as embryonic stem cells (ESCs) which are commonly used to study the molecular and genetic processes of early embryonic development, as well as having applications in the production of chimeric embryos (Kaufman and Evans, 1981; Martin, 1981). However, whilst the inner cell mass appears to be similar in morphology between mouse and human, the developmental potential of the resulting ESCs suggests that there are differential molecular processes at this equivalent stage (Weinberger *et al.*, 2016; Messmer *et al.*, 2019). The potency of mouse embryonic stem cells is described as having "naïve pluripotency", meaning there doesn't appear to be extensive restriction of the cell

types that can differentiate from this population. On the other hand, human ESCs exhibit "primed pluripotency"; there is some restriction on the cell types that they can differentiate into. Human ESCs have been described as being comparable in potency to stem cells derived from later stage mouse epiblasts; EpiSCs (Tesar *et al.*, 2007; Weinberger *et al.*, 2016). The basis of this difference in differentiation potency are still being elucidated and are of interest to the field of human regenerative medicine (Bates and Silva, 2017; Karagiannis and Takashima, 2017; Giulitti *et al.*, 2019).

Another area of eutherian embryo development that is variable is the extent to which the placenta invades the maternal tissue. Mouse and human for example have very invasive interstitial placentation but this is rare in eutherians with most species studied having less invasive placentation where the blastocyst is not entirely encapsulated by the endometrium (McGowen *et al.*, 2014).

3.4 ETCHbox genes

The Cone-Rod homeobox (*CRX*) gene is a highly conserved *Otx* family member residing on chromosome 19 in human and chromosome 7 in mouse. Expression of human and mouse *CRX* is restricted to developed and developing photoreceptor cells and has been shown to be vital to their differentiation (Freund *et al.*, 1997; Mansergh *et al.*, 2015; Masuda, Berlinicke and Zack, 2015; L. Wang *et al.*, 2018). Although *CRX* sequence and expression patterns are generally conserved throughout vertebrates, duplicates of *CRX* occurring in the lineage to eutherian mammals are much more variable. Originally discovered through studies of gene gain in the human genome (Booth and Holland, 2007) and subsequently defined as

CRX duplicates (Maeso *et al.*, 2016) the Eutherian Totipotent Cell homeobox (ETCHbox) genes *ARGFX*, *PARGFX*, *TPRX1*, *TPRX2*, *LEUTX*, and *DPRX* are highly variable in sequence and copy number (Maeso *et al.*, 2016)(figure 1.4). As their name suggests, expression of ETCHbox genes is highly specific and limited to a spike of expression in the totipotent eight-cell embryo and morula of cows (*Bos taurus*) and humans (*Homo sapiens*)(Jouhilahti *et al.*, 2016; Maeso *et al.*, 2016; Katayama *et al.*, 2018) with expression not found in any other healthy adult or embryonic tissue. The eight-cell to morula stage of preimplantation development is dynamic as described above. At these stages the embryo is undergoing compaction and the beginnings of the first cell lineage decision, as well as EGA in humans (Xue *et al.*, 2013; Freour and Vassena, 2017; Godini and Fallahi, 2018)(figure 1.3).

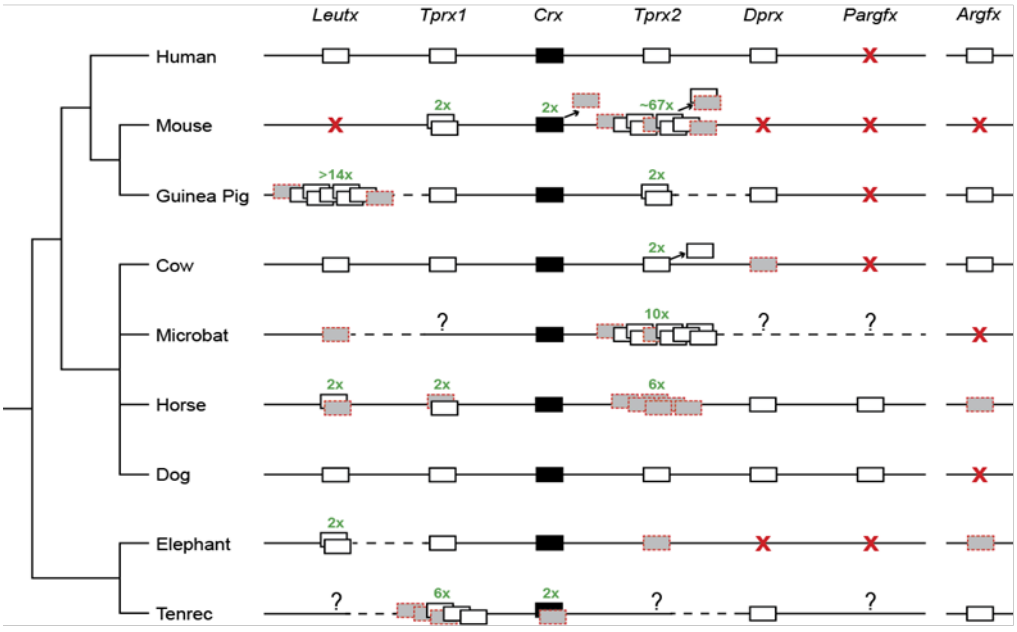


Figure 1.4 ETCHbox gene retention across mammals is highly variable. A cross indicates a gene loss, a solid box indicates a gene with a complete gene structure, a grey box indicates pseudogenisation, and a question mark indicates an incomplete assembly or other reason for incomplete understanding. Image edited from Maeso *et al.*, 2016.

3.4.1 Current understanding of human ETCHbox genes

The human genome has retained more ETCHbox genes than the mouse, with all known ETCHbox genes other than *PARGFX* retained (Maeso *et al.*, 2016). A screen of public RNA-sequencing datasets by Maeso and colleagues revealed the highly specific expression of ETCHbox genes in both human in cow in the preimplantation embryo, peaking at the eight-cell and morula stages which is indicative of their function. Over-expression of three of the human ETCHbox coding sequences (*LEUTX*, *TPRX1*, and *ARGFX*) revealed differential expression of a group of genes with a specific expression profile also in the eight-cell/morula embryonic stages (Maeso *et al.*, 2016). Interestingly this is not necessarily indicative of redundant functions as *LEUTX/TPRX1* expression caused down-regulation of gene in this profile whereas *ARGFX* expression caused up-regulation. Within this profile are genes with developmental importance such as *TGFβ*-signalling component *RHOB*, and embryo-maternal signalling molecule *HBEGF* – further evidence of the influence of human ETCHbox genes in key preimplantation processes. Furthermore, in a study of some ETCHbox genes in human embryonic stem cells it has been suggested that over-expression of human *DPRX* generally causes repression in target genes and that multiple PRD-like transcription factors (incl. ETCHbox) influence Yamanaka factor expression (Madissoon *et al.*, 2016). More globally, a motif analysis indicated that human ETCHbox genes are involved in the EGA which may be expected given their temporal expression profiles (Töhönen *et al.*, 2015).

There has been specific interest in the human *LEUTX* gene. Following the definition of a new transcriptional start site, over-expression of *LEUTX* appears to cause up-regulation of down-stream genes reminiscent of human EGA (Jouhilahti *et*

al., 2016). This important function of LEUTX is supported by a variant analysis which shows that mutations that potentially implicate the homeodomain are not common and are only found in heterozygotes (Katayama *et al.*, 2018). While the exact function of human ETCHbox genes is relevant to human biology, the variable nature of ETCHbox genes means that any conclusions on function are difficult to apply broadly to their function across eutherian mammals. For this reason, it is interesting to both study mammals other than humans and to compare human function to other eutherian ETCHbox function.

3.4.2 Current understanding of mouse ETCHbox genes

In comprehensive descriptions of homeobox genes, the mouse genome has been reported to be missing all ETCHbox genes including *TPRX1* and *TPRX2* (Holland, Booth and Bruford, 2007; Wilming, Boychenko and Harrow, 2015). It was hypothesised that the homeobox genes in the equivalent loci were consequences of mouse specific duplications (Zhong and Holland, 2011; Wilming, Boychenko and Harrow, 2015). However, their equivalent position on their respective chromosomes, flanked either side by conserved mammalian genes (*SEPW1/SELENOW* and *SULT2A*) suggested a common evolutionary history (figure 1.5). The orthology of *Tprx1/Crxos* and *Tprx2/*Obox has now been confirmed through increased taxon sampling (Maeso *et al.*, 2016). However, in the light of this orthology it is evident that these loci have diverged heavily in mouse.

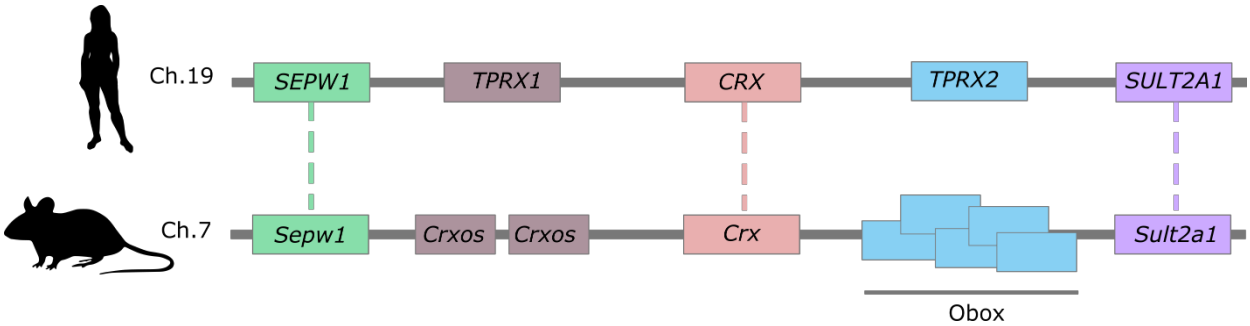


Figure 1.5 Genomic organisation of *TPRX* genes of human and mouse. The orthology of flanking genes *SEPW1*, *CRX*, and *SULT2A1* provided strong evidence that "lineage specific" *Crxos* and Obox genes were divergent *Tprx1* and *Tprx2* respectively. This has now been confirmed through increased taxon sampling.

In addition to this, the Obox cluster consists of upwards of 60 related genes and pseudogenes in a ~1.4 mega-base-pair genomic region (Zhong and Holland, 2011; Wilming, Boychenko and Harrow, 2015). The next largest cluster of PRD-like homeobox genes is the *Rhox* cluster with half as many loci (Holland, Booth and Bruford, 2007). This high level of duplication allows for the possibility that natural selection hasn't had adequate time to refine function after expansion, alternatively subfunctionalisation may have led to each expressed gene taking on a subsection of the functions of the ancestral gene.

Previous studies of *Crxos* revealed that the locus has duplicated and gives rise to three distinct transcripts; a full length transcript containing two homeodomains, and two variants consisting of one of the two homeodomains each (Saito *et al.*, 2009, 2010). Functional studies of this loci have suggested that *Crxos* is involved in cell pluripotency in early, preimplantation development in mice with the N-terminal shorter variant inhibiting differentiation markers in embryonic stem cells, and the C-terminal variant "stabilising" the pluripotent state (Saito *et al.*, 2010; Iha *et al.*, 2012; Sato *et al.*, 2015). Other studies by the same group suggest that the double-homeobox variant is involved in specifying the primitive endoderm lineage

(Soma *et al.*, 2012). Using transcript knock-down techniques, it has also been suggested that *Crxos* may be involved in correct ICM formation in the mouse blastocyst (Cui *et al.*, 2016). Even though there are a number of research articles elucidating the function of *Crxos* in the mouse embryo, results seem complicated and ectopic expression experiments generally limited to the pluripotent cell context which is itself variable and complex.

At the time of writing the Obox cluster on mouse chromosome 7 consists of six annotated genes (*Obox1, Obox2, Obox3, Obox5, Obox6, and Obox8*) in the most recent genome assembly (GRCm38/mm10, annotations from RefSeq and geneSymbol on UCSC browser). Previously, an *Obox4* had been annotated however this is no longer predicted as a gene. There are over 60 Obox-related loci however, that are not annotated as such, leading to a heavily expanded cluster (Appendix A1; Zhong and Holland, 2011; Maeso *et al.*, 2016). The Obox genes were initially identified as Oocyte-specific (Rajkovic *et al.*, 2002) however it is now known that expression can be detected throughout early murine development (Cheng *et al.*, 2007). Using RT-PCR, Cheng and co-authors (Cheng *et al.*, 2007) state the *Obox1* and *Obox2* are sequence-wise and transcriptionally equivalent and so are treated as one gene. They also find that *Obox1/2*, *3* and *5* are most highly expressed in the one-cell stage embryo and expression decreases until no expression can be detected in the morula. The *Obox6* gene however has elevated expression between the two-cell and morula stages. Interestingly, when *Obox6* was knocked out in mice by homologous recombination, offspring exhibited normal development and were fertile. This may suggest a redundancy mechanism and compensation by another gene.

During the progression of this thesis, studies investigating the potential role of Obox genes in the reprogramming cascade have been published (Schiebinger *et al.*, 2017; Wu *et al.*, 2017). Using inducible over-expression systems, it has been suggested that the presence of *Obox1* and *Obox6* can markedly increase the efficiency of reprogramming to pluripotency. This is of interest evolutionarily as the same Yamanaka factors act to reprogram differentiated cells in a number of mammals including human, mouse, rat and platypus (Whitworth *et al.*, 2018). On the other hand, there are no genes with similar sequence to Obox outside of mouse, indicating that research into inducing pluripotency could benefit from taking a more evolutionary viewpoint.

4. Project aims and objectives

The ETCHbox genes are of interest for multiple reasons. Firstly, they represent highly divergent homeobox genes with massively variable primary sequence and copy number across different species with seemingly uniform progression of preimplantation development. Secondly, their expression profile is unusual in that they are expressed in a short burst and then seemingly never utilised again.

The focus of this thesis is initially the function and evolution of ETCHbox genes, before asking whether our understanding can be applied to the field of regenerative medicine. The broad question I will answer is "How can species evolve such varied repertoires of highly and early expressed homeobox genes but seem to have relatively similar progression of preimplantation embryogenesis?". I will, for

the most part, concentrate on the mouse genome, which has lost four ETCHbox genes leaving only *Tprx1* and *Tprx2* which have each duplicated. I use a combination of bioinformatics, molecular techniques, and cell culture techniques to explore the evolution, function, and potential applications of ETCHbox genes. Specifically, I study the mouse repertoire of ETCHbox genes in detail and reveal how orthologue function compares across species (Chapter 2). I then study through RNA-sequencing the relationship between the function of ETCHbox genes and their ancestor *CRX*, presenting a hypothesis of subfunctionalisation during ETCHbox gene evolution in the eutherian lineage (Chapter 3). I conduct a search for ETCHbox genes across rodents to identify the timing of the extreme gene loss and divergence seen in mouse ETCHbox genes and ask if mice really are just weird (Chapter 4). Finally, I investigate the potential applications of these newly described homeobox genes in the field of somatic cell reprogramming and induced pluripotency and reveal some interplay between ETCHbox genes and the generation of the totipotent state (Chapter 5).

More broadly, the work presented in this thesis contributes to our understanding of gene evolution and genetic redundancy in the eutherian lineage and I discuss how vital gene functions coupled with duplications could facilitate gene divergence.

5. References

Allendorf, F. W. and Thorgaard, G. H. (1984) 'Tetraploidy and the Evolution of Salmonid Fishes', in *Evolutionary Genetics of Fishes*. Boston, MA: Springer US, pp. 1–53. doi: 10.1007/978-1-4684-4652-4_1.

Bailey, W. J., Kim, J., Wagner, G. P. and Ruddle, F. H. (1997) 'Phylogenetic reconstruction of vertebrate Hox cluster duplications', *Molecular Biology and Evolution*, 14(8), pp. 843–853. doi: 10.1093/oxfordjournals.molbev.a025825.

Baker, C. L. and Pera, M. F. (2018) 'Capturing Totipotent Stem Cells', *Cell Stem Cell*. Elsevier Inc., 22(1), pp. 25–34. doi: 10.1016/j.stem.2017.12.011.

Barton-Owen, T. B., Ferrier, D. E. K. and Somorjai, I. M. L. (2018) 'Pax3/7 duplicated and diverged independently in amphioxus, the basal chordate lineage', *Scientific Reports*, 8(1), pp. 1–11. doi: 10.1038/s41598-018-27700-x.

Bates, L. E. and Silva, J. C. (2017) 'Reprogramming human cells to naïve pluripotency: how close are we?', *Current Opinion in Genetics and Development*, 46, pp. 58–65. doi: 10.1016/j.gde.2017.06.009.

Boiani, M., Casser, E., Fuellen, G. and Christians, E. S. (2019) 'Totipotency continuity from zygote to early blastomeres – a model under revision', *Reproduction*, 10249, pp. 1–42. doi: doi.org/10.1530/REP-18-0462.

Booth, H. A. F. and Holland, P. (2004) 'Eleven daughters of NANOG', *Genomics*, 84(2), pp. 229–238. doi: 10.1016/j.ygeno.2004.02.014.

Booth, H. A. F. and Holland, P. W. H. (2007) 'Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line', *Gene*, 387(1–2), pp. 7–14. doi: 10.1016/j.gene.2006.07.034.

Boroviak, T., Stirparo, G. G., Dietmann, S., Herraez, I. H., Mohammed, H., Reik, W., Smith, A., Sasaki, E., Nichols, J. and Bertone, P. (2018) 'Global transcriptome analysis reveals conserved and divergent features of primate preimplantation development', *Development, in revision*.

Briscoe, A. D. (2001) 'Functional diversification of lepidopteran opsins following gene duplication', *Molecular Biology and Evolution*, 18(12), pp. 2270–2279. doi: 10.1093/oxfordjournals.molbev.a003773.

Bürglin, T. R. (2011a) 'Homeodomain Subtypes and Functional Diversity', in Hughes, T. R. (ed.) *A Handbook of Transcription Factors*. Dordrecht: Springer Netherlands, pp. 95–122. doi: 10.1007/978-90-481-9069-0_5.

Bürglin, T. R. (2011b) 'Homeodomain Subtypes and Functional Diversity', in Hughes, T. R. (ed.) *A Handbook of Transcription Factors*. Dordrecht: Springer Netherlands, pp. 95–122. doi: 10.1007/978-90-481-9069-0_5.

Casser, E., Israel, S., Witten, A., Schulte, K., Schlatt, S., Nordhoff, V. and Boiani, M. (2017) 'Totipotency segregates between the sister blastomeres of two-cell stage

mouse embryos', *Scientific Reports*. Springer US, 7(1), pp. 1–15. doi: 10.1038/s41598-017-08266-6.

Casser, E., Wdowik, S., Israel, S., Witten, A., Schlatt, S., Nordhoff, V. and Boiani, M. (2019) 'Differences in blastomere totipotency in 2-cell mouse embryos are a maternal trait mediated by asymmetric mRNA distribution', *Molecular Human Reproduction*, pp. 1–49. doi: 10.1093/molehr/gaz051.

Cheng, W. C., Hsiu, M. H. L., Yeh, Y. J. and Li, H. (2007) 'Mice lacking the Obox6 homeobox gene undergo normal early embryonic development and are fertile', *Developmental Dynamics*, 236(9), pp. 2636–2642. doi: 10.1002/dvdy.21261.

Condic, M. L. (2013) 'Totipotency: What It Is and What It Is Not', *Stem Cells and Development*, 23(8), pp. 796–812. doi: 10.1089/scd.2013.0364.

Conrad, B. and Antonarakis, S. E. (2007) 'Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease', *Annual Review of Genomics and Human Genetics*, 8(1), pp. 17–35. doi: 10.1146/annurev.genom.8.021307.110233.

Cui, W., Dai, X., Marcho, C., Han, Z., Zhang, K., Tremblay, K. D. and Mager, J. (2016) 'Towards Functional Annotation of the Preimplantation Transcriptome: An RNAi Screen in Mammalian Embryos', *Scientific Reports*. Nature Publishing Group, 6(November), pp. 1–12. doi: 10.1038/srep37396.

Dai, Y. and Holland, P. W. H. (2019) 'The Interaction of Natural Selection and GC Skew May Drive the Fast Evolution of a Sand Rat Homeobox Gene', *Molecular Biology and Evolution*. Edited by B. Su, 36(7), pp. 1473–1480. doi: 10.1093/molbev/msz080.

Dehal, P. and Boore, J. L. (2005) 'Two rounds of whole genome duplication in the ancestral vertebrate', *PLoS Biology*, 3(10). doi: 10.1371/journal.pbio.0030314.

Dunwell, T. L., Paps, J. and Holland, P. W. H. (2017) 'Novel and divergent genes in the evolution of placental mammals', *Proceedings of the Royal Society B: Biological Sciences*, 284(1864). doi: 10.1098/rspb.2017.1357.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. and Postlethwait, J. (1999) 'Preservation of Duplicate Genes by Complementary , Degenerative Mutations', *Genetics*, pp. 1531–1545. Available at: https://www.ncbi.nlm.nih.gov/pubmed/10101175.

Frankenberg, S. R., de Barros, F. R. O., Rossant, J. and Renfree, M. B. (2016) 'The mammalian blastocyst', *Wiley Interdisciplinary Reviews: Developmental Biology*, 5(2), pp. 210–232. doi: 10.1002/wdev.220.

Frankenberg, S., Shaw, G., Freyer, C., Pask, A. J. and Renfree, M. B. (2013) 'Early cell lineage specification in a marsupial: a case for diverse mechanisms among mammals', *Development*, 140(5), pp. 965–975. doi: 10.1242/dev.091629.

Freour, T. and Vassena, R. (2017) 'Transcriptomics analysis and human preimplantation development', *Journal of Proteomics*. Elsevier B.V., 162, pp. 135–140. doi: 10.1016/j.jprot.2016.10.004.

Freund, C. L., Gregory-Evans, C. Y., Furukawa, T., Papaioannou, M., Looser, J., Ploder, L., Bellingham, J., Ng, D., Herbrick, J. A. S., Duncan, A., Scherer, S. W., Tsui, L. C., Loutradis-Anagnostou, A., Jacobson, S. G., Cepko, C. L., Bhattacharya, S. S. and McInnes, R. R. (1997) 'Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor', *Cell*, 91(4), pp. 543–553. doi: 10.1016/S0092-8674(00)80440-7.

Germot, A., Lecointre, G., Plouhinec, J. L., Le Mentec, C., Girardot, F. and Mazan, S. (2001) 'Structural evolution of Otx genes in craniates.', *Molecular biology and evolution*, 18(9), pp. 1668–1678.

Giulitti, S., Pellegrini, M., Zorzan, I., Martini, P., Gagliano, O., Mutarelli, M., Ziller, M. J., Cacchiarelli, D., Romualdi, C., Elvassore, N. and Martello, G. (2019) 'Direct generation of human naive induced pluripotent stem cells from somatic cells in microfluidics', *Nature Cell Biology*. Springer US, 21(2), pp. 275–286. doi: 10.1038/s41556-018-0254-5.

Glasauer, S. M. K. and Neuhauss, S. C. F. (2014) 'Whole-genome duplication in teleost fishes and its evolutionary consequences', *Molecular Genetics and Genomics*, 289(6), pp. 1045–1060. doi: 10.1007/s00438-014-0889-2.

Godini, R. and Fallahi, H. (2018) 'Dynamics changes in the transcription factors during early human embryonic development', *Journal of Cellular Physiology*. doi: 10.1002/jcp.27386.

Hargreaves, A. D., Zhou, L., Christensen, J., Marlétaz, F., Liu, S., Li, F., Jansen, P. G., Spiga, E., Hansen, M. T., Pedersen, S. V. H., Biswas, S., Serikawa, K., Fox, B. A., Taylor, W. R., Mulley, J. F., Zhang, G., Heller, R. S. and Holland, P. W. H. (2017) 'Genome sequence of a diabetes-prone rodent reveals a mutation hotspot around the ParaHox gene cluster', *Proceedings of the National Academy of Sciences*, 114(29), pp. 7677–7682. doi: 10.1073/pnas.1702930114.

Hittinger, C. T. and Carroll, S. B. (2007) 'Gene duplication and the adaptive evolution of a classic genetic switch', *Nature*, 449(7163), pp. 677–681. doi: 10.1038/nature06151.

Holland, L. Z. and Ocampo Daza, D. (2018) 'A new look at an old question: When did the second whole genome duplication occur in vertebrate evolution?', *Genome Biology*. Genome Biology, 19(1), pp. 2–5. doi: 10.1186/s13059-018-1592-0.

Holland, P., Ingham, P. and Krauss, S. (1992) 'Mice and flies head to head', *Nature*, 358(6388), pp. 627–628. doi: 10.1038/358627a0.

Holland, P. W. H. (2013) 'Evolution of homeobox genes', *Wiley Interdisciplinary Reviews: Developmental Biology*, 2(1), pp. 31–45. doi: 10.1002/wdev.78.

Holland, P. W. H., Booth, H. A. F. and Bruford, E. A. (2007) 'Classification and nomenclature of all human homeobox genes', *BMC Biology*, 5, pp. 1–28. doi: 10.1186/1741-7007-5-47.

Van Hoof, A. (2005) 'Conserved functions of yeast genes support the duplication,

degeneration and complementation model for gene duplication', *Genetics*, 171(4), pp. 1455–1461. doi: 10.1534/genetics.105.044057.

Iha, M., Watanabe, M., Kihara, Y., Sugawara, S., Saito, K., Soma, M., Sato, S., Mori, Y., Kasuga, K., Kojima, I., Sasamura, R., Murata, J. and Kobayashi, M. (2012) 'Effect of ectopic expression of homeoprotein EGAM1C on the cell morphology, growth, and differentiation in a mouse embryonic stem cell line, MG1.19 cells', *Reproduction*, 143(4), pp. 477–489. doi: 10.1530/rep-11-0379.

Jouhilahti, E.-M., Madissoon, E., Vesterlund, L., Töhönen, V., Krjutškov, K., Plaza Reyes, A., Petropoulos, S., Månsson, R., Linnarsson, S., Bürglin, T., Lanner, F., Hovatta, O., Katayama, S. and Kere, J. (2016) 'The human PRD-like homeobox gene *LEUTX* has a central role in embryo genome activation', *Development*, 143(19), pp. 3459–3469. doi: 10.1242/dev.134510.

Karagiannis, P. and Takashima, Y. (2017) 'Surface Markers Guide the Journey toward Naive Pluripotency', *Cell Stem Cell*. Elsevier Inc., 20(6), pp. 737–738. doi: 10.1016/j.stem.2017.05.004.

Katayama, S., Ranga, V., Jouhilahti, E.-M., Airenne, T. T., Johnson, M. S., Mukherjee, K., Bürglin, T. R. and Kere, J. (2018) 'Phylogenetic and mutational analyses of human LEUTX, a homeobox gene implicated in embryogenesis', *Scientific Reports*, 8(1), p. 17421. doi: 10.1038/s41598-018-35547-5.

Kaufman, M. H. and Evans, M. J. (1981) 'Establishment in culture of pluripotential cells from mouse embryos', *Nature*, 292(July), pp. 154–156.

Kuraku, S., Meyer, A. and Kuratani, S. (2009) 'Timing of genome duplications relative to the origin of the vertebrates: Did cyclostomes diverge before or after?', *Molecular Biology and Evolution*, 26(1), pp. 47–59. doi: 10.1093/molbev/msn222.

Ladstätter, S. and Tachibana, K. (2018) 'Genomic insights into chromatin reprogramming to totipotency in embryos', *The Journal of Cell Biology*, p. jcb.201807044. doi: 10.1083/jcb.201807044.

Larroux, C., Luke, G. N., Koopman, P., Rokhsar, D. S., Shimeld, S. M. and Degnan, B. M. (2008) 'Genesis and expansion of metazoan transcription factor gene classes', *Molecular Biology and Evolution*, 25(5), pp. 980–996. doi: 10.1093/molbev/msn047.

Leidenroth, A. and Hewitt, J. E. (2010) 'A family history of DUX4: Phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene', *BMC Evolutionary Biology*, 10(1). doi: 10.1186/1471-2148-10-364.

Leite, D. J., Baudouin-Gonzalez, L., Iwasaki-Yokozawa, S., Lozano-Fernandez, J., Turetzek, N., Akiyama-Oda, Y., Prpic, N. M., Pisani, D., Oda, H., Sharma, P. P. and McGregor, A. P. (2018) 'Homeobox gene duplication and divergence in arachnids', *Molecular Biology and Evolution*, 35(9), pp. 2240–2253. doi: 10.1093/molbev/msy125.

Li, W. H., Yang, J. and Gu, X. (2005) 'Expression divergence between duplicate genes',

Trends in Genetics, 21(11), pp. 602–607. doi: 10.1016/j.tig.2005.08.006.

Liu, C., Ma, Y., Shang, Y., Huo, R. and Li, W. (2018) 'Post-translational regulation of the maternal-to-zygotic transition', *Cellular and Molecular Life Sciences*. Springer International Publishing, 75(10), pp. 1707–1722. doi: 10.1007/s00018-018-2750-y.

Lu, X., Gao, Z., Qin, D. and Li, L. (2017) 'A Maternal Functional Module in the Mammalian Oocyte-To-Embryo Transition', *Trends in Molecular Medicine*. Elsevier Ltd, 23(11), pp. 1014–1023. doi: 10.1016/j.molmed.2017.09.004.

MacLean, J. A., Chen, M. A., Wayne, C. M., Bruce, S. R., Rao, M., Meistrich, M. L., Macleod, C. and Wilkinson, M. F. (2005) 'Rhox: A new homeobox gene cluster', *Cell*, 120(3), pp. 369–382. doi: 10.1016/j.cell.2004.12.022.

Madissoon, E., Jouhilahti, E. M., Vesterlund, L., Töhönen, V., Krjutškov, K., Petropoulous, S., Einarsdottir, E., Linnarsson, S., Lanner, F., Månsson, R., Hovatta, O., Bürglin, T. R., Katayama, S. and Kere, J. (2016) 'Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos', *Scientific Reports*. Nature Publishing Group, 6(July), pp. 1–14. doi: 10.1038/srep28995.

Maeso, I., Dunwell, T. L., Wyatt, C. D. R., Marlétaz, F., Veto, B., Bernal, J. A., Quah, S., Irimia, M. and Holland, P. W. H. (2016) 'Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals', *BMC Biology*, 14(1), pp. 1–14. doi: 10.1186/s12915-016-0267-0.

Mansergh, F. C., Carrigan, M., Hokamp, K. and Farrar, G. J. (2015) 'Gene expression changes during retinal development and rod specification.', *Molecular vision*, 21(January), pp. 61–87. doi: 10.1097/MCC.0000000000000501.

Maoz, N. and Buganim, Y. (2017) 'Moving towards totipotency without a single miR-acle', *Cell Research*. Nature Publishing Group, 27(5), pp. 600–601. doi: 10.1038/cr.2017.30.

Marques, A. C., Dupanloup, I., Vinckenbosch, N., Reymond, A. and Kaessmann, H. (2005) 'Emergence of young human genes after a burst of retroposition in primates', *PLoS Biology*, 3(11), pp. 1970–1979. doi: 10.1371/journal.pbio.0030357.

Martin, G. R. (1981) 'Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells', *Proceedings of the National Academy of Sciences*, 78(12), pp. 7634–7638. doi: 10.1073/pnas.78.12.7634.

Masuda, T., Berlinicke, C. A. and Zack, D. J. (2015) 'Gene regulation: It matters who you hang out with', *Neuron*. Elsevier Inc., 86(1), pp. 7–9. doi: 10.1016/j.neuron.2015.03.057.

McGowen, M. R., Erez, O., Romero, R. and Wildman, D. E. (2014) 'The evolution of embryo implantation', *International Journal of Developmental Biology*, 58(2–4), pp. 155–161. doi: 10.1387/ijdb.140020dw.

McGregor, A. P. (2005) 'How to get ahead: The origin, evolution and function of bicoid', *BioEssays*, 27(9), pp. 904–913. doi: 10.1002/bies.20285.

Messmer, T., von Meyenn, F., Savino, A., Santos, F., Mohammed, H., Lun, A. T. L., Marioni, J. C. and Reik, W. (2019) 'Transcriptional Heterogeneity in Naive and Primed Human Pluripotent Stem Cells at Single-Cell Resolution', *Cell Reports*, 26(4), p. 815–824.e4. doi: 10.1016/j.celrep.2018.12.099.

Niakan, K. K. and Eggan, K. (2013) 'Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse', *Developmental Biology*. Elsevier, 375(1), pp. 54–64. doi: 10.1016/j.ydbio.2012.12.008.

Nishioka, N., Inoue, K. ichi, Adachi, K., Kiyonari, H., Ota, M., Ralston, A., Yabuta, N., Hirahara, S., Stephenson, R. O., Ogonuki, N., Makita, R., Kurihara, H., Morin-Kensicki, E. M., Nojima, H., Rossant, J., Nakao, K., Niwa, H. and Sasaki, H. (2009) 'The Hippo Signaling Pathway Components Lats and Yap Pattern Tead4 Activity to Distinguish Mouse Trophectoderm from Inner Cell Mass', *Developmental Cell*. Elsevier Ltd, 16(3), pp. 398–410. doi: 10.1016/j.devcel.2009.02.003.

Ohno, S. (1970) *Evolution by Gene Duplication*, *Springer-Verlag*. New York. doi: doi.org/10.1007/978-3-642-86659-3.

Onichtchouk, D. V. and Voronina, A. S. (2015) 'Regulation of Zygotic Genome and Cellular Pluripotency', *Biochemistry (Moscow)*, 80(13), pp. 1723–1733. doi: 10.1134/S0006297915130088.

Ostertag, E. M. and Kazazian, H. H. (2001) 'Biology of Mammalian L1 Retrotransposons', *Annual Review of Genetics*, 35, pp. 501–538. doi:10.1146/annurev.genet.35.102401.091032.

Plouhinec, J. L., Sauka-Spengler, T., Germot, A., Le Mentec, C., Cabana, T., Harrison, G., Pieau, C., Sire, J. Y., Véron, G. and Mazan, S. (2003) 'The mammalian Crx genes are highly divergent representatives of the Otx5 gene family, a gnathostome orthology class of orthodenticle-related homeogenes involved in the differentiation of retinal photoreceptors and circadian entrainment', *Molecular Biology and Evolution*, 20(4), pp. 513–521. doi: 10.1093/molbev/msg085.

Posfai, E., Rovic, I. and Jurisicova, A. (2019) 'The mammalian embryo's first agenda: Making trophectoderm', *International Journal of Developmental Biology*, 63(3–5), pp. 157–170. doi: 10.1387/ijdb.180404ep.

Prince, V. E. and Pickett, F. B. (2002) 'Splitting pairs: The diverging fates of duplicated genes', *Nature Reviews Genetics*, 3(11), pp. 827–837. doi: 10.1038/nrg928.

Rajkovic, A., Yan, C., Yan, W., Klysik, M. and Matzuk, M. M. (2002) 'Obox, a family of homeobox genes preferentially expressed in germ cells', *Genomics*, 79(5), pp. 711–717. doi: 10.1006/geno.2002.6759.

Rastogi, S. and Liberles, D. A. (2005) 'Subfunctionalization of duplicated genes as a

transition state to neofunctionalization', *Medizinische Monatsschrift fur Pharmazeuten*, 28(5), pp. 522–528. doi: 10.1186/1471-2148-5-28.

Rayon, T., Menchero, S., Nieto, A., Xenopoulos, P., Crespo, M., Cockburn, K., Cañon, S., Sasaki, H., Hadjantonakis, A. K., de la Pompa, J. L., Rossant, J. and Manzanares, M. (2014) 'Notch and Hippo Converge on Cdx2 to Specify the Trophectoderm Lineage in the Mouse Blastocyst', *Developmental Cell*, 30(4), pp. 410–422. doi: 10.1016/j.devcel.2014.06.019.

Renfree, M. B. (2010) 'Review: Marsupials: Placental Mammals with a Difference', *Placenta*. Elsevier Ltd, 31(SUPPL.), pp. S21–S26. doi: 10.1016/j.placenta.2009.12.023.

Rossant, J. (2015) 'Mouse and human blastocyst-derived stem cells: Vive les differences', *Development (Cambridge)*, 142(1), pp. 9–12. doi: 10.1242/dev.115451.

Rossant, J. (2016) 'Genetic Control of Early Lineages in Mammalian Embryos', *Annu Rev Genet*, 51(1), p. annurev-genet-120116-024544. doi: 10.1146/annurev-genet-120116.

Saini, D. and Yamanaka, Y. (2018) 'Cell Polarity-Dependent Regulation of Cell Allocation and the First Lineage Specification in the Preimplantation Mouse Embryo', in, pp. 11–35. doi: 10.1016/bs.ctdb.2017.10.008.

Saito, K., Abe, H., Nakazawa, M., Irokawa, E., Watanabe, M., Hosoi, Y., Soma, M., Kasuga, K., Kojima, I. and Kobayashi, M. (2010) 'Cloning of Complementary DNAs Encoding Structurally Related Homeoproteins from Preimplantation Mouse Embryos: Their Involvement in the Differentiation of Embryonic Stem Cells', *Biology of Reproduction*, 82(4), pp. 687–697. doi: 10.1095/biolreprod.108.075697.

Saito, R., Yamasaki, T., Nagai, Y., Wu, J., Kajiho, H., Yokoi, T., Noda, E., Nishina, S., Niwa, H., Azuma, N., Katada, T. and Nishina, H. (2009) 'CrxOS maintains the self-renewal capacity of murine embryonic stem cells', *Biochemical and Biophysical Research Communications*, 390(4), pp. 1129–1135. doi: http://dx.doi.org/10.1016/j.bbrc.2009.09.118.

Sankoff, D. (2001) 'Gene and genome duplication', *Current Opinion in Genetics and Development*, 11(6), pp. 681–684. doi: 10.1016/S0959-437X(00)00253-7.

Sasaki, H. (2010) 'Mechanisms of trophectoderm fate specification in preimplantation mouse development', *Development Growth and Differentiation*, 52(3), pp. 263–273. doi: 10.1111/j.1440-169X.2009.01158.x.

Sato, S., Nakazawa, M., Kihara, Y., Kubo, Y., Sato, Y., Kikuchi, T., Nonaka, A., Sasaki, A., Iwashita, J., Murata, J., Hosaka, M. and Kobayashi, M. (2015) 'Partial inhibition of differentiation associated with elevated protein levels of pluripotency factors in mouse embryonic stem cells expressing exogenous EGAM1N homeoprotein', *Journal of Bioscience and Bioengineering*, 120(5), pp. 562–569. doi: http://dx.doi.org/10.1016/j.jbiosc.2015.03.001.

Sauka-Spengler, T., Baratte, B., Shi, D. and Mazan, S. (2001) 'Structure and

expression of an Otx5-related gene in the dogfish Scyliorhinus canicula: Evidence for a conserved role of Otx5 and Crx genes in the specification of photoreceptors', *Development Genes and Evolution*, 211(11), pp. 533–544. doi: 10.1007/s00427-001-0191-2.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Liu, S., Lin, S., Berube, P., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A. and Lander, E. (2017) 'Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming.', *bioRxiv (preprint)*, p. 191056. doi: 10.1101/191056.

Sebé-Pedrós, A. and de Mendoza, A. (2015) 'Transcription Factors and the Origin of Animal Multicellularity', in Ruiz-Trillo, I. and Nedelcu, A. M. (eds) *Evolutionary Transitions to Multicellular Life: Principles and mechanisms*. Dordrecht: Springer Netherlands, pp. 379–394. doi: 10.1007/978-94-017-9642-2_18.

Simeone, A. (1998) 'Otx1 and Otx2 in the development and evolution of the mammalian brain', *EMBO*, 17(23), pp. 6790–6798.

Simeone, A., Acampora, D., Gulisano, M., Stornaiuolo, A. and Boncinelli, E. (1992) 'Nested expression domains of four homeobox genes in developing rostral brain', *Nature*, 358(6388), pp. 687–690. doi: 10.1038/358687a0.

Soma, M., Iha, M., Kihara, Y., Sato, S., Sato, Y., Sato, S., Mori, Y., Sugawara, S., Kasuga, K., Kojima, I. and Kobayashi, M. (2012) 'Preferential emergence of cell types expressing markers for primitive endoderm lineages in mouse embryonic stem cells expressing exogenous EGAM1 homeoprotein', *Journal of Bioscience and Bioengineering*, 114(3), pp. 342–346. doi: http://dx.doi.org/10.1016/j.jbiosc.2012.04.012.

Stauber, M., Jackle, H., Schmidt-Ott, U., Jäckle, H. and Schmidt-Ott, U. (1999) 'The anterior determinant bicoid of Drosophila is a derived Hox class 3 gene.', *Proceedings of the National Academy of Sciences of the United States of America*, 96(7), pp. 3786–9. doi: 10.1073/pnas.96.7.3786.

Suvorov, A., Jensen, N. O., Sharkey, C. R., Fujimoto, M. S., Bodily, P., Wightman, H. M. C., Ogden, T. H., Clement, M. J. and Bybee, S. M. (2017) 'Opsins have evolved under the permanent heterozygote model: insights from phylotranscriptomics of Odonata', *Molecular Ecology*, 26(5), pp. 1306–1322. doi: 10.1111/mec.13884.

Suwińska, A., Czołowska, R., Ozdzeński, W. and Tarkowski, A. K. (2008) 'Blastomeres of the mouse embryo lose totipotency after the fifth cleavage division: Expression of Cdx2 and Oct4 and developmental potential of inner and outer blastomeres of 16- and 32-cell embryos', *Developmental Biology*, 322(1), pp. 133–144. doi: 10.1016/j.ydbio.2008.07.019.

Takahashi, K. and Yamanaka, S. (2006) 'Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors', *Cell*, 126(4), pp. 663–676. doi: 10.1016/j.cell.2006.07.024.

Tesar, P. J., Chenoweth, J. G., Brook, F. A., Davies, T. J., Evans, E. P., Mack, D. L.,

Gardner, R. L. and McKay, R. D. G. (2007) 'New cell lines from mouse epiblast share defining features with human embryonic stem cells', *Nature*, 448(7150), pp. 196–199. doi: 10.1038/nature05972.

Theunissen, T. W., Friedli, M., He, Y., Planet, E., O'Neil, R. C., Markoulaki, S., Pontis, J., Wang, H., Iouranova, A., Imbeault, M., Duc, J., Cohen, M. A., Wert, K. J., Castanon, R., Zhang, Z., Huang, Y., Nery, J. R., Drotar, J., Lungjangwa, T., Trono, D., Ecker, J. R. and Jaenisch, R. (2016) 'Molecular Criteria for Defining the Naive Human Pluripotent State', *Cell Stem Cell*, 19(4), pp. 502–515. doi: 10.1016/j.stem.2016.06.011.

Töhönen, V., Katayama, S., Vesterlund, L., Jouhilahti, E. M., Sheikhi, M., Madissoon, E., Filippini-Cattaneo, G., Jaconi, M., Johnsson, A., Bürglin, T. R., Linnarsson, S., Hovatta, O. and Kere, J. (2015) 'Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development', *Nature Communications*, 6, pp. 1–9. doi: 10.1038/ncomms9207.

Tsai, I. J., Zarowiecki, M., Holroyd, N., Garciarrubio, A., Sanchez-Flores, A., Brooks, K. L., Tracey, A., Bobes, R. J., Fragoso, G., Sciutto, E., Aslett, M., Beasley, H., Bennett, H. M., Cai, J., Camicia, F., Clark, R., Cucher, M., De Silva, N., Day, T. A., Deplazes, P., Estrada, K., Fernández, C., Holland, P. W. H., Hou, J., Hu, S., Huckvale, T., Hung, S. S., Kamenetzky, L., Keane, J. A., Kiss, F., Koziol, U., Lambert, O., Liu, K., Luo, X., Luo, Y., MacChiaroli, N., Nichol, S., Paps, J., Parkinson, J., Pouchkina-Stantcheva, N., Riddiford, N., Rosenzvit, M., Salinas, G., Wasmuth, J. D., Zamanian, M., Zheng, Y., Cai, X., Soberon, X., Olson, P. D., Laclette, J. P., Brehm, K., Berriman, M., Morett, E., Portillo, T., Jose, M. V., Carrero, J. C., Larralde, C., Morales-Montor, J., Limon-Lason, J., Cevallos, M. A., Gonzalez, V., Ochoa-Leyva, A., Landa, A., Jimenez, L. and Valdes, V. (2013) 'The genomes of four tapeworm species reveal adaptations to parasitism', *Nature*, 496(7443), pp. 57–63. doi: 10.1038/nature12031.

Vignali, R., Colombetti, S., Lupo, G., Zhang, W., Stachel, S., Harland, R. M. and Barsacchi, G. (2000) 'Xotx5b, a new member of the Otx gene family, may be involved in anterior and eye development in Xenopus laevis', *Mechanisms of Development*, 96(1), pp. 3–13. doi: 10.1016/S0925-4773(00)00367-1.

Wagner, A. (2005) 'Distributed robustness versus redundancy as causes of mutational robustness', *BioEssays*, 27(2), pp. 176–188. doi: 10.1002/bies.20170.

Wang, H. and Dey, S. K. (2006) 'Roadmap to embryo implantation: Clues from mouse models', *Nature Reviews Genetics*, 7(3), pp. 185–199. doi: 10.1038/nrg1808.

Wang, J., Wang, L., Feng, G., Wang, Y., Li, Y., Li, X., Liu, C., Jiao, G., Huang, C., Shi, J., Zhou, T., Chen, Q., Liu, Z., Li, W. and Zhou, Q. (2018) 'Asymmetric Expression of LincGET Biases Cell Fate in Two-Cell Mouse Embryos', *Cell*. Elsevier Inc., 175(7), p. 1887–1901.e18. doi: 10.1016/j.cell.2018.11.039.

Wang, L., Qi, A., Pan, H., Liu, B., Feng, J., Chen, W. and Wang, B. (2018) 'A novel CRX frameshift mutation causing cone-rod dystrophy in a Chinese family: A case report', *Medicine*, 97(32), p. e11499. doi: 10.1097/MD.0000000000011499.

Weinberger, L., Ayyash, M., Novershtern, N. and Hanna, J. H. (2016) 'Dynamic stem cell states: Naive to primed pluripotency in rodents and humans', *Nature Reviews*

Molecular Cell Biology. Nature Publishing Group, 17(3), pp. 155–169. doi: 10.1038/nrm.2015.28.

Weltner, J., Balboa, D., Katayama, S., Bespalov, M., Krjutškov, K., Jouhilahti, E. M., Trokovic, R., Kere, J. and Otonkoski, T. (2018) 'Human pluripotent reprogramming with CRISPR activators', *Nature Communications*, 9(1), pp. 1–12. doi: 10.1038/s41467-018-05067-x.

Whitworth, D. J., Limnios, I., Gauthier, M., Weeratunga, P., Ovchinnikov, D., Baillie, G., Grimmond, S., Graves, J. and Wolvetang, E. J. (2018) 'Platypus Induced Pluripotent Stem Cells: the Unique Pluripotency Signature of a Monotreme', *bioRxiv*, p. 377010. doi: 10.1101/377010.

Wilming, L. G., Boychenko, V. and Harrow, J. L. (2015) 'Comprehensive comparative homeobox gene annotation in human and mouse', *Database*, 2015(2), pp. 1–12. doi: 10.1093/database/bav091.

Wu, L., Wu, Y., Peng, B., Hou, Z., Dong, Y., Chen, K., Guo, M., Li, H., Chen, X., Kou, X., Zhao, Y., Bi, Y., Wang, Y., Wang, H., Le, R., Kang, L. and Gao, S. (2017) 'Oocyte-Specific Homeobox 1, Obox1, Facilitates Reprogramming by Promoting Mesenchymal-to-Epithelial Transition and Mitigating Cell Hyperproliferation', *Stem Cell Reports*, 9(5), pp. 1692–1705. doi: 10.1016/j.stemcr.2017.09.012.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C. Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y. E., Liu, J. Y., Horvath, S. and Fan, G. (2013) 'Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing', *Nature*. Nature Publishing Group, 500(7464), pp. 593–597. doi: 10.1038/nature12364.

Zhang, J. (2003) 'Evolution by gene duplication: An update', *Trends in Ecology and Evolution*, 18(6), pp. 292–298. doi: 10.1016/S0169-5347(03)00033-8.

Zhang, J., Zhang, Y. ping and Rosenberg, H. F. (2002) 'Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey', *Nature Genetics*, 30(4), pp. 411–415. doi: 10.1038/ng852.

Zhong, Y. F. and Holland, P. W. H. (2011) 'The dynamics of vertebrate homeobox gene evolution: Gain and loss of genes in mouse and human lineages', *BMC Evolutionary Biology*, 11(1). doi: 10.1186/1471-2148-11-169.

Ziomek, C. A., Johnson, M. H. and Handyside, A. H. (1982) 'The developmental potential of mouse 16-cell blastomeres', *Journal of Experimental Zoology*, 221(3), pp. 345–355. doi: 10.1002/jez.1402210310.

Chapter 2: Mouse Obox and *Crxos* modulate preimplantation transcriptional profiles revealing similarity between paralogous mouse and human homeobox genes

AH Royall, I Maeso, TL Dunwell, PWH Holland (2018).

Introduction to manuscript:

This chapter has been published as a first author research article in BMC *EvoDevo* (doi.org/:10.1186/s13227-018-0091-4; Appendix A2) with only minor formatting changes prior to inclusion in this thesis (under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/40). The authors on this manuscript in order are Amy H Royall, Ignacio Maeso, Thomas L Dunwell*, Peter WH Holland* (*corresponding authors).

One of the intriguing aspects of ETCHbox genes is that we do not know their function, only that the genes analysed until now are expressed highly and specifically at a vital period of development. Additionally, it has been shown that there are over 60 genes in mouse, but we don't know whether they are all expressed, what all their expression patterns are, or what their targets might be. Relevant functional information is also useful in order to understand how evolutionary pressures affect the highly variable repertoires of ETCHbox genes seen in extant eutherians. In this chapter I focus on unravelling the complex distribution of the mouse ETCHbox genes, and the function of mouse ETCHbox genes using temporal expression analysis and cell-culture followed by transcriptomics. Comparative analysis allowed me to present a hypothesis for how this group of homeobox genes can evolve to be so dynamic.

Authors' contributions

AHR performed molecular and cell biology experiments, transcriptomics and statistical analyses of data. IM undertook genomic annotation of *Obox* genes. AHR, TLD and PWHH designed the ectopic expression experiments and analysis strategy.

AHR and PWHH wrote the manuscript with input from TLD and IM. All authors reviewed and approved the final manuscript.

Availability of data and materials

The RNA sequencing raw datasets generated in this study are available in the NCBI Gene Expression Omnibus repository under accession GSE108060. The mouse developmental stage RNA sequencing data analyzed in this study were obtained from NCBI SRA project SRP034543 (Fan *et al.*, 2015). All transformed data are included in this published article and its supplementary files. All supplementary material, including large datasets not attached to this thesis, are available with the online publication (see section 7).

Acknowledgements and Funding

We acknowledge the Oxford Genomics Centre, Wellcome Trust Centre for Human Genetics, for transcriptome sequencing, and the Flow Cytometry Facility, Experimental Medicine Division, University of Oxford, for assistance with cell sorting. We thank Adam Hargreaves, Daniel Lunn and Monica McAndrews for discussions and advice. Funding was provided through the Oxford Interdisciplinary Bioscience BBSRC Doctoral Training Partnership (BB/M011224/1), the EPA Cephalosporin Fund and a Royal Society International Exchanges grant.

Abbreviations

CMV, cytomegalovirus; DMEM, Dulbecco's Modified Eagle Medium; ECM, extracellular matrix; FACS, Fluorescence-activated cell sorting; FCS, foetal calf serum; FPKM, fragments per kilobase of transcript per million mapped reads; NCBI, National Center for Biotechnology Information; SRA, sequence read archive; ZGA, zygotic genome activation.

Abstract

ETCHbox genes are eutherian-specific homeobox genes expressed during preimplantation development at a time when the first cell lineage decisions are being made. The mouse has an unusual repertoire of ETCHbox genes with several gene families lost in evolution and the remaining two, *Crxos* and *Obox*, greatly divergent in sequence and number. Each has undergone duplication to give a double homeodomain *Crxos* locus and a large cluster of over 60 *Obox* loci. The gene content differences between species raise important questions about how evolution can tolerate loss of genes implicated in key developmental events.

We find that *Crxos* internal duplication occurred in the mouse lineage while *Obox* duplication was stepwise, generating subgroups with distinct sequence and expression. Ectopic expression of three *Obox* genes and a *Crxos* transcript in primary mouse embryonic cells followed by transcriptome sequencing allowed investigation into their functional roles. We find distinct transcriptomic influences for different *Obox* sub-groups and *Crxos*, including modulation of genes related to zygotic genome activation and preparation for blastocyst formation. Comparison with similar experiments performed using human homeobox genes reveals striking overlap between genes downstream of mouse *Crxos* and genes downstream of human *ARGFX*.

Mouse *Crxos* and human *ARGFX* homeobox genes are paralogous rather than orthologous, yet they have evolved to regulate a common set of genes. This suggests there was compensation of function alongside gene loss through co-option of a different locus. Functional compensation by non-orthologous genes with dissimilar

sequences is unusual but may indicate underlying distributed robustness. Compensation may be driven by the strong evolutionary pressure for successful early embryo development.

Keywords

Gene duplication, gene loss, homeodomain, PRD class, *ARGFX*, transcription factor, compensation, blastocyst

1.Introduction

Members of the homeobox superclass are widespread across eukaryotes and their encoded proteins function mainly as transcription factors (Holland, 2013). The proteins contain a DNA-binding domain known as the homeodomain which is highly variable at the primary sequence level. In animals, homeobox genes can be assigned to eleven classes, containing ~100 gene families in humans (Holland, 2013). The largest homeobox gene class in animals is the ANTP class, with the PRD class being second largest. Some homeobox genes are highly conserved throughout animals; others, such as the mammalian X-linked *Rhox* cluster, have been evolving rapidly through duplication and sequence divergence(MacLean *et al.*, 2005). Some homeobox gene families have been lost from certain evolutionary lineages secondarily (Zhong and Holland, 2011; Tsai *et al.*, 2013).

Ten years ago, several PRD class homeobox genes with highly divergent sequences were identified in the human genome (*TPRX1, TPRX2, DPRX, LEUTX* and *ARGFX*); it is now known that these are specific to eutherian mammals and arose through duplication of an Otx family gene *Crx* (Booth and Holland, 2007; Holland, Booth and Bruford, 2007; Zhong and Holland, 2011; Maeso *et al.*, 2016). Each of these genes has expression limited to the germ line and early embryos in humans and collectively they have been named Eutherian Totipotent Cell Homeobox (ETCHbox) genes (Booth and Holland, 2007; Töhönen *et al.*, 2015; Maeso *et al.*, 2016). Ectopic expression of several human ETCHbox genes revealed roles in transcriptional regulation of genes that have a peak of expression in the eight cell embryo and morula stage, around the time of embryo compaction and following

embryonic genome activation in humans (Madissoon *et al.*, 2016; Maeso *et al.*, 2016).

Until recently, these genes were thought to be lost from the mouse genome due to absence of genes with highly similar homeobox sequences to human ETCHbox genes. The genomic regions syntenic to *TPRX1* and *TPRX2,* flanking the *Crx* gene on mouse chromosome 7, contain homeobox genes but these are so distinct in sequence from the human genes that it was hypothesized they had arisen through independent tandem gene duplications in mouse ancestry (Zhong and Holland, 2011; Wilming, Boychenko and Harrow, 2015). Each of these mouse genes, *Crx*-Opposite-Strand (*Crxos*)(Alfano *et al.*, 2005) and Oocyte-Specific Homeobox (*Obox*), have also undergone additional tandem duplication events not evident at the human locus. However, with increased sampling of rodents, the long phylogenetic branch-lengths leading to *Crxos* and *Obox* could be broken and the mouse genes were shown to be highly divergent orthologues of *TPRX1* and *TPRX2* respectively (Maeso *et al.*, 2016). The mouse genome has no orthologous homeobox genes in loci syntenic to any other ETCHbox gene (Figure 2.1; (Maeso *et al.*, 2016). This unique genome organisation is quite different from human and most other mammals, and raises important questions about the function of these genes in the mouse. For example, does extensive sequence divergence of the mouse genes imply they have acquired distinct functions from their human orthologues? What is the functional consequence of secondary duplications experienced by these genes in mice? With loss of *Dprx*, *Leutx*, *Pargfx* and *Argfx* in mice, have any compensatory mechanisms evolved such that *Crxos* or *Obox* genes take over the function of the lost genes?

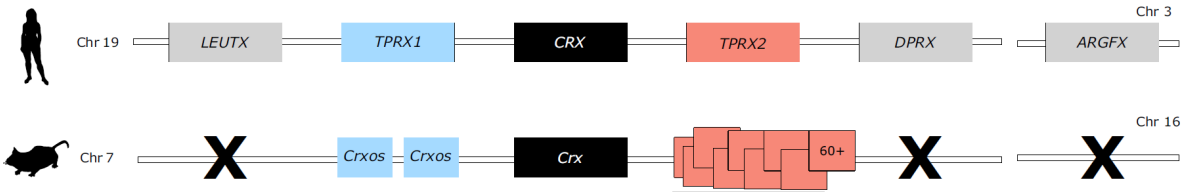


Figure 2.1 ETCHbox genes in human and mouse. The ETCHbox group of homeobox genes has six members: *LEUTX, TPRX1, TPRX2, DPRX, ARGFX* and *PARGFX*. In human, *PARGFX* has been lost in evolution. In mouse, only *Tprx1* and *Tprx2* orthologues remain and are referred to as *Crxos* and *Obox* respectively; each has undergone sequence divergence and gene duplication.

Previous studies of mouse *Crxos* revealed that the locus gives rise to three distinct transcripts: a long transcript encompassing the two duplicate genes (and two homeoboxes), and one transcript from each duplicate. The complete *Crxos* locus has six exons with homeobox sequences spanning exons 2 to 3, and 5 to 6. The long transcript consists of exons 1, 2, 3 (partially), 5 and 6, with the shorter transcripts containing either exons 1, 2 and 3, or exons 4, 5 and 6 (Saito *et al.*, 2009, 2010). Experiments involving ectopic expression in embryonic stem cells followed by quantification of candidate downstream targets have suggested that *Crxos* is involved in cell pluripotency with each shorter variant inhibiting differentiation markers (Saito *et al.*, 2010; Iha *et al.*, 2012; Cui *et al.*, 2016). Furthermore, an RNAi screen has identified *Crxos* as important for correct formation of the inner cell mass in the blastocyst and for hatching and outgrowth from the surrounding zona pellucida (Soma *et al.*, 2012). Another study has suggested that the long isoform is involved in specifying the primitive endoderm lineage (Sato *et al.*, 2015).

The *Obox* genes were initially described as transcripts in unfertilised mouse eggs (Rajkovic *et al.*, 2002) and expression has now been reported throughout early

murine development (Cheng *et al.*, 2007; Ge, 2017). Secondary duplication of *Obox* genes generated multiple loci. In a recent genome assembly (GRCm38/mm10) an array of five *Obox* loci are annotated, *Obox1, Obox2, Obox3, Obox5* and *Obox6*, with an additional locus *Obox7* given 'provisional' status. Previously, *Obox4* was annotated, but this is now listed as a 'partial' and unplaced annotation. Cheng et al., (2007) showed that *Obox1* and *Obox2* have high sequence similarity and suggested they have the same expression pattern in the preimplantation embryo; they also found that *Obox1/2*, *Obox3* and *Obox5* are most highly expressed in the one-cell stage embryo with expression decreasing until no expression is detected in the morula. The *Obox6* locus, in contrast, has elevated expression between the two-cell and morula stages. When *Obox6* was knocked out in mice by homologous recombination, offspring exhibited normal development and were fertile (Cheng *et al.*, 2007). More recent genome exploration has revealed that the *Obox* duplications were far more extensive than previously recognized, and have generated over 60 distinct loci (Maeso *et al.*, 2016). Not all loci have complete homeobox sequences, but it is clear that the diversity of *Obox* genes had been vastly underestimated. This raises new questions regarding the function of the *Obox* cluster in the light of these extensive duplications.

In this study, we use ectopic expression and transcriptomic analysis to investigate the function of mouse ETCHbox genes (*Crxos* and *Obox*). Our goals were to investigate whether these divergent genes have similar or distinct developmental functions to their human homologues, and whether *Crxos* or *Obox* genes have taken over roles associated with ETCHbox genes secondarily lost in mouse evolution. We find that ectopic expression of *Crxos* or *Obox* genes in cultured mouse embryonic fibroblasts induces large transcriptomic changes which can be related to notable

events in the preimplantation embryo. We also argue that *Crxos* functions have evolved to compensate partially for loss of the *Argfx* homeobox gene.

2.Materials and Methods

2.1 Embryonic expression and phylogenetic analyses

To examine endogenous expression patterns, RNA sequencing files were acquired from the SRA database (SRA identifier SRP034543; Supplementary 2.8) and aligned to the GRCm38/mm10 *M. musculus* genome using the STAR alignment tool (Dobin *et al.*, 2013). The Cufflinks tool was used to extract FPKM expression values (Supplementary file 2.9). To investigate evolutionary relationships between genes, maximum likelihood phylogenies were generated from nucleotide or deduced amino acid sequences using RAxML with 500 bootstraps (Stamatakis, 2014).

2.2 Ectopic gene expression

Mouse embryonic fibroblasts (Sciencell #M7540-57) were cultured in Dulbecco's Modified Eagle Medium (DMEM, ThermoFisher #41966-029) supplemented with 10% Foetal Calf Serum (FCS) and 1% pen-strep in 5% CO_2 at 37˚C on 0.001% poly-L-lysine coated plasticware. Coding sequences of *Oboxa1*, *Oboxa4*, *Oboxa7* and *Crxos* were obtained from GenScript and cloned into a vector co-expressing daGFP; genes were C-terminally tagged with V5 and under control of a CMV promoter (Oxford Genetics #OG244). For ectopic expression, 10^6 cells in 100 µl Opti-Mem (Gibco 11058-021) were electroporated (NEPA GENE; poring pulse: 175 volts, 5 ms, 4 pulses) with 10µg of plasmid DNA and seeded in DMEM, 10% FCS, 1% Pen Strep antibiotic (ThermoFisher #15140-122). Media was changed after 24 h and cells collected for FACS sorting after 48 h to enrich for transfected cells. RNA was extracted using the RNeasy micro kit (Qiagen) and quality checked on an Experion electrophoresis station (BIO-RAD). RNA sequencing was performed on three replicates from each experimental condition and control (empty vector

transfection); libraries were prepared using Illumina TruSeq and paired-end RNA sequencing reads generated on the Hi-Seq 4000 platform (Oxford Genomics Centre), yielding 38.2 to 58.2 million reads per sample. Sequence reads were aligned to the GRCm38/mm10 *Mus musculus* genome using the STAR alignment tool, and gene expression levels for protein-coding genes assessed according to NCBI annotations (ftp://ftp.ncbi.nlm.nih.gov/genomes/Mus_musculus/ accessed 13 July 2016, FPKM read-outs in Supplementary file 9). The ectopically expressed genes were expressed at a level (assessed by FPKM) comparable to their embryonic expression levels (Supplementary file 2.1L).

2.3 Differential gene expression analysis

DESeq2 was used to identify genes differentially expressed in response to homeobox gene transfection following FeatureCounts to retrieve raw read counts (Supplementary file 2.10) (Love, Huber and Anders, 2014). Experiments were matched by date to minimise the effects of day-to-day variation. The Cufflinks tool was used to estimate mean FPKM expression values (Supplementary file 2.11). Lists of up- and down-regulated genes for each experimental condition were produced using criteria of $p<0.05$ (with Benjamini-Hochberg correction) and expression fold-change greater than 1.25 or -1.25 (Supplementary file 2.3). Transcripts with an FPKM >2 were considered to be actively expressed.

To generate sets of genes with similar temporal expression profiles across preimplantation mouse development (oocyte, zygote, 2-cell, 4-cell, 8-cell, morula and blastocyst) we took mouse RNAseq data processed as above, filtered to identify genes with a FPKM variance >5 across preimplantation development, and clustered these into profiles using Mfuzz (Kumar and Futschik, 2007; Fan *et al.*, 2015). This

analysis generated 150 initial temporal profiles of gene expression. Sets of profiles with a correlation coefficient of over 0.95 were merged into composite clusters which are identified by IDs >200 (Supplementary files 2.1G and 2.4). When high correlation coefficients were not all reciprocal within a set of profiles, an expression dendrogram was used to guide the process of merging.

2.4 Profile enrichment

Pearson's Chi-square test was used to test the null hypothesis that the number of differentially expressed genes is proportional across all temporal profiles. If the proportions are not equal between profiles ($P<0.05$), Pearson's statistic was used to identify the contribution of each profile to the overall difference. After removal of profiles inferred to contribute to the difference, Fisher's Exact test was used to verify that the differentially expressed genes were proportionally distributed between the remaining profiles and to find statistical difference between occupation of enriched and non-enriched profiles.

3.Results

3.1 Sequence and expression diversity within expanded mouse ETCHbox clusters

The orthologues of both *TPRX1* (*Crxos*) and *TPRX2* (*Obox*) genes are duplicated in mouse. We previously identified 67 *Obox* loci (including pseudogenes) clustered on chromosome 7 in the region syntenic to the *TPRX2* locus in humans (Maeso *et al.*, 2016). To investigate the pathway of duplication and search for groupings of highly similar loci, we conducted phylogenetic analyses using *Obox* nucleotide sequences (Supplementary 2.1A). One locus previously identified was found to not contain a homeodomain or high sequence similarity to annotated *Obox* genes. This sequence was therefore disregarded resulting in a total of 66 loci. Phylogenetic analysis revealed three main categories, named here OboxA, OboxB and OboxD (supplementary 2.1A) and containing 13, 26 and 26 loci respectively, arranged in an interspersed manner (Table 2.1). These include putatively functional loci and pseudogenes. Since there are many more loci than previously named, we propose a new Obox nomenclature system based on DNA sequence (see Appendix A1). We identify one additional locus that does not fit within these main groups, *OboxC*.

Of these 66 *Obox* loci, we suggest that 28 have potential to be translated into protein sequences with a full homeodomain (Supplementary file 2.2). These deduced protein sequences were then used in a second phylogenetic analysis which indicated three subgroups within the OboxA group (Table 2.1; Figure 2.2B; supplementary file 2.1D). Analysis of RNA sequencing data from preimplantation mouse embryos revealed that temporal expression is similar within *Obox* groups or

subgroups, but can be subtly different between them (Figure 2.2B). Expression of the vast majority loci within the OboxD group was not detected in the early embryo, with the exception of *Oboxd10* which shows low level expression (FPKM 4.09) at the 2-cell stage (Table 2.1). Overall, we identified four expression profiles: oocyte to 2-cell (OboxAb), 2-cell to 4-cell (OboxAa, OboxB), 2-cell to 8-cell (OboxC), and 2-cell to blastula (OboxAc) (Figure 2.2B). The observed expression patterns for *Oboxa1*, *Oboxa2*, *Oboxa3*, *Oboxa4* and *Oboxa6* are consistent with previous studies (Rajkovic *et al.*, 2002; Cheng *et al.*, 2007; Ge, 2017). Based on sequence and expression similarity, we hypothesise there is likely to be functional redundancy within, but not usually between, *Obox* groups and subgroups.

Table 2.1 Groupings of mouse *Obox* genes with corresponding expression patterns. *Exact number uncertain because of 100% sequence similarity.

Human orthologue	Group	Sub-group	Members with ORF & HD	Members expressed in embryo	Example	Expression
TPRX2	OboxA	OboxAa	5	5	*Oboxa7*	Two cell to four cell
		OboxAb	4	4	*Oboxa4*	Oocyte to two cell
		OboxAc	1	1	*Oboxa1*	Two cell to morula
	OboxB	OboxB	1	1	*Oboxb2*	Two cell to four cell
	OboxC	OboxC	1	1	*Oboxc*	Two cell to eight cell
	OboxD	OboxD	16	1*	*Oboxd10*	Two cell
TPRX1	Crxos	NA	3	3	*Crxos*	Two cell to blastocyst

The murine gene *Crxos*, orthologous to human *TPRX1*, has duplicated in tandem with the locus generating three distinct transcripts (Figure 2.2A). To date the duplication, we used a conserved domain search to examine the genomic region between *Crx* and *SepW1* loci (the location of *TPRX1*/*Crxos*) for homeobox sequences.

In cow (*Bos taurus*), guinea pig (*Cavia porcellus*), naked mole rat (*Heterocephalus glaber*) and brown rat (*Rattus norvegicus*) only one homeobox was found (figure 2.2A, Supplementary 2.1E), although *R. norvegicus* has an apparently independent duplication of the region containing extra copies of both *Crxos* and *SepW1*. In contrast, using TBLASTN in the Algerian mouse (*Mus spretus*, Jackson laboratory strain #001146*)*, we found two homeobox sequences and regions homologous to all *M. musculus* exons (supplementary 2.1E and 2.1F). These data reveal that the internal duplication to generate a double-homeobox *Crxos* gene occurred in the ancestry of the *Mus* genus after the split of the mouse and rat lineages.

Analysis of RNA sequencing data from mouse preimplantation embryos revealed that the 3' single-homeobox *Crxos* transcript has higher expression than the 5' single-homeobox transcript, although the temporal profile is the same (Figure 2.2C). The 3' *Crxos* transcript has also been previously reported to encode sequences essential for nuclear localisation of the protein (Sato *et al.*, 2013).

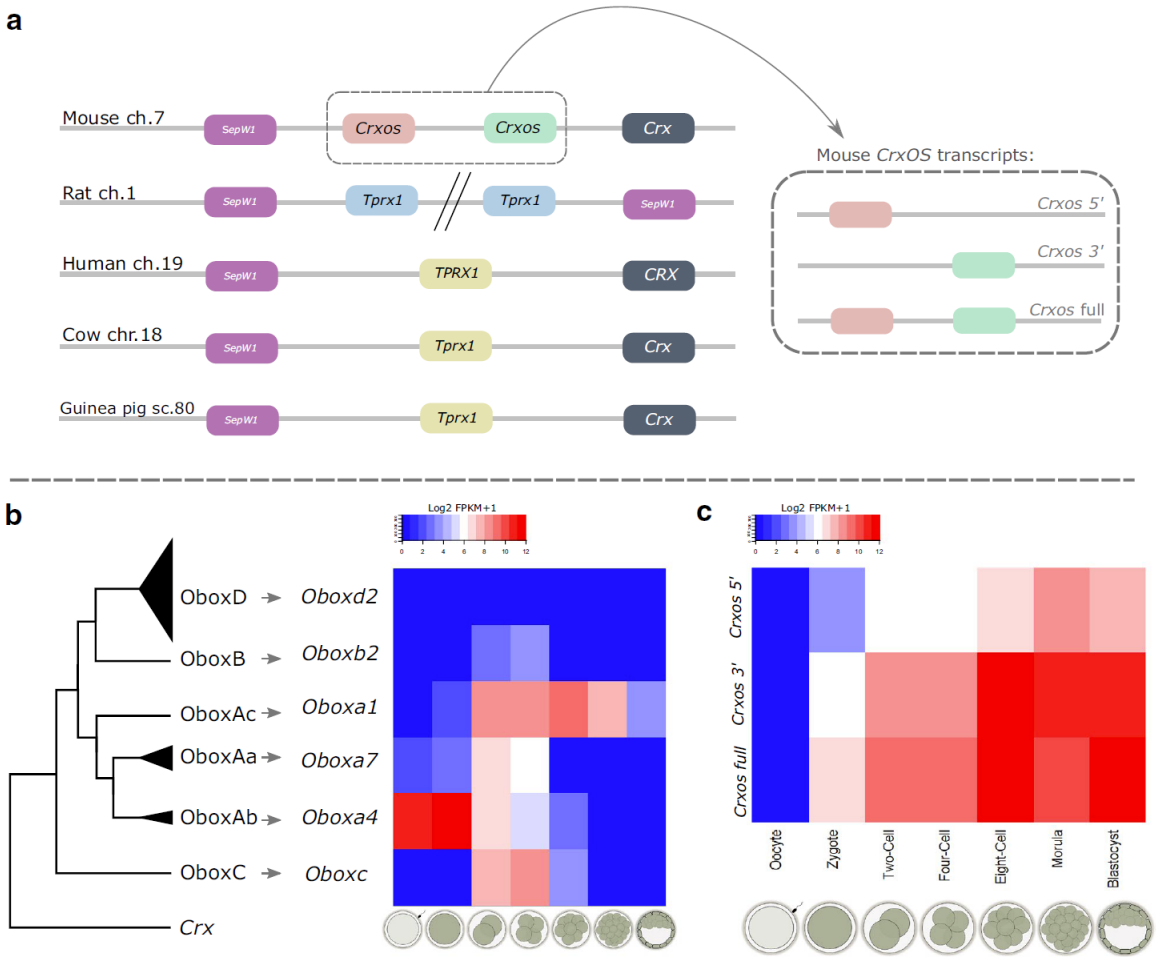


Figure 2.2 ETCHbox organisation and expression in preimplantation development. (A) The *Crxos* locus has undergone a mouse-specific duplication and generates three transcripts: two with a single homeobox and a composite double homeobox transcript. *Crxos* is the *Tprx1* orthologue although originally named as a distinct gene. (B) Analysis of protein and nucleotide sequences of 66 *Obox* genes in mouse identified four groupings; sequences and expression profiles enable OboxA to be split into three sub-groups. (C) The three *Crxos* transcripts have the same preimplantation temporal expression pattern, with the 3' single homeobox transcript generally having higher RNA levels.

3.2 Ectopic expression of *Crxos* and *Obox* genes induces transcriptomic changes

As homeobox genes encode transcription factors, we asked whether ectopic expression of mouse ETCHbox genes could induce transcriptional changes and how such changes are related to possible *in vivo* roles. Due to the complex genomic organisation, we used the above evolutionary analyses to inform the choice of genes. We selected three *Obox* genes with different expression patterns and from different subgroups to test the hypothesis that genes from different subgroups have distinct activities. We also selected the 3' *Crxos* single homeodomain transcript as the most highly expressed splice variant of *Crxos;* this will now be referred to as '*Crxos*' throughout.

We expressed ectopically *Oboxa1* (sub-group OboxAc, *Obox6* in earlier nomenclature), *Oboxa4* (sub-group OboxAb, formerly *Obox1*), *Oboxa7* (sub-group OboxAa) or *Crxos* under a constitutive promoter in primary mouse embryonic fibroblasts, and assayed transcriptome-wide effects through RNA sequencing. In each case, we determined catalogues of genes that were up- or down-regulated 48 hours after ectopic expression (Supplementary file 2.3). These catalogues could include direct and indirect targets. To identify embryonic processes potentially influenced by *Crxos, Oboxa1, Oboxa4*, and *Oboxa7* we examined if each catalogue of up- or down-regulated genes had overlap with defined 'temporal expression profiles' (these are sets of genes grouped on the basis of expression pattern in mouse preimplantation development; shown in Supplementary files 2.1G and 2.4). We also used Gene Ontology (GO terms) to provide further functional information

on groups of genes. The numbers of genes affected significantly by each treatment, and the overlaps between experiments, are shown in Supplementary file 2.1H.

3.2.1 *Crxos* up- and down- regulated profiles

The catalogue of genes down-regulated following *Crxos* ectopic expression is enriched in genes from a single temporal expression profile, number 5 (Fisher's test p=0.0007; Figure 3A, Supplementary file 2.5). Profile 5 contains genes with high levels of mRNA in the oocyte, which decrease as development proceeds (Figure 2.3A). Genes up-regulated by *Crxos* ectopic expression are enriched for genes in profile 59 (Fisher's test p=0.02, supplementary file 2.5) composed of genes with a sharp peak of expression in the two-cell embryo, then lower expression from four-cell to morula, before a second, larger expression peak in blastocyst. There is also enrichment for profile 216, but this profile is not consistent with normal *Crxos* expression so may be off-target. Given that expression of *Crxos* begins in the two-cell embryo and extends to the blastocyst, we suggest that the second, larger, peak of expression in profile 59 includes *in vivo* downstream targets of *Crxos* (Figure 2.3A). Together, these results suggest that when *Crxos* transcription is initiated in the two-cell stage, it serves a role in preparing the embryo for blastocyst formation.

3.2.2 Obox up- and down-regulated profiles

The catalogue of genes up-regulated by ectopic expression of *Oboxa1*, *Oboxa4* or *Oboxa7* showed extensive overlap. Each of these *Obox* genes elicited higher expression of genes found within temporal expression profile 101 (*Oboxa1* p=0.007,

Oboxa4 p=0.005, *Oboxa7* p=0.001; Figure 2.3B, supplementary file 2.5). In mouse development, genes in profile 101 are not expressed during the earliest stages of preimplantation development but show a sharp pulse of expression at the 8-cell stage. These genes are likely to have roles before the earliest cell fate decisions. The finding that three *Obox* genes, but not *Crxos*, elicit increased expression of similar sets of genes may be a consequence of the three closely-related Obox proteins activating target genes through recognition of the same enhancer motifs (Ge, 2017). Thus, after ectopic expression, an *Obox* gene may up-regulate targets usually regulated by a different *Obox* gene. Considering the expression profiles of *Oboxa1*, *Oboxa4* and *Oboxa7*, we suggest that genes within profile 101 are possible *in vivo* downstream targets of *Oboxa1* (OboxAc subclass) or *Oboxa7* (OboxAa subclass).

Considered across all temporal profiles, the three *Obox* genes examined up-regulate a common set of 343 genes and down-regulate a common set of 268 genes (Figure 2.4). GO analysis reveals extracellular matrix ($p = 4.9 \times 10^{-33}$) and focal adhesion ($p= 7.2 \times 10^{-8}$) as enriched functions among the commonly up-regulated genes, and DNA replication ($p = 1.7 \times 10^{-5}$) enriched among the down-regulated genes (Supplementary file 2.1I).

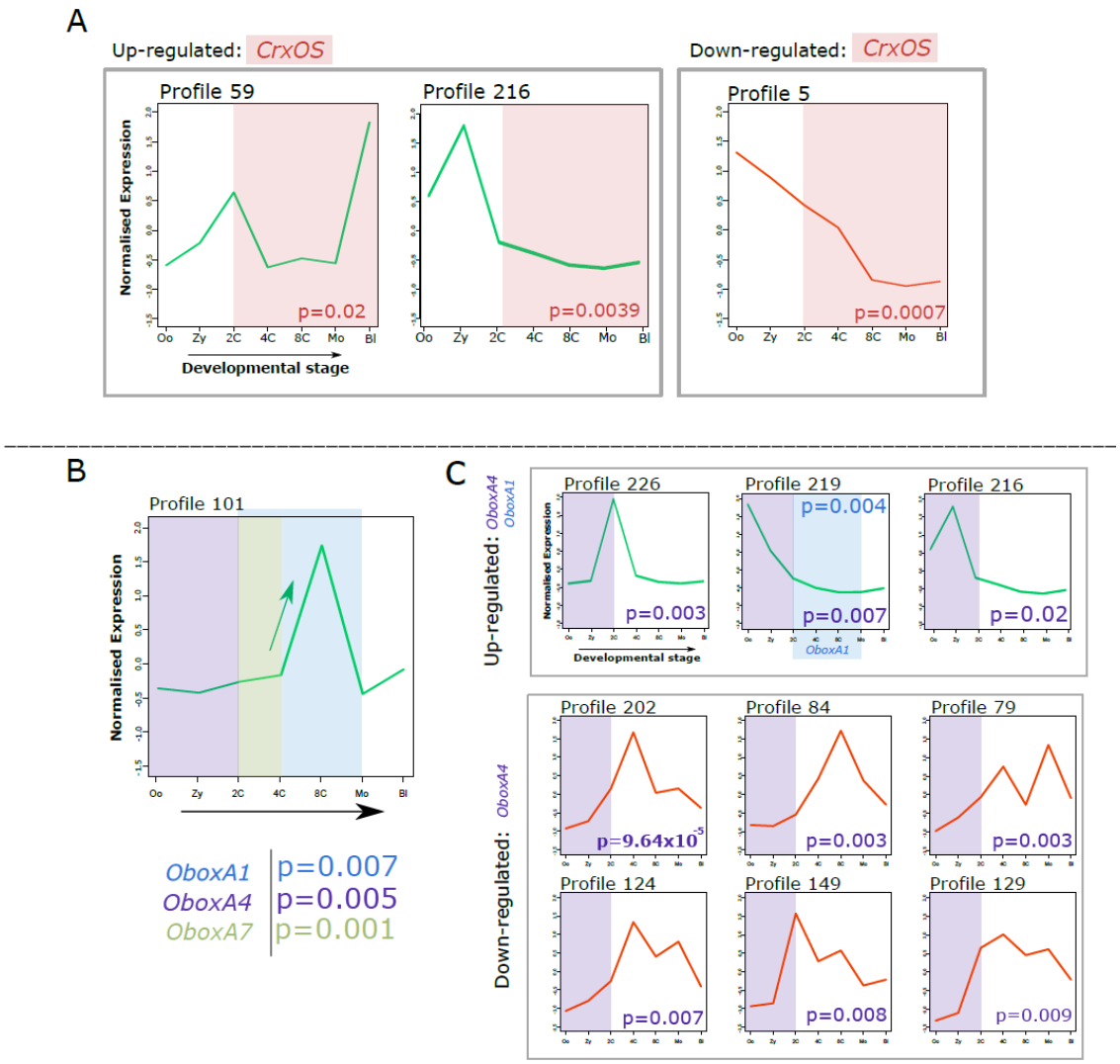


Figure 2.3 Enriched profiles following *Crxos* or *Obox* ectopic expression. Following over-expression of *Crxos*, *Oboxa1, Oboxa4 or Oboxa7* in cultured mouse embryonic cells, we identified genome-wide transcriptomic changes. Each set of up- or down-regulated genes was compared to sets of genes assigned to distinct temporal expression profiles to test for enrichment (Fisher's Exact test, corrected p-values shown). (A) Genes up-regulated following *Crxos* ectopic expression are enriched for two profiles, 59 and 216, although we suggest profile 216 is an off-target effect. Genes down-regulated are enriched for profile 5. Pink shading indicates time of *Crxos* expression. (B) Genes in profile 101 are up-regulated by *Oboxa1*, *Oboxa4* and *Oboxa7*. Blue, purple or green shading indicates time of expression for *Oboxa1*, *Oboxa4* or *Oboxa7* respectively. (C) Genes up-regulated by *Oboxa4* ectopic expression are enriched for two profiles not affected by other *Obox* genes (226, 216) and one profile also affected by *Oboxa1* (219). Genes down-regulated by *Oboxa4* ectopic expression are enriched for six profiles not affected by other *Obox* genes (202, 84, 79, 124, 149, 129). Green curves relate to up-regulated genes; red curves relate to down-regulated genes. Oo = Oocyte, Zy = Zygote, 2C = Two-cell embryo, 4C = Four-cell embryo, 8C = Eight-cell embryo, Mo = Morula, Bl = Blastocyst.

The comparison of all downstream genes also reveals that ectopic expression of *Oboxa4* has the most distinct effects of the three Obox genes tested, affecting more than 600 genes not significantly affected by the other Obox genes (Figure 2.4). The distinctiveness of *Oboxa4* is also evident at the level of temporal profiles. In addition to the shared profile 101, genes up-regulated following ectopic expression of *Oboxa4* were enriched for three further temporal profiles: 216, 219 and 226 (Figure 2.3C, Supplementary file 2.5). The first two of these comprise genes with *in vivo* expression peaks in the zygote and 2-cell embryo. Since *Oboxa4* is predominantly expressed as a maternal transcript, these profiles are consistent with being downstream *in vivo* direct or indirect targets (Figure 2.3C). Genes down-regulated following *Oboxa4* ectopic expression are significantly enriched for six expression profiles, each of which has an mRNA peak after zygotic genome activation (ZGA; (Palfy, Shai and Vastenhouw, 2017)): profiles 202, 84, 79, 124, 149, 129; collective Fisher's test $p = 2.7 \times 10^{-13}$ (Figure 2.3B, supplementary file 2.5). These profiles, therefore, are consistent with being negatively-regulated downstream *in vivo* targets, with maternal *Oboxa4* suppressing expression until after ZGA. A gene ontology (GO) analysis of the regulated genes in these profiles shows enrichment for ribosome biogenesis functions ($P<0.001$). Profile 219 is also enriched in the genes up-regulated downstream of *Oboxa1*, although considering the temporal expression of *Oboxa1* in the embryo this is more likely to be an *in vivo* target of *Oboxa4* (Figure 2.3C).

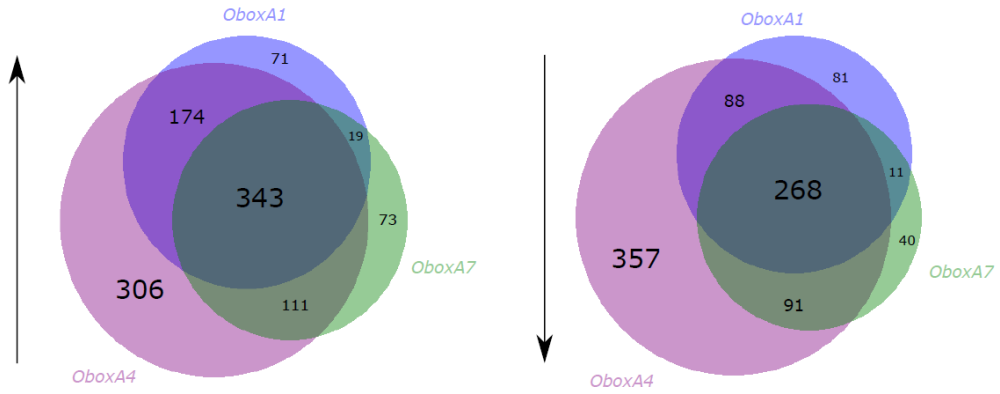


Figure 2.4 Overlap between genes downstream of Obox genes. Comparison of genes up- or down-regulated following ectopic expression of Obox genes; ectopic expression of the maternally expressed *Oboxa4* gene affects expression of additional downstream genes not affected by *Oboxa1* or *Oboxa7* expression.

3.3 Comparison between human and mouse ETCHbox gene functions

The evolutionary loss of *Argfx*, *Leutx* and *Dprx* ETCHbox genes in mouse, and duplication of the remaining *Tprx1* and *Tprx2* genes, raises questions about functional similarities and differences between mouse and human. We compared the sets of genes significantly up- and down-regulated after ectopic expression of human *TPRX1*, *ARGFX*, *DPRX* and *LEUTX* (Maeso *et al.*, 2016) with each of the gene sets up- or down-regulated by mouse *Crxos*, *Oboxa1*, *Oboxa4,* and *Oboxa7*. Two of these comparisons revealed striking overlap (Figure 2.5A; Supplementary files 2.6 and 2.7). We found that for 96 of the genes up-regulated by *ARGFX* in human cells, their mouse orthologues were also up-regulated by *Crxos* in mouse cells (Figure 2.5B; Fisher's test p = $3x10^{-52}$; Supplementary file 2.7). Similarly, the set of genes down-regulated by *ARGFX* in human cells showed significant similarity to the set of genes down-regulated by *Crxos* in mouse cells (Figure 2.5B; 125 one-to-one orthologues, p = 2.4 $x10^{-102}$; Supplementary file 2.7). Of these, 98% of the jointly

down-regulated orthologues are expressed in the mouse preimplantation embryo, as are 50% of the jointly up-regulated genes. These data suggest close similarity of function between human *ARGFX* and mouse *Crxos*, despite these being non-orthologous proteins. The direct orthologue of mouse *Crxos* is human *TPRX1,* not *ARGFX*; these orthologous genes seem to have contrasting functions following ectopic expression since many genes down-regulated by *Crxos* over-expression are up-regulated by *TPRX1* ($p= 1.58^{-36}$; Supplementary file 2.6). There are also additional significant overlaps observed between genes (figure 2.5A), suggesting further overlapping functions of human and mouse ETCHbox genes.

3.4 Targets with strong response are associated with embryonic milestones

The analysis of function using temporal gene profiles pays equal attention to genes with low and high expression, and does not distinguish between mildly or strongly up- and down-regulated targets. We therefore asked which target genes were most strongly affected by ectopic expression of mouse *Crxos* and *Obox* genes (highest fold change up-regulated in Table 2.2; highest fold change down-regulated Supplementary file 2.1J). Three of the 10 most strongly up-regulated genes have functions related to the extracellular matrix (ECM). Similarly, when ordered by fold-change, three of the genes up-regulated most strongly following *Oboxa1* and *Oboxa4* ectopic expression are ECM components or function as mediators of ECM communication: *Col8a2 (Oboxa1* and *Oboxa4), Eln (Oboxa1* and *Oboxa4), Egfl6 (Oboxa1)* and *Aoc3 (Oboxa4).*

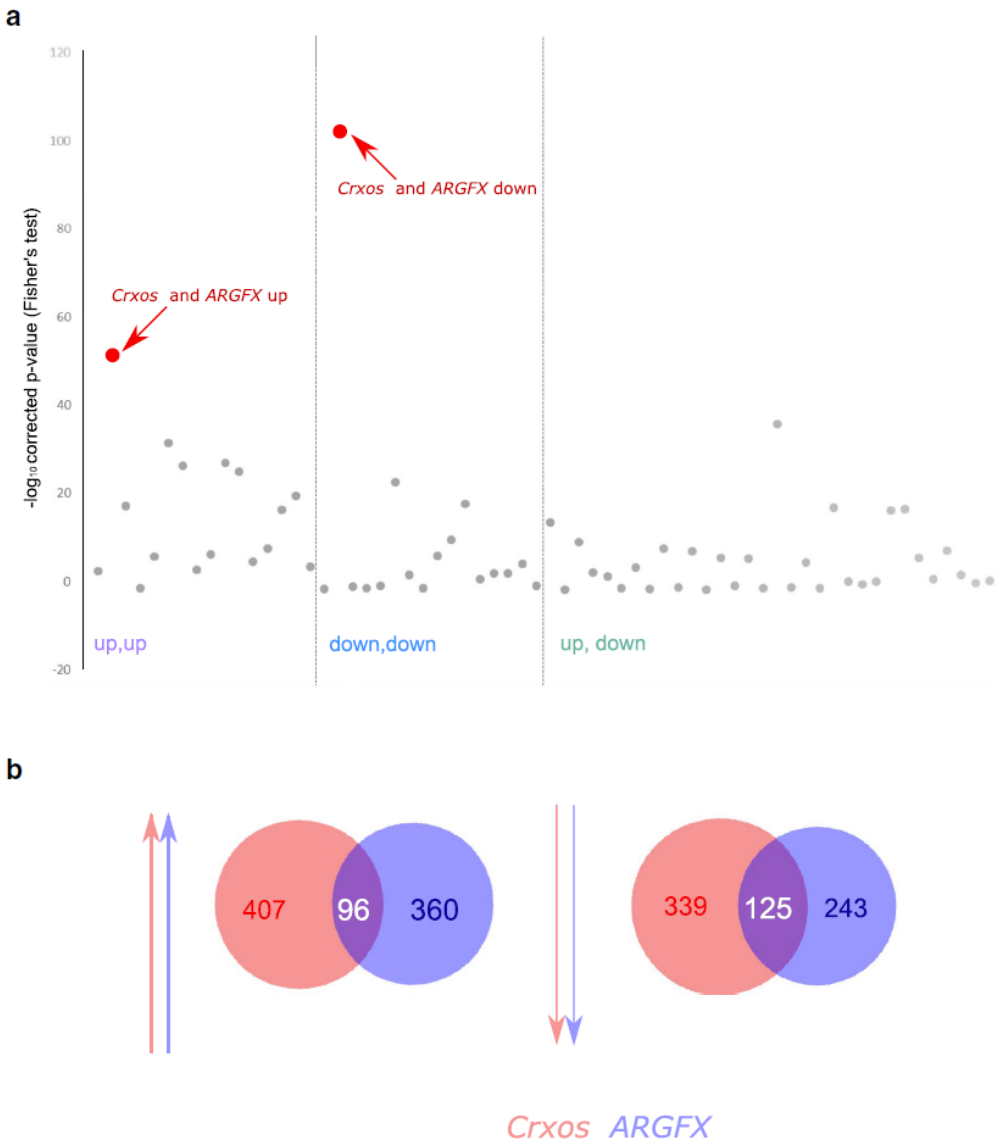


Figure 2.5 Similarity between genes downstream of human *ARGFX* and mouse *Crxos*. (A) The sets of genes up- or down-regulated following ectopic expression are compared between each mouse and human ETCHbox gene after filtering for one-to-one orthologues (human data from (Maeso *et al.*, 2016); y-axis shows -log(p-values) derived from pairwise Fisher's Exact Test. Many comparisons are significant; the most striking similarities are between downstream targets of mouse *Crxos* and human *ARGFX* (up-regulated genes $P = 3 \times 10^{-52}$; down-regulated genes $P = 4.2 \times 10^{-102}$). Order of comparisons, from left to right, given in supplementary File 2.6. (B) Proportional Venn diagrams showing extent of overlap between human and mouse one-to-one orthologues affected by *CrxOS* ectopic expression in primary mouse embryonic fibroblasts and *ARGFX* in primary adult human fibroblasts.

Table 2.2 Genes with highest up-regulated expression fold-change following ectopic expression of mouse ETCHbox genes.

	Oboxa1		***Oboxa4***		***Oboxa7***		***Crxos***	
Rank	**Gene**	**Log_2 Fold Change**	**Gene**	**Log_2 Fold Change**	**Gene**	**Log_2 Fold Change**	**Gene**	**Log_2 Fold Change**
1	*Lgr5*	2.02	*Megf6*	2.24	*Fmo1*	2.45	*Apod*	1.85
2	*Megf6*	1.82	*Lgr5*	1.91	*Megf6*	1.65	*Serpina3g*	1.64
3	*Myh1*	1.66	*Myh1*	1.87	*Myh11*	1.58	*Tnn*	1.56
4	*Myh11*	1.54	*Eln*	1.78	*Egfl6*	1.57	*Megf6*	1.53
5	*Col8a2*	1.48	*Gja5*	1.78	*Cdo1*	1.55	*Cxcl14*	1.53
6	*Eln*	1.47	*Gdpd2*	1.71	*Rbp3*	1.44	*Mfap4*	1.53
7	*Gja5*	1.38	*Col8a2*	1.71	*Apoe*	1.39	*Dpt*	1.52
8	*Egfl6*	1.38	*Gucy1a3*	1.69	*Chit1*	1.35	*Egfl6*	1.49
9	*Itga11*	1.36	*Aoc3*	1.68	*Gja5*	1.32	*Ch25h*	1.48
10	*Epha3*	1.36	*Hmcn1*	1.64	*Slpi*	1.31	*Serping1*	1.43

The most strongly down-regulated genes have functions in processes distinct from the up-regulated genes (Supplementary file 2.1J). Amongst the top 10 down-regulated genes following *Oboxa1* ectopic expression is *Sox2*, a transcription factor involved in specification of the inner cell mass of the blastocyst, and deployed in iPS cell production as one of the Yamanaka factors (Takahashi and Yamanaka, 2006). Similarly, a gene strongly down-regulated after *Crxos* ectopic expression is implicated in trophectoderm specification and development, *Slco2a1* (Log_2 fold change of -0.94, (Assou *et al.*, 2012; Reinholt *et al.*, 2017)). The effect of ectopic ETCHbox expression on other genes with known roles in preimplantation development is given in supplementary file 2.1K.

4.Discussion

Mammalian preimplantation development encompasses several processes common to a wide range of species, such as generation of a hollow blastocyst and implantation into maternal endometrial tissue. Mice have long been used judiciously as models for human embryogenesis (Sulik, Johnston and Webb, 1981; Juriloff and Harris, 2000; Rossant and Cross, 2001; Wang and Dey, 2006), with caution urged about extrapolation between species (Ménézo and Hérubel, 2002). Furthermore, consistent with the hourglass model which describes propensity for evolutionary change early and late in development, transcriptomic analyses have highlighted variation in the earliest stages of development between different mammalian species (Kuijk, Geijsen and Cuppen, 2015; Breschi, Gingeras and Guigó, 2017). Similarly, several of the ETCHbox genes implicated in regulation of human preimplantation gene expression have been secondarily lost in mice (Maeso *et al.*, 2016). These findings raise questions about how far pathways and regulatory networks discovered in mice can be applied to human preimplantation development, and vice versa. We have studied the mouse ETCHbox genes which, like their human orthologues, are expressed specifically in the preimplantation embryo but which have very different numbers and genomic composition to humans. Extensive gene loss, sequence divergence and duplication of the remaining genes in mouse allowed us to investigate how differences at the genomic level relate to species-specific differences or similarities in preimplantation development.

In mouse, four of the six ancestral ETCHbox gene families have been lost leaving orthologues of just *Tprx1* and *Trpx2*. The mouse *Crxos* gene is the orthologue of human *TPRX1*. *Crxos* has duplicated and is processed to give three transcripts

with common temporal expression profiles in the preimplantation mouse embryo. The 66 *Obox* loci are orthologous to human *TPRX2* and can be divided into four groups and six sub-groups, with expression profiles mirroring molecular phylogenetic classification. The concordance between sequence and expression suggests there may be functional redundancy within *Obox* subgroups during preimplantation mouse development.

To investigate downstream activities and functional similarity to human genes, we ectopically expressed three *Obox* genes from different subgroups and the most highly expressed *Crxos* transcript in mouse embryonic fibroblasts. It could be argued that these cell types are very different from cells of the preimplantation embryo, but to conduct evolutionarily meaningful comparisons with data previously obtained using human fibroblasts (Maeso *et al.*, 2016) it is important to use comparable cell types. Additionally, ectopic expression of ETCHbox genes in primary cells is comparable to reprogramming experiments, which commonly use primary fibroblasts as the initial cell population.

We found that each of the genes elicited transcriptomic changes related to gene expression profiles of preimplantation stages, suggesting we have partially recapitulated the *in vivo* roles of mouse ETCHbox genes. For example, we find that over-expression of *Crxos* modifies the transcriptome of embryonic fibroblasts to partially mimic the transcriptome of the blastocyst, down-regulating transcripts that fall in abundance during cleavage stages and up-regulating genes with a blastocyst peak. These results are exciting since the blastocyst is composed of the descendent cell types from the first cell lineage decision: trophectoderm and pluripotent inner cell mass. It is also the time when the embryonic secretome

communicates with maternal tissue before subsequent blastocyst invasion into the endoderm (Wang and Dey, 2006). A role in formation of these cell types is emphasised by the finding that orthologues of the genes most strongly up-regulated by *Crxos* are implicated in implantation in human development: *Apod* and *Serpina3* (Table 2.2).

Similarly, over-expression of *Oboxa4* caused transcriptomic changes mirroring those of the early embryo. *Oboxa4* is expressed maternally in the oocyte and high levels of RNA are detected in the zygote. Ectopic expression of this gene caused down-regulation of genes that are not detectable as RNA in the earliest developmental stages, and up-regulation of those that are. In mouse development, zygotic genome activation (ZGA) is initiated earlier than in humans, with an early major wave of activation at the 2-cell stage and further waves of activation occurring until the morula stage (Xue *et al.*, 2013). The expression profiles of genes downstream of *Oboxa4* suggest that it may have a role in delaying or suppressing expression of embryonic genes, affecting timing of ZGA (Figure 2.6).

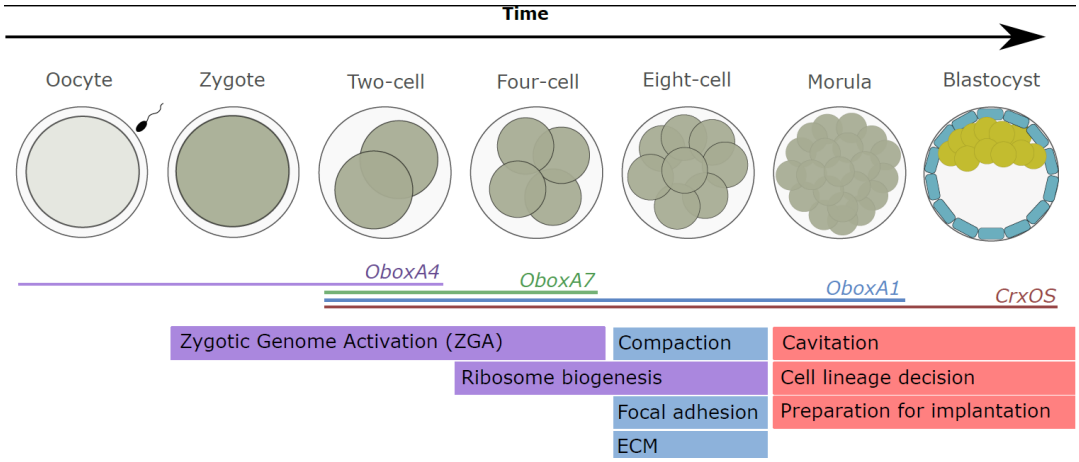


Figure 2.6 Cellular and embryonic processes potentially regulated by mouse ETCHbox genes. Global transcriptomic changes elicited by ectopic expression, and the embryonic expression profiles of ETCHbox genes themselves, highlights possible developmental milestones regulated by ETCHbox genes. We suggest that *Crxos* is involved in preparing the embryo for the first cell fate decision prior to the early blastocyst stage. *Obox* genes likely regulate a range of biological processes *in vivo*; we suggest *Oboxa4* is involved with early milestones including induction of zygotic gene expression, whereas *Oboxa1* and *Oboxa7* are involved in later events such as embryo compaction. Expression of the ETCHbox genes tested are represented by coloured lines (*Crxos* red, *Oboxa1* blue, *Oboxa4* purple, *Oboxa7* green).

For all *Obox* genes tested, we detect evidence of regulation of genes expressed in the 8-cell embryo, and thus substantially earlier than the effect driven by *Crxos* ectopic expression. However, because this effect is seen for three different *Obox* genes, this may partially reflect cross-regulation between targets of closely related genes. On the basis of the normal expression profiles of *Obox* genes, we suggest the common effect is most likely to reflect *in vivo* function of *Oboxa1* or *Oboxa7* (Table 2.1). The 8-cell stage of mouse development, when many putative *Obox* target genes are expressed, is a particularly interesting developmental period. At this stage, the cells of the embryo increase their cell-cell contacts and elongate, and the embryo compacts in a critical morphological change which prepares the embryo for implantation (Figure 2.6)(Johnson and Maro, 1986; Landry *et al.*, 2006; Fierro-González *et al.*, 2013). Functions of genes affected include interactions with the ECM, adhesion to the external environment and DNA replication.

One of the most striking findings of this study was the discovery of highly significant overlap between the set of genes up- and down-regulated by ectopic mouse ETCHbox genes with the set of genes affected by ectopic expression of human ETCHbox genes. We find many comparisons show significant overlaps (Figure 2.5A), which may reflect widespread overlapping functions between the majority of ETCHbox genes. Most notably, many of the inferred downstream targets of *Crxos* in mouse are orthologous to the inferred downstream targets of *ARGFX* in human (96 orthologues commonly up-regulated, 125 commonly down-regulated). Much smaller overlap was detected for other pairs of mouse and human ETCHbox genes. There are differences between the biological activities of the two genes, however. For example, expressing *ARGFX* in human fibroblasts induced transcriptional changes that mirror a pulse of expression in the 8-cell human embryo (Maeso *et al.*,

2016). In contrast, *Crxos*-induced transcriptional changes more closely parallel changes occurring at the blastocyst stage. Furthermore, *Crxos* has broader expression in mouse preimplantation development than does *ARGFX* in human. Together, these data suggest that before or after the loss of *ARGFX* in murid evolution, the *Crxos* gene took over roles originally undertaken by *ARGFX,* and has also acquired additional targets and biological functions (Figure 2.7). It is particularly intriguing that this functional compensation involved deployment of a paralogous rather than an orthologous homeobox gene. Indeed, *Crxos* seems to have contrasting transcriptomic effects to its direct orthologue *TPRX1*, with significant overlap between genes down-regulated by *Crxos* and those up-regulated by *TPRX1* ($p= 1.58^{-36}$, supplementary file 2.7).

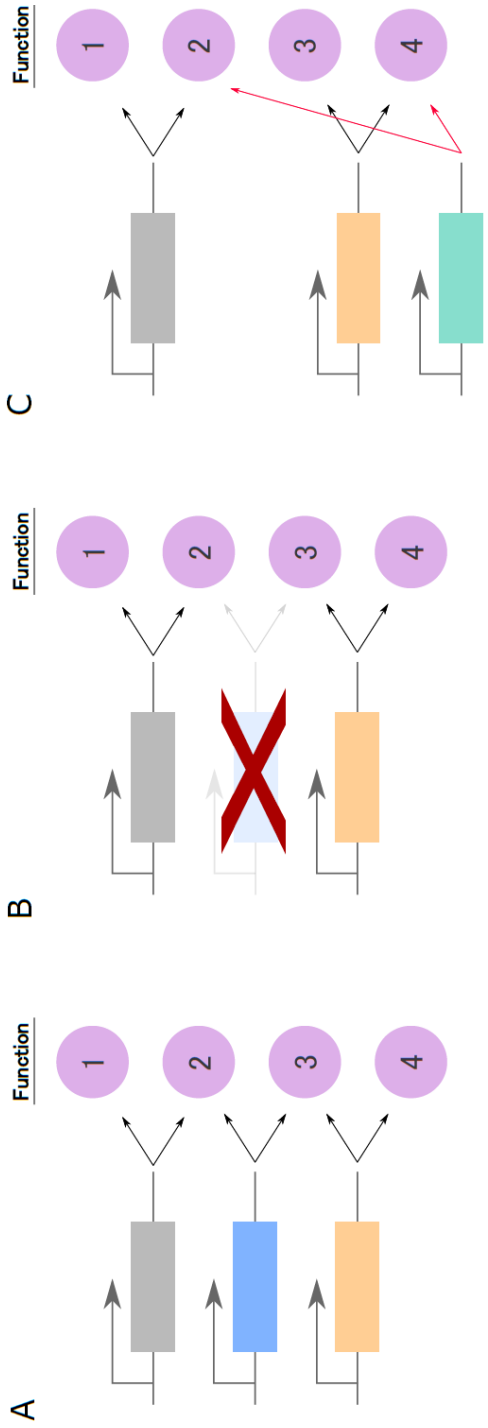


Figure 2.7 Simplified scenario for the evolution of functional redundancy. (A) A group of genes contribute to several biological processes with partial overlap of functions. (B) Partial functional overlap permits gene loss to persist as a temporary state. (C) Redundancy can be restored through gene duplication and divergence. The temporal sequence of (B) and (C) can be reversed. Partial redundancy affords buffering against somatic failure of a gene or process.

The convergence of function between human and mouse ETCHbox genes provides an intriguing example of compensation and refinement of gene function alongside dynamic lineage specific gene loss and expansions (Figure 2.7). Such compensation would be favoured if the system displays 'distributed robustness' as defined by Wagner (Wagner, 2005). Under this model, individual genes may have overlapping rather than redundant functions, such that collectively the system has multiple routes to achieve the same endpoints. Extrapolating, we suggest that a variable set of ETCHbox genes regulates a suite of functions in eutherian preimplantation embryos that are shared between species, regardless of the particular repertoire of ETCHbox genes retained in a genome.

The evolutionary history of ETCHbox genes, with extensive gene duplication, loss and sequence change during mammalian radiation, contrasts sharply to that of most other homeobox genes. We show that ectopic expression of these genes in mouse embryonic fibroblasts elicits transcriptomic changes that mirror transcriptomic changes in early development. For example, the maternally-expressed *Oboxa4* gene can reduce expression of genes normally active after ZGA, other *Obox* genes elevate expression of genes that peak at the 8-cell stage, and *Crxos* expression causes transcriptomic changes mirroring pre-blastocyst development. Most strikingly, we find close similarity between the gene downstream of mouse *Crxos* and human *ARGFX*, despite these not being orthologous homeobox genes. These results point to an evolutionarily labile system where a set of regulatory genes can duplicate and be lost in evolution, but shift functionally to compensate for gene loss. Functional replacement by a paralogous gene might be favoured when

evolution builds an inherently redundant system to ensure the robustness of critical developmental events.

5.References

Alfano, G. *et al.* (2005) 'Natural antisense transcripts associated with genes involved in eye development', 14(7), pp. 913–923. doi: 10.1093/hmg/ddi084.

Assou, S. *et al.* (2012) 'Transcriptome Analysis during Human Trophectoderm Specification Suggests New Roles of Metabolic and Epigenetic Genes', *PLoS ONE*, 7(6). doi: 10.1371/journal.pone.0039306.

Booth, H. A. F. and Holland, P. W. H. (2007) 'Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line', *Gene*, 387(1–2), pp. 7–14. doi: 10.1016/j.gene.2006.07.034.

Breschi, A., Gingeras, T. R. and Guigó, R. (2017) 'Comparative transcriptomics in human and mouse', *Nature Reviews Genetics*. Nature Publishing Group, 18(7), pp. 425–440. doi: 10.1038/nrg.2017.19.

Cheng, W. C. *et al.* (2007) 'Mice lacking the Obox6 homeobox gene undergo normal early embryonic development and are fertile', *Developmental Dynamics*, 236(9), pp. 2636–2642. doi: 10.1002/dvdy.21261.

Cui, W. *et al.* (2016) 'Towards Functional Annotation of the Preimplantation Transcriptome: An RNAi Screen in Mammalian Embryos', *Scientific Reports*, 6. doi: 10.1038/srep37396.

Dobin, A. *et al.* (2013) 'STAR: Ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Fan, X. *et al.* (2015) 'Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos', *Genome Biology*. Genome Biology, pp. 1–17. doi: 10.1186/s13059-015-0706-1.

Fierro-González, J. C. *et al.* (2013) 'Cadherin-dependent filopodia control preimplantation embryo compaction', *Nature Cell Biology*. Nature Publishing Group, 15(12), pp. 1424–1433. doi: 10.1038/ncb2875.

Ge, S. X. (2017) 'Exploratory bioinformatics investigation reveals importance of " junk " DNA in early embryo development'. BMC Genomics, pp. 1–19. doi: 10.1186/s12864-017-3566-0.

Holland, P. W. H. (2013) 'Evolution of homeobox genes', *Wiley Interdisciplinary Reviews: Developmental Biology*, 2(1), pp. 31–45. doi: 10.1002/wdev.78.

Holland, P. W. H., Booth, H. A. F. and Bruford, E. A. (2007) 'Classification and nomenclature of all human homeobox genes', *BMC Biology*, 5, pp. 1–28. doi: 10.1186/1741-7007-5-47.

Iha, M. *et al.* (2012) 'Effect of ectopic expression of homeoprotein EGAM1C on the cell morphology, growth, and differentiation in a mouse embryonic stem cell line,

MG1.19 cells', *Reproduction*, 143(4), pp. 477–489. doi: 10.1530/rep-11-0379.

Johnson, M. H. and Maro, B. (1986) 'The role of cell adhesion in the synchronization and orientation of polarization in 8-cell mouse blastomeres', 255, pp. 239–255.

Juriloff, D. M. and Harris, M. J. (2000) 'Mouse models for neural tube closure defects', 9(6).

Kuijk, E., Geijsen, N. and Cuppen, E. (2015) 'Pluripotency in the light of the developmental hourglass', *Biological Reviews*, 90(2), pp. 428–443. doi: 10.1111/brv.12117.

Kumar, L. and Futschik, M. (2007) 'Bioinformation Mfuzz : A software package for soft clustering of microarray data Bioinformation', pp. 5–7.

Landry, D. *et al.* (2006) 'Hypocellularity and Absence of Compaction as Criteria for Embryonic Deathle', *Regenerative Medicine*, 1(3), pp. 367–371. doi: 10.2217/17460751.1.3.367.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', pp. 1–21. doi: 10.1186/s13059-014-0550-8.

MacLean, J. A. *et al.* (2005) 'Rhox: A new homeobox gene cluster', *Cell*, 120(3), pp. 369–382. doi: 10.1016/j.cell.2004.12.022.

Madissoon, E. *et al.* (2016) 'Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos', *Scientific Reports*. Nature Publishing Group, 6(July), pp. 1–14. doi: 10.1038/srep28995.

Maeso, I. *et al.* (2016) 'Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals', *BMC Biology*, 14(1), pp. 1–14. doi: 10.1186/s12915-016-0267-0.

Ménézo, Y. J. R. and Hérubel, F. (2002) 'Mouse and bovine models for human IVF.', *Reproductive biomedicine online*, 4(2), pp. 170–175. doi: 10.1016/S1472-6483(10)61936-0.

Palfy, M., Shai, J. and Vastenhouw, N. (2017) 'The timing of Zygotic genome activation', *Current Opinion in Genetics and Development*, 43, pp. 53–60. doi: 10.1016/j.gde.2016.12.001.

Rajkovic, A. *et al.* (2002) 'Obox, a family of homeobox genes preferentially expressed in germ cells', *Genomics*, 79(5), pp. 711–717. doi: 10.1006/geno.2002.6759.

Reinholt, B. M. *et al.* (2017) 'General and Comparative Endocrinology Tissue organization alters gene expression in equine induced trophectoderm cells', *General and Comparative Endocrinology*. Elsevier Inc., 247, pp. 174–182. doi: 10.1016/j.ygcen.2017.01.030.

Rossant, J. and Cross, J. C. (2001) 'Placental Development: Lessons From Mouse Mutants', *Nature Reviews Genetics*, 2, pp. 538–548.

Saito, K. *et al.* (2010) 'Cloning of Complementary DNAs Encoding Structurally Related Homeoproteins from Preimplantation Mouse Embryos: Their Involvement in the Differentiation of Embryonic Stem Cells', *Biology of Reproduction*, 82(4), pp. 687–697. doi: 10.1095/biolreprod.108.075697.

Saito, R. *et al.* (2009) 'CrxOS maintains the self-renewal capacity of murine embryonic stem cells', *Biochemical and Biophysical Research Communications*, 390(4), pp. 1129–1135. doi: http://dx.doi.org/10.1016/j.bbrc.2009.09.118.

Sato, S. *et al.* (2013) 'Intact structure of EGAM1 homeoproteins and basic amino acid residues in the common homeodomain of EGAM1 and EGAM1C contribute to their nuclear localization in mouse embryonic stem cells p p', *Journal of Bioscience and Bioengineering*. Elsevier Ltd, 116(2), pp. 141–146. doi: 10.1016/j.jbiosc.2013.02.007.

Sato, S. *et al.* (2015) 'Partial inhibition of differentiation associated with elevated protein levels of pluripotency factors in mouse embryonic stem cells expressing exogenous EGAM1N homeoprotein', *Journal of Bioscience and Bioengineering*, 120(5), pp. 562–569. doi: http://dx.doi.org/10.1016/j.jbiosc.2015.03.001.

Soma, M. *et al.* (2012) 'Preferential emergence of cell types expressing markers for primitive endoderm lineages in mouse embryonic stem cells expressing exogenous EGAM1 homeoprotein', *Journal of Bioscience and Bioengineering*, 114(3), pp. 342–346. doi: http://dx.doi.org/10.1016/j.jbiosc.2012.04.012.

Stamatakis, A. (2014) 'RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies', 30(9), pp. 1312–1313. doi: 10.1093/bioinformatics/btu033.

Sulik, K., Johnston, M. and Webb, M. (1981) 'Fetal Alcohol Syndrome : Embryogenesis in', *Science*, 214(4523), pp. 936–939. doi: 10.1126/science.6795717.

Takahashi, K. and Yamanaka, S. (2006) 'Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors', *Cell*, 126(4), pp. 663–676. doi: 10.1016/j.cell.2006.07.024.

Töhönen, V. *et al.* (2015) 'Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development', *Nature Communications*, 6, pp. 1–9. doi: 10.1038/ncomms9207.

Tsai, I. J. *et al.* (2013) 'The genomes of four tapeworm species reveal adaptations to parasitism', *Nature*, 496(7443), pp. 57–63. doi: 10.1038/nature12031.

Wagner, A. (2005) 'Distributed robustness versus redundancy as causes of

mutational robustness', *BioEssays*, 27(2), pp. 176–188. doi: 10.1002/bies.20170.

Wang, H. and Dey, S. K. (2006) 'Roadmap to embryo implantation: Clues from mouse models', *Nature Reviews Genetics*, 7(3), pp. 185–199. doi: 10.1038/nrg1808.

Wilming, L. G., Boychenko, V. and Harrow, J. L. (2015) 'Comprehensive comparative homeobox gene annotation in human and mouse', *Database*, 2015(2), pp. 1–12. doi: 10.1093/database/bav091.

Xue, Z. *et al.* (2013) 'Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing', *Nature*. Nature Publishing Group, 500(7464), pp. 593–597. doi: 10.1038/nature12364.

Zhong, Y. F. and Holland, P. W. H. (2011) 'The dynamics of vertebrate homeobox gene evolution: Gain and loss of genes in mouse and human lineages', *BMC Evolutionary Biology*, 11(1). doi: 10.1186/1471-2148-11-169.

6. Supplementary files

2.1

(A) *Obox* phylogeny using nucleotide sequences.

(B) Genomic organisation of *Mus musculus* ETCHbox genes

(C) Proposed new *Obox* nomenclature. The term 'partial' indicates an Obox sequence lacking a complete homeobox and the term is not part of proposed nomenclature.

(D) *Obox* phylogeny using protein sequences that contain a complete homeodomain

(E) Summary of number of homeoboxes in *Tprx1* syntenic region in different species. * In rat, the two homeobox sequences reflect an independent regional duplication

(F) Alignment of *M. musculus* and *M. spretus Tprx1*/*Crxos* homeodomains

(G) Normalised expression graphs of each temporal profile generated by Mfuzz

(H) Venn diagrams shown numbers of genes affected significantly by each treatment, and overlaps between experiments

(I) Enriched GO terms for genes affected by *Obox* ectopic expression

(J) Genes most strongly down-regulated following ectopic expression

(K) Effect of ectopic ETCHbox expression on selected genes with notable preimplantation or stem cell functions

(L) In vivo expression levels of ETCHbox genes at mouse preimplantation stages compared to ectopic expression levels of transfected genes

2.2 Nucleotide and protein sequences used for Obox phylogenetic analyses

2.3 Lists of genes significantly up- or down-regulated following ectopic expression

2.4 Mfuzz clusters and genes within each cluster. Additional

2.5 Observed and expected numbers of significantly regulated genes populating temporal expression profiles used for enrich- ment analysis. Additional

2.6 List of one-to-one orthologues similarly regulated fol- lowing ectopic expression of human ARGFX and mouse Crxos. Additional

2.7 Data used for statistical tests of overlap between sets of one-to-one orthologues regulated by mouse and human ETCHbox genes

2.8 SRA data information used to generate embryonic gene expression data. Additional

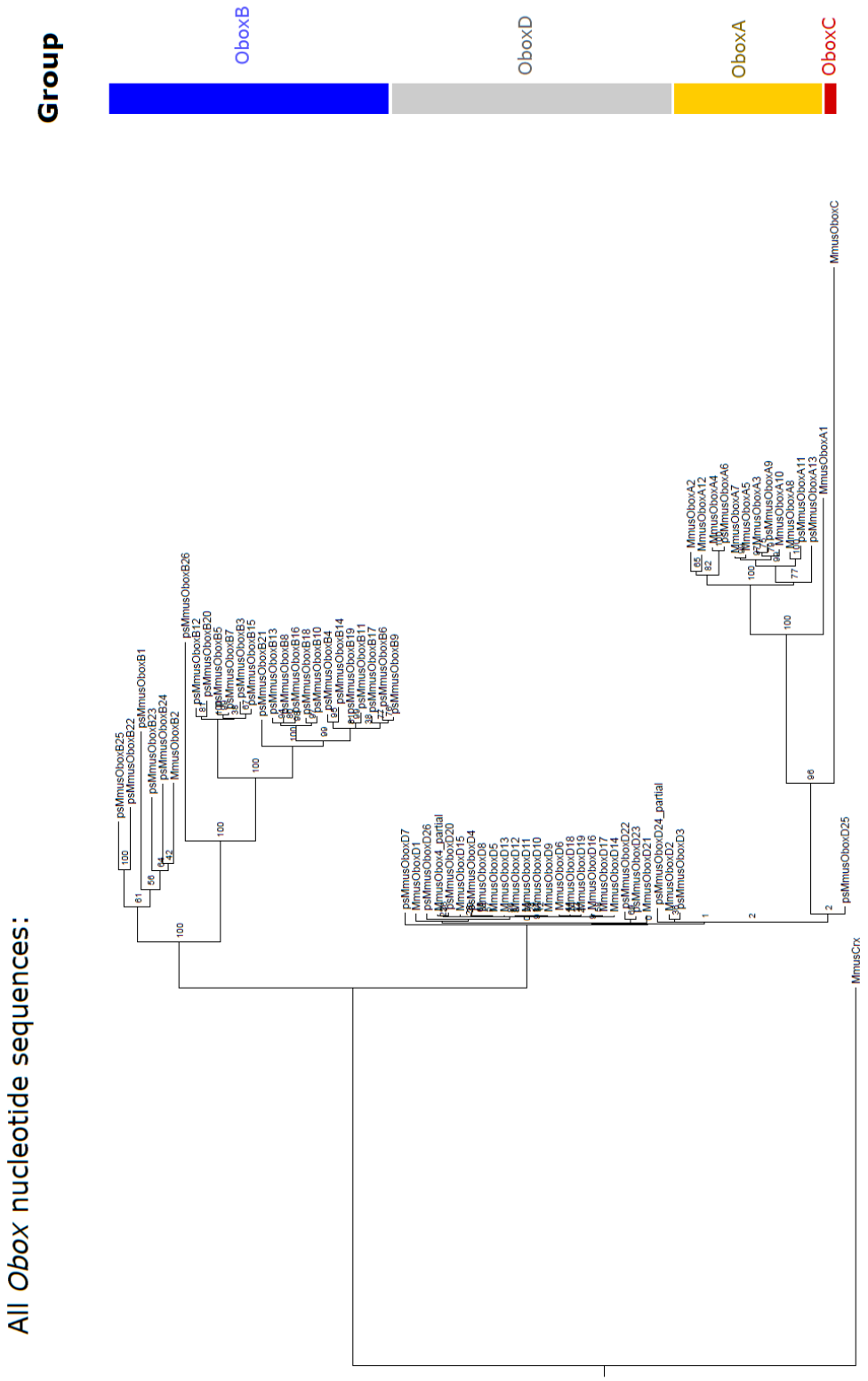
2.9 Gene expression (FPKM) across mouse embryonic stages. Additional

2.10 Gene expression (raw reads) from triplicate RNA sequencing of mouse embryonic fibroblasts following ectopic expression of mouse ETCHbox genes. Additional

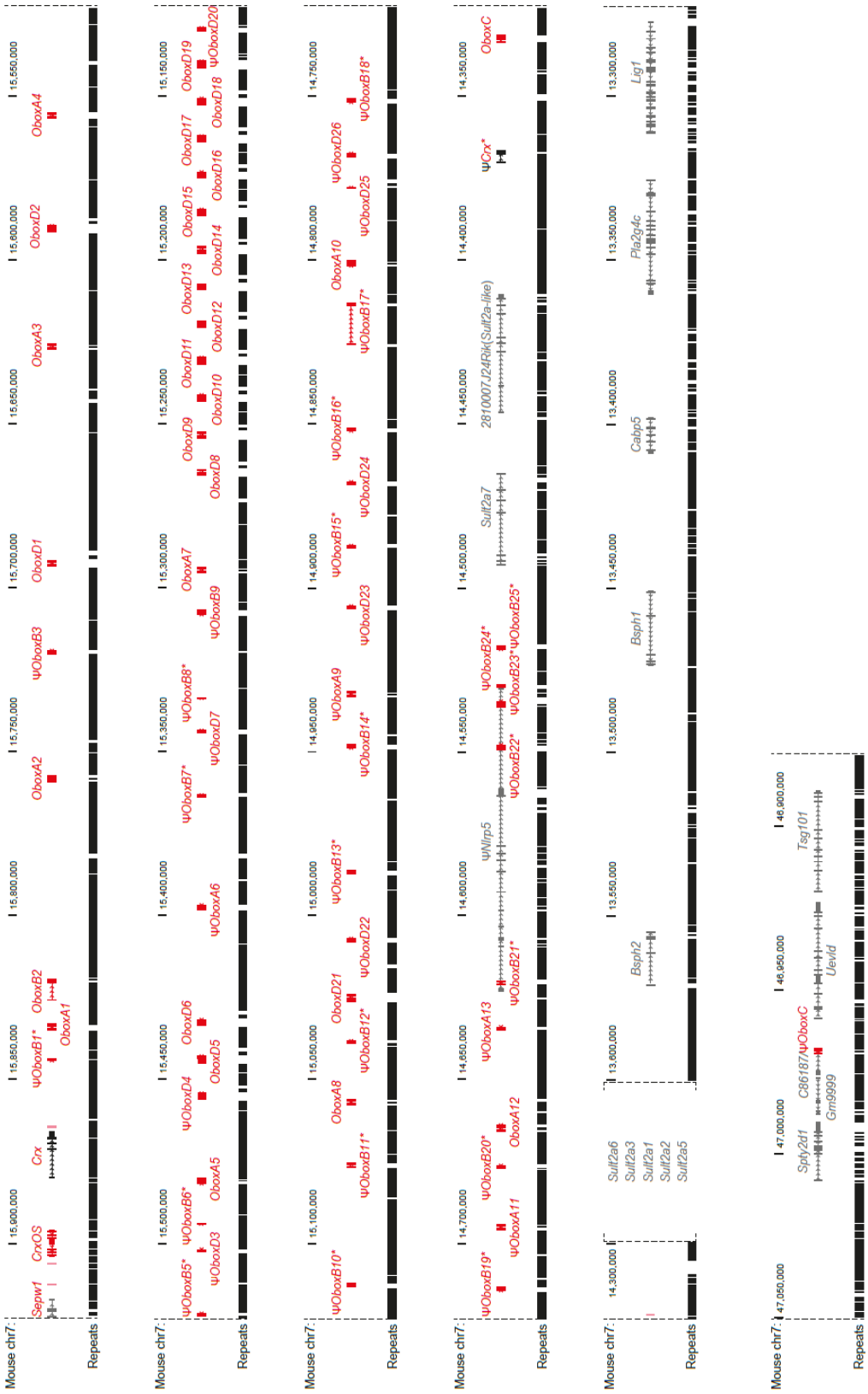
2.11 Average FPKM values for experimental datasets.

2.1

(A) *Obox* phylogeny using nucleotide sequences.



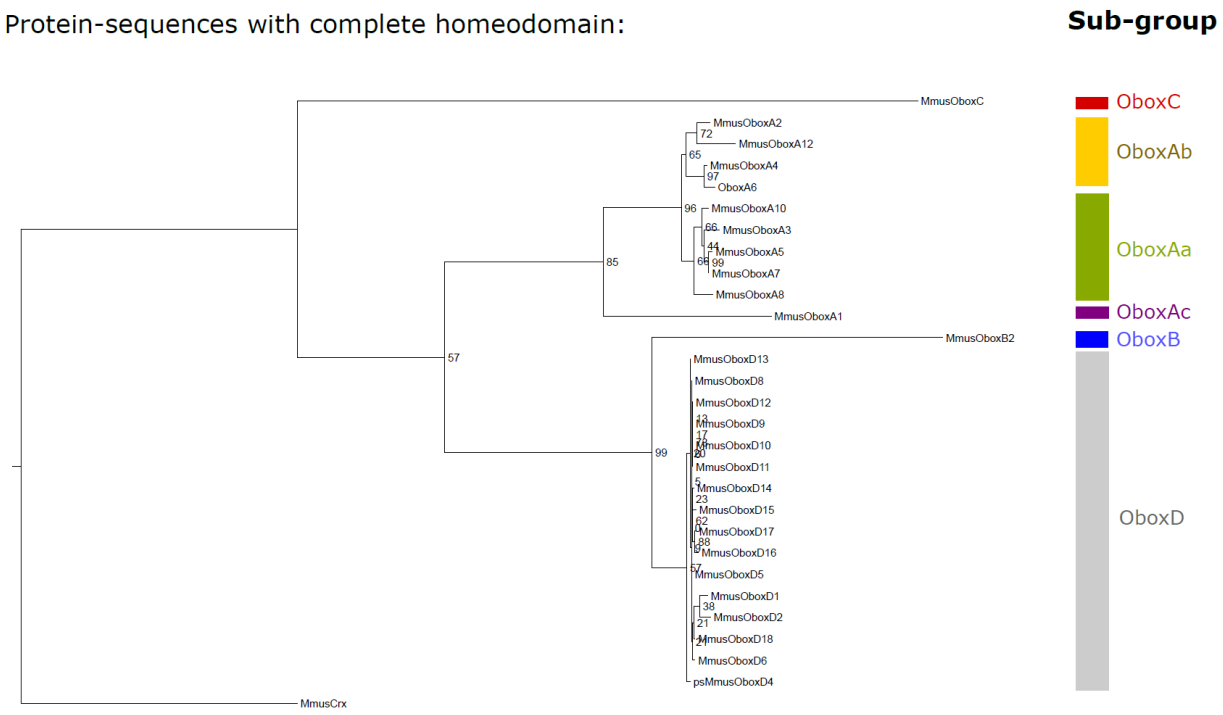
(B) Genomic organisation of *Mus musculus* ETCHbox genes.



(C) Proposed new *Obox* nomenclature. The term 'partial' indicates an Obox sequence lacking a complete homeobox and the term is not part of proposed nomenclature.

Found in Appendix A1.

(D) *Obox* phylogeny using protein sequences that contain a complete homeodomain.



(E) Summary of number of homeoboxes in *Tprx1* syntenic region in different species. * In rat, the two homeobox sequences reflect an independent regional duplication.

Species	Genome assembly	Number of homeodomains in *Tprx1* region
Algerian mouse	GCA_001624865.1	2
Naked mole rat	HetGla_female_1.0	1
Guinea pig	Broad/cavPor3	1
Cow	UMD 3.1.1	1
Rat	RGSC Rnor_6.0	2*

(F) Alignment of *M. musculus* and *M. spretus Tprx1/Crxos* homeodomains.

```
CLUSTAL O(1.2.4) multiple sequence alignment


CrxOS2spretus       MEEIDKMIIQMRLKDSKTVLISKTELTDLQFQKLRKHFETDRYPNEETLQAFAEELKLQK
CrxOS2musculus      MEEIDGMIVQMRLKDSKTVLISKTELTDEQFQKLRKHFETDRCPNEETLQAFAEELKLRK
CrxOS1spretus       MEAS------PRSLTSYTLGPLDQKFSWEQLWELEAYFKMEPYPDLQARKIMATRLKLKE
CrxOS1musculus      MEAS------PRSLTSCTLGPLDQKFSWEQLSELEAYFKVEPYPDLQDRKIMATRLKLKE
                    **         *   * *:   . :::  *: :*. :*: :  *: :  : :* .***::

CrxOS2spretus       AVIRPWFITQRHRMIGYRRLFMRYY-------------------RDWKTSREYSTTRRFD
CrxOS2musculus      DLIRSWFITQRHRMRGYRRLFMRYY-------------------RDWKTSREYSTTRSFD
CrxOS1spretus       EQVEAWFIQRSLEEEM-RPPLARLQQSARDDTPS--------------------------
CrxOS1musculus      EQVEAWFIQRSLEEEM-RPPLARLQQSALDGTSSPSHKALCCRPPSWKYRL-----IPIN
                      :. *** :  .    *  : *

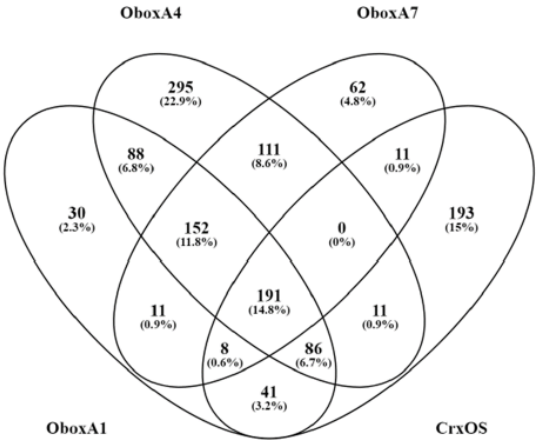
CrxOS2spretus       RQKNSKECSQNDPGLPEALEALKRLKLSSGY-SRDGMSQDF
CrxOS2musculus      RQKNSKECSQNDPGLPEALEALKRLKLSSGYQSRDGMSQDF
CrxOS1spretus       -----------------------------------------
CrxOS1musculus      PPESSTSCKHSC-----------------------------
```

(G) Normalised expression graphs of each temporal profile generated by Mfuzz.

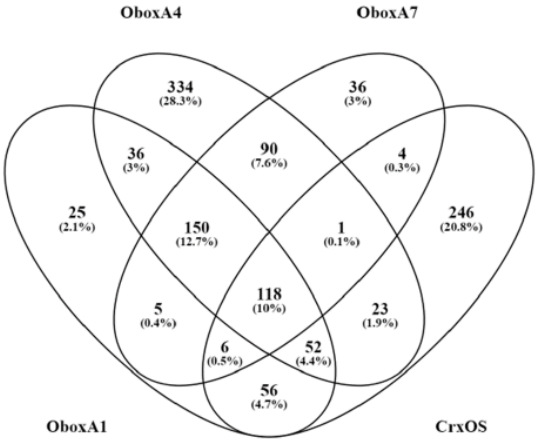
Available online at https://evodevojournal.biomedcentral.com/articles/10.1186/s13227-018-0091-4

(H) Venn diagrams shown numbers of genes affected significantly by each treatment, and overlaps between experiments.

Up-regulated following ectopic expression



Down-regulated following ectopic expression



(I) Enriched GO terms for genes affected by *Obox* ectopic expression.

Available online at https://evodevojournal.biomedcentral.com/articles/10.1186/s13227-018-0091-4

(J) Genes most strongly down-regulated following ectopic expression.

Top 10 genes by fold change down-regulated following over-expression

	OboxA1		*OboxA4*		*OboxA7*		*CrxOS*	
	Gene	Log_2 Fold Change	Gene	Log_2 Fold Change	Gene	Log_2 Fold Change	Gene	Log_2 Fold Change
1	*Blnk*	-1.28	*Ptpn22*	-2.36	*Mmp13*	-1.94	*Dynap*	-1.54
2	*Igsf11*	-1.18	*Mmp13*	-2.14	*Ptpn22*	-1.85	*Ahrr*	-1.43
3	*Hs3st1*	-1.14	*Itga2*	-1.87	*Mmp3*	-1.57	*Krt19*	-1.26
4	*Gprc5a*	-1.13	*Dynap*	-1.64	*Hs3st1*	-1.24	*Itga3*	-1.22
5	*Clic5*	-1.12	*Fam110c*	-1.63	*Clic5*	-1.2	*Ccbe*	-1.1
6	*Itga6*	-1.12	*Areg*	-1.62	*Nptx1*	-1.17	*Aqp5*	-1.08
7	*Ngef*	-1.08	*Mmp3*	-1.53	*Fam110c*	-1.08	*Rasgrp3*	-1.02
8	*Esm1*	-1.08	*Clic5*	-1.5	*Lgr6*	-1.07	*Coch*	-1.01
9	*Sox2*	-0.99	*Serpinb2*	-1.49	*Slc14a1*	-1.06	*Nqo1*	-0.98
10	*Hmga1-rs1*	-0.99	*Upp1*	-1.42	*Areg*	-1.04	*Ngef*	-0.97

(K) Effect of ectopic ETCHbox expression on selected genes with notable preimplantation or stem cell functions.

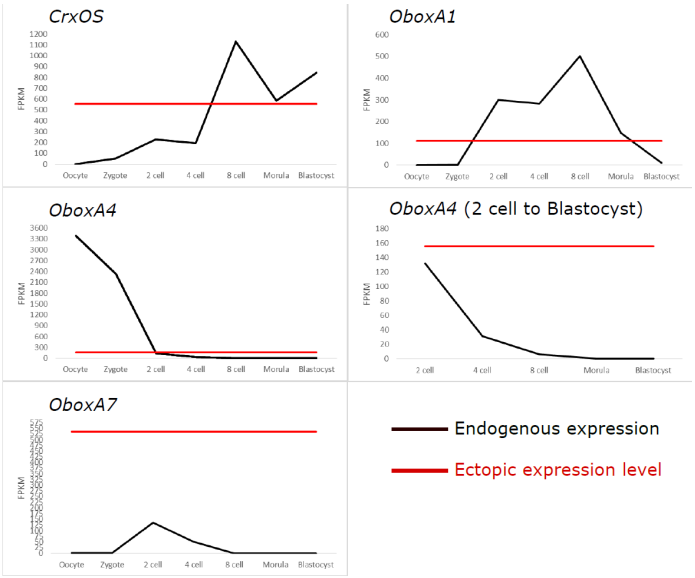
OboxA1		
Gene	**Up/Down regulated**	**Corrected P-value**
Nanog	NA	NA
Oct4/Pou5F1	NA	NA
Sox2	Down	7.23E-10
Klf4	NA	NA
Myc	NA	NA
Tead4	Down	2.64E-05
Cdx2	NA	NA
Gata3	Down	2.51E-08

OboxA4		
Gene	**Up/Down regulated**	**Corrected P-value**
Nanog	NA	NA
Oct4/Pou5F1	NA	NA
Sox2	Down	4.22E-13
Klf4	Down	9.47E-07
Myc	Down	1.83E-31
Tead4	Down	5.07E-06
Cdx2	NA	NA
Gata3	Down	1.31E-07

OboxA7		
Gene	**Up/Down regulated**	**Corrected P-value**
Nanog	NA	NA
Oct4/Pou5F1	NA	NA
Sox2	Down	2.92E-11
Klf4	NA	NA
Myc	Down	0.000000212
Tead4	Down	0.00056
Cdx2	NA	NA
Gata3	Down	0.0000216

CrxOS		
Gene	**Up/Down regulated**	**Corrected P-value**
Nanog	NA	NA
Oct4/Pou5F1	NA	NA
Sox2	Down	0.03
Klf4	NA	NA
Myc	NA	NA
Tead4	Down	0.006
Cdx2	NA	NA
Gata3	Down	0.006041

(L) In vivo expression levels of ETCHbox genes at mouse preimplantation stages compared to ectopic expression levels of transfected genes.



Supplementary files 2.2- 2.11

Available online at https://evodevojournal.biomedcentral.com/articles/10.1186/s13227-018-0091-4

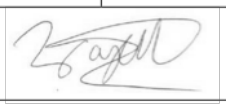
Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Mouse *Obox* and *Crxos* modulate preimplantation transcriptional profiles revealing similarity between paralogous mouse and human homeobox genes
Publication Status	Published
Publication Details	Royall AH, Maeso I, Dunwell TL, Holland PWH, (2018), Mouse *Obox* and *Crxos* modulate preimplantation transcriptional profiles revealing similarity between paralogous mouse and human homeobox genes, EvoDevo 9, DOI: 10.1186/s13227-018-0091-4

Student Confirmation

Student Name:	Amy H Royall		
Contribution to the Paper	Statement of contribution within the paper		
Signature		Date	08/09/2019

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Peter Holland, Linacre Professor of Zoology		
Supervisor comments Amy Royall was primary contributor to this paper. Exact details of contributions detailed in the paper		
Signature 	Date	12/9/19

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 3: Of Eyes and Embryos; Subfunctionalisation of the *CRX* Homeobox Gene in Mammalian Evolution.

AH Royall, S Frankenberg, AJ Pask, PWH Holland (2019).

Introduction to manuscript:

The vast majority of this chapter (excluding RNA sequencing data in sections 2.4, 3.2, and figure 3.3) has been published as a first author entirely Open Access research article in *Proceedings of the Royal Society B* (Appendix A3; doi.org/10.1098/rspb.2019.0830) with minor formatting changes prior to inclusion in this thesis (under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/40). The authors on this manuscript in order are Amy H Royall, Stephen Frankenberg, Andrew J Pask, Peter WH Holland (corresponding author).

The variability in gene copy number and sequence of the ETCHbox genes sparks questions relating to how such variability can evolve alongside function. We know that the ETCHbox genes' closest relative is the highly conserved *CRX* gene, which in eutherian mammals is very specific to the developing eye and adult retina, whereas ETCHbox themselves are not conserved and exhibit completely different characteristics. How can such a constrained gene, conserved across vertebrates and with highly specific expression patterns, have duplicates with such variability, as well as specificity for an entirely different domain of expression? In this manuscript I explore these evolutionary questions using comparative analyses and transcriptomics to analyse *CRX* evolution in eutherian mammals and their closest outgroup, metatherians. My and my fellow authors' findings in this chapter culminate in a hypothesised mechanism for ETCHbox evolution from *CRX* based on an adjusted subfunctionalisation model we name DDC+.

Authors' contributions

AHR and PWHH conceived and designed the study. AHR performed sequence analyses, molecular and cell biology experiments, transcriptomics and statistical analyses of data. SF and AJP carried out dunnart embryo sequencing and analysis. AHR and PWHH wrote the manuscript and all authors contributed to and approved the final manuscript.

Data

The RNA-sequencing data following ectopic expression in Mouse Embryonic Fibroblasts can be accessed under GSE129456 through the NCBI Gene Expression Omnibus (GEO). RNA sequencing of adult dunnart tissues are unpublished but can be accessed on the NCBI Sequence Read Archive (Project ID:PRJNA******, private until article publication, contact AHR or PWHH for access to raw read files).

Acknowledgements

We thank Rodrigo Pracana, Yichen Dai, Aris Katzourakis, Sounak Sahu, Aziz Aboobaker, Anish Dattani, Sonia Trigueros, Tim Davies and Paul Fairchild for discussions and advice.

Funding

We acknowledge BBSRC funding to AHR through the Oxford Interdisciplinary Bioscience Doctoral Training Partnership (BB/M011224/1) and Discovery Project funding (DP160103683) from the Australian Research Council to AJP and SF.

Abstract

ETCHbox genes are fast-evolving homeobox genes present only in eutherian (placental) mammals which originated by duplication and divergence from a conserved homeobox gene, *Cone-rod homeobox* (*CRX*). While expression and function of *CRX* is restricted to the retina in eutherian mammals, ETCHbox gene expression is specific to pre-implantation embryos. This dramatic difference could reflect the acquisition of new functions by duplicated genes or subfunctionalisation of pleiotropic roles between *CRX* and ETCHbox genes. To resolve between these hypotheses, we compared expression, sequence and inferred function between *CRX* of metatherian (marsupial) mammals and ETCHbox genes of eutherians. We find the metatherian *CRX* homeobox gene is expressed in early embryos and in eyes, unlike eutherian *CRX*, and distinct amino acid substitutions were fixed in the metatherian and eutherian evolutionary lineages consistent with altered transcription factor specificity. We find that metatherian *CRX* is capable of regulating embryonically-expressed genes in cultured cells in a comparable way to eutherian ETCHbox. The data are consistent with *CRX* having a dual role in eyes and embryos of metatherians, providing an early embryonic function comparable to that of eutherian ETCHbox genes; we propose that subfunctionalisation of pleiotropic functions occurred after gene duplication along the placental lineage, followed by functional elaboration.

1.Introduction

Gene duplication is postulated to facilitate evolutionary change and innovation through providing new genetic material which can gain or lose functions over evolutionary time. Although most homeobox gene families are highly conserved across Metazoa, there are examples of homeobox genes that have undergone duplication followed by extreme sequence divergence in some lineages.

The Otx homeobox gene family provides a striking example. Following characterisation of the *Drosophila orthodentical* (*otd*) gene, two homologues were rapidly identified in mammals: *OTX1* and *OTX2* (Simeone *et al.*, 1992). Subsequently, a third gene family member was identified, named *OTX5* or *CRX* (Vignali *et al.*, 2000; Plouhinec *et al.*, 2003). These three Otx family genes are descendents of the whole genome duplications that occurred in vertebrate ancestry; they are found across jawed vertebrates and share high sequence similarity. More recently it was shown that a series of highly divergent homeobox genes, specific to eutherian mammals, originated later in evolution by tandem duplication from the *CRX* gene followed by extreme sequence divergence away from a canonical Otx-type sequence. Their pattern of sequence divergence has been described as asymmetric, meaning that the 'parental' gene (*CRX*) has undergone far less sequence change than the 'daughter' genes (Germot *et al.*, 2001; Plouhinec *et al.*, 2003; Maeso *et al.*, 2016; Holland *et al.*, 2017). The descendent genes are so distinct in sequence from Otx family genes that they were given their own gene family names: *Tprx*, *Dprx*, *Leutx* and *Argfx* (Booth and Holland, 2004) and collectively ETCHbox (Maeso *et al.*, 2016). This innovation and sequence change is of note as lineage-specific genes could have important roles in evolution(Villanueva-Cañas *et al.*, 2017). In eutherian mammals, the 'parental'

CRX gene locus is flanked by *Tprx1* (called *Crxos* in mouse) and *Tprx2* (Obox genes in mouse); the other duplicates are more distantly located or on another chromosome (Maeso *et al.*, 2016; Royall *et al.*, 2018). There is also variation between eutherian species as to which ETCHbox are retained or lost, and which have been duplicated further [5].

Divergent sequence evolution is mirrored by distinct expression patterns and probable functions. The ETCHbox genes of human and cow are expressed predominantly in early development, in a sharp pulse around the 8-cell to 16-cell stages (Töhönen *et al.*, 2015; Madissoon *et al.*, 2016; Maeso *et al.*, 2016). Mouse ETCHbox genes are more variable in expression, but are also specific to the very early embryo (Rajkovic *et al.*, 2002; Saito *et al.*, 2011; Royall *et al.*, 2018). Ectopic expression in cell culture, followed by RNA sequencing (RNA-seq) has uncovered putative roles in controlling a suite of downstream genes, including genes peaking in expression before blastocsyst and implicated in cell fate and foetal-maternal interactions (Maeso *et al.*, 2016; Royall *et al.*, 2018).

In contrast, the *CRX* gene encodes a transcription factor implicated in specification of photoreceptor fate in the developing eye (Plouhinec *et al.*, 2003; Goldshmit *et al.*, 2010; Andzelm *et al.*, 2015). There is also some evidence for trace expression of *CRX* in the early embryo in human and mouse, but the levels are too low to be identified by whole embryo RNA-sequencing (Blakeley *et al.*, 2015; Maeso *et al.*, 2016; Royall *et al.*, 2018). Mouse models with perturbed *CRX* function or *CRX* gene deletion are characterised by near or complete absence of vision from birth, and mutations in humans can cause cone-rod dystrophy (Freund *et al.*, 1997; Goldshmit *et al.*, 2010; Wang *et al.*, 2018). Expression in the eye has also been

reported for *Crx*/*Otx5* in frog, chicken, zebrafish (two genes), and dogfish (Vignali *et al.*, 2000; Viczian, 2003; Shen and Raymond, 2004; Plouhinec *et al.*, 2005). However, in some non-mammalian vertebrates, *CRX* expression has also been found in embryonic tissues. For example, in the frog *Xenopus laevis, crx (Xotx5)* RNA is detected in the Spemann organiser of the gastrula and in anterior neuroectoderm (Kuroda *et al.*, 2000; Vignali *et al.*, 2000); in dogfish *Scyliorhinus canicula*, *Crx (Otx5)* expression can be detected in the embryonic brain by *in situ* hybridisation in addition to the developing eye (Sauka-Spengler *et al.*, 2001; Plouhinec *et al.*, 2005).

It is unclear whether the vertebrate *CRX* gene had ancestral roles in the early embryo, which have been secondarily lost in the eutherian mammal lineage (Plouhinec *et al.*, 2003), or alternatively whether some sites of embryonic expression were independently gained in *Xenopus* and dogfish. This question is closely tied to understanding the evolution of ETCHbox genes, because these embryo-specific genes were derived from *CRX* in eutherian mammal evolution. If *CRX* was pleiotropic with roles in embryos and eyes prior to the emergence of eutherian mammals, the distinct functions in eutherians could reflect subfunctionalisation after gene duplication, followed by gradual and progressive optimisation of multiple ETCHbox roles to subtly different embryonic roles (Royall *et al.*, 2018). In contrast, if *CRX* was ancestrally eye-specific, ETCHbox genes and their functions would be very radical innovations acquired by eutherian mammals only, thus marking a clear distinction to other vertebrates.It is not clear which scenario is more likely, since eutherian mammal embryogenesis has a mixture of plesiomorphic and derived features. The eutherian mammal lineage encompasses the emergence of the extant Placentalia, characterised by extended embryogenesis and development of a highly invasive placenta that contributes to embryo nutrient

exchange throughout embryogenesis. A comparison to non-eutherian mammals is key to resolving these questions. Marsupials (metatherians) are the immediate outgroup to eutherian mammals, and their embryos do establish maternal interactions albeit with a less invasive placenta than in eutherian mammals (Renfree, 2010). Furthermore, metatherians do not possess ETCHbox genes, as these arose from *CRX* specifically on the eutherian stem lineage.

In this study we ask whether a metatherian *CRX* gene is eye-specific or whether it also has expression and possible function in the early embryo, comparable to ETCHbox genes. Using sequence comparisons, we first show that specific amino acid changes occurred in the *CRX* proteins of metatherians and eutherians, compatible with alterations to transcription factor function during mammalian evolution. We examine expression of the *CRX* gene in metatherian development, detecting expression in early embryos and in eye. To enable comparisons to previously characterised functions of eutherian ETCHbox genes, we expressed metatherian and eutherian *CRX* genes ectopically in cell culture, and examined transcriptomic responses using RNA-seq and qPCR. These experiments uncovered similarities in activity between metatherian *CRX* and eutherian ETCHbox genes, consistent with subfunctionalisation and progressive specialisation after gene duplication in eutherian mammal evolution.

2. Materials and Methods

2.1 Protein sequence analysis

Deduced protein sequences were gathered from NCBI and aligned using MAFFT to identify amino acid substitutions in the homeodomain. The inferred vertebrate ancestral sequence was estimated using FastML (Ashkenazy *et al.*, 2012, supplementary material 3. 1A). We tested for positive selection using CodeML implemented in PAML (Yang, 2007), using a branch-model to estimate the dN/dS ratio by assigning two independent ratios, specifying the branch leading to eutherian mammals (model = 2, NSsites = 0).

2.2 Gene expression analysis

To examine mouse gene expression, publicly available RNA-seq datasets (supplementary material 3. 2A) from across preimplantation development were analysed using kallisto v0.42.4 using GRCm38 coding sequences and default settings to obtain number of transcripts-per-million mapped reads (TPM)(Bray *et al.*, 2016). Coding sequences for *Oboxa7*, *Oboxb2*, and *Oboxd2* were manually added to the GRCm38 coding sequence dataset as they are currently unannotated. Other mouse transcripts examined were *Crx* (ENSMUST00000044434.12), *Crxos* (ENSMUST00000171280.2), *Oboxa1* (ENSMUST00000108513.4), *Oboxa4* (ENSMUST00000067288.14), *Otx1* (ENSMUST00000006071.13), and *Otx2* (ENSMUST00000226501.1).

2.3 *Sminthopsis crassicaudata* conceptus expression

Dunnart single-cell RNA-seq data were derived from a larger study to be published in a separate manuscript. Dissociated single cells from a range of conceptus stages were processed using Clontech SMART-Seq v4 3' DE kit, with sample volumes scaled down five-fold. Libraries were prepared using the Illumina Nextera XT DNA Sample Preparation Kit and sequenced at 1 million reads per cell (12 cells per multiplexed library) on an Illumina Nextseq 500 platform. Using the Galaxy web-based analysis platform (https://usegalaxy.org/), de-multiplexed reads were mapped to dunnart genomic scaffolds (unpublished) using Bowtie for Illumina (default parameters) and assembled into transcripts using Stringtie and Stringtie Merge. Stringtie was then reapplied to the mapped reads using the Stringtie Merge output as a transcripts reference file to produce the TPM (transcripts per million) dataset reported here. *CRX* was identified among the Stringtie Merge assembled transcripts using BLASTn.

2.4 Gene expression in adult *Sminthopsis crassicaudata* tissues.

Fat-tailed dunnart tissue collections were approved by the University of Melbourne Animal Experimentation Ethics Committees and were in accordance with the National Health and Medical Research Council of Australia (2004) guidelines. Total RNA was extracted from frozen adult female dunnart tissues using the RNeasy Mini Kit™ (Qiagen), followed by treatment with DNA-*free*™ (Ambion) to eliminate genomic DNA. Oligo(dt)-primed first-strand cDNA was reverse transcribed using Superscript III™ (Invitrogen). PCR was performed using GoTaq Green™ (Promega) in a 20-µL reaction volume for 40 cycles of 95°C for 15 sec, 61°C

for 20 sec, and 72°C for 20 sec, using template equivalent to 50 µg of the original total RNA sample. Dunnart-specific primer sequences were: *CRX*- forward AAGATCAATCTCCCAGAATCCCGAG; *CRX*-reverse AGTCAGTGCATAGGAAGAGGAGG; *TBP*- forward TCCCCAATGACCCCTATGACTC; *TBP*-reverse TTTCTGGCTGCTAACCTGGAC. Products were analysed by electrophoresis. RNA harvested for sequencing by SF and sequenced using BGISeq500 platform with 75bp-paired end sequencing yielding between 57 and 75 million reads per sample. In the absence of an accessible genome, the raw sequences were assembled into individual *de novo* transcriptomes using Trinity (Grabherr *et al.*, 2011) and TBLASTN used to identify *CRX* transcripts within the assemblies (Data available on SRA following article publication, contact AHR or PWHH for access).

2.5 Ectopic expression and transcriptomic responses

Plasmids containing V5-tagged human *CRX,* V5-tagged opossum *CRX* (RefSeq: XM_001372308.2, Genscript CloneID: Omc23124) or no homeobox gene, along with a puromycin resistance gene (Oxford Genetics; OG3422), were electroporated using a NEPA21 system (NEPAGENE) into Mouse Embryonic Fibroblasts (MEFs, ScienCell: M7540-57) and allowed to recover for 48 hours in DMEM (Gibco: 11965092; 1% PenStrep (Gibco: 15140122), 10% FBS). Cells for immunocytochemistry were fixed in 4% formaldehyde in PBS for twenty minutes before permeabilisation using 0.1% Triton X in 1x PBS, 0.1% Tween-20. Antibodies and visualisation are described previously (Royall *et al.*, 2018).

Cells for RNA harvesting had media replaced for DMEM containing puromycin (0.5 µg/mL). Puromycin was kept in the media for 5 days with a media

change on day 3. Puromycin was washed off using multiple short (5 min) incubations in DMEM and cultured overnight. RNA was harvested in triplicate per sample (control, metatherian *CRX*, eutherian *CRX*) using RNeasy Micro Plus kit (Qiagen) following manufacturer's instructions. RNA was prepared for sequencing using the Illumina TruSeq library preparation kit and size of the fragments checked (Agilent D1000 ScreenTape). RNA was sequenced using the Illumina NextSeq500 platform using 2x75bp paired end reads giving between 68 and 121 million reads per sample (Data available under GSE129456). Transcriptome reads were aligned used STAR aligner v2.4.0 (genome assembly: mm10 GRCm38; (Dobin *et al.*, 2013)) and Cufflinks v2.2.1 software used to deduce FPKM values; FeatureCounts v1.4.6 produced raw read counts (Royall *et al.*, 2018). Clustering using Mfuzz software (Kumar and Futschik, 2007) and gene list overlap analysis was performed as described in Royall *et al.* (2018). Quantitative PCR (qPCR) experiments used New England BioScience Luna Universal One-Step qPCR kit according to manufacturer's instructions and Prime Pro 48 Real-Time qPCR system (Techne). qPCR analysis was performed relative to the *Gapdh* housekeeping gene and relative expression was calculated by comparison to control MEFs transfected with the 'no homeobox gene' plasmid. Mouse-specific qPCR primers are: *Gapdh*-forward CGGGTTCCTATAAATACGGACTG, *Gapdh*-reverse AATACGGCCAAATCCGTTCA, *Hbegf*-forward TCCCTCTTGCAAATGCCTC, *Hbegf*-reverse GGACGACAGTACTACAGCCA, *Ptgs2*-forward TGGGTGTGAAGGGAAATAAGG, *Ptgs2*-reverse GAAGTGCTGGGCAAAGAATG, *Smarca1*-forward TATGCCCTTGAAAGCAGACC, *Smarca1*-reverse TGTTGGAGACTTCTGTGCTG.

To enable comparisons between transcriptomic responses following *CRX* and ETCHbox ectopic expression, differential expression analysis was carried out

on FPKM counts using R package DESeq2 (Love, Huber and Anders, 2014). The resulting list of differentially expressed genes was compared with those differentially expressed following ETCHbox ectopic expression (previously obtained using the same methodology as in Royall *et al.*, 2018) and overlapping responses calculated as a percentage of total differentially regulated genes in this study. The extent of overlap was assessed using Fisher's Exact Test. The GO term enrichment analysis used functional annotation chart from DAVID Bioinformatics Resources v6.8 (Huang, Brad T Sherman and Lempicki, 2009; Huang, Brad T. Sherman and Lempicki, 2009).

3. Results

3.1 Eutherian-specific substitutions in CRX homeodomain

We hypothesized that if a pleiotropic CRX protein underwent subfunctionalisation during the evolution of eutherian mammals, with some functions partitioned to ETCHbox genes, this may have resulted in novel amino acid substitutions associated with CRX specialisation in the eutherian lineage. We searched for CRX sequences across vertebrates and confirmed the identity of opossum *CRX* using synteny to mouse and human loci (figure 3.1). As *CRX* encodes a transcription factor, we focussed on the homeodomain sequence responsible for sequence-specific DNA binding and target gene specificity. Comparison of the CRX homeodomain sequence from a range of vertebrates revealed high conservation, but with substitutions in some lineages (figure 3.1A).

We find substitutions at eight amino acid sites in the CRX homeodomain of eutherian mammals, all completely conserved between eutherians examined (figure 3.1A). By contrast, metatherian CRX has undergone substitutions at three sites (two conserved); other lineages have undergone limited change. A phylogenetic tree of CRX protein sequences clearly separates metatherian and eutherian *CRX* genes (supplementary material 3.2B). We tested if the eutherian-specific substitutions were driven by positive selection, as indicated by an elevated dN/dS (ω) ratio, but found the long divergence times were associated with saturation of dS and gave unreliable ω estimation. Nonetheless, the amino acid changes in eutherian mammals are consistent with an altered function compared to other vertebrates, compatible with the subfunctionalisation hypothesis. We find no evidence for reciprocal compensatory changes in ETCHbox homeodomains (supplementary material 3. 2C).

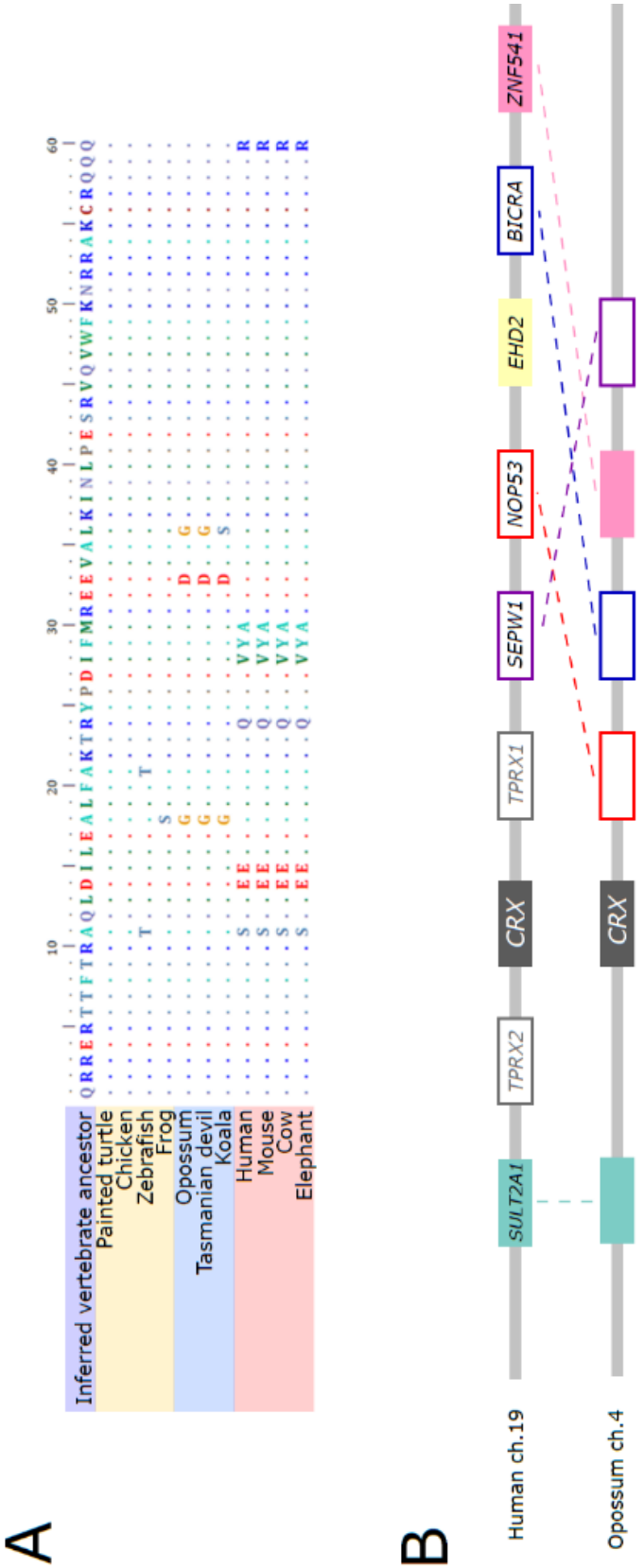


Figure 3.1 Organisational and amino acid changes in mammalian CRX. (A) Eight amino acid changes are present in all eutherian mammal CRX homeodomains and absent in non-eutherian vertebrates. Two conserved amino acid changes are metatherian-specific and not found outside of metatherians. (B) Synteny of opossum *CRX* relative to orthologous human genomic region confirms identity of opossum *CRX* locus. Assembly of dunnart genome is not complete enough to describe the syntenic region.

3.2 *CRX* expression in metatherian embryos and eyes

If subfunctionalisation occurred after gene duplication, with retinal functions retained by eutherian *CRX* and embryonic roles retained by diverging ETCHbox genes, then the ancestral mammalian *CRX* gene would have had expression in eyes and in early embryos. If neofunctionalisation occurred, with ETCHbox genes acquiring novel roles, the ancestral gene would be eye specific. To resolve between these alternatives we examined *CRX* gene expression in a metatherian.

We analysed single cell RNA sequencing data from early developmental stages of the fat-tailed dunnart (*Sminthopsis crassicaudata*), a marsupial in the family Dasyuridae. We found expression of *CRX* in the early embryo (figure 3.2), with mean transcripts per million (TPM) of 79.6 at 16-cell, 14.1 at 32-cell, and 40.8 at 64-cell (supplementary material 3. 1B). In contrast, comparable analysis of mouse and human whole embryo RNA-seq data revealed no expression in the blastocyst or earlier stages, when ETCHbox genes are specifically expressed (Maeso *et al.*, 2016) (Figure 3.2). Using RT-PCR we confirmed the expected, widely conserved, expression of *CRX* in the adult metatherian eye (supplementary material 3. 2D). RNA-sequencing of adult dunnart tissues confirmed specific expression of *CRX* in the adult eye, with an absence of expression in the tesis, ovary, and fetal membranes (figure 3.3). A transcript is identified in the mixed uterus and oviduct sample which could be explained by preganancy of the sampled animal (figure 3.3). This also allowed us to predict the dunnart homeodomain sequence (figure 3.3). These data imply that the ancestral mammalian *CRX* gene was expressed in embryos and eyes specifically, and these properties were separated after gene duplication.

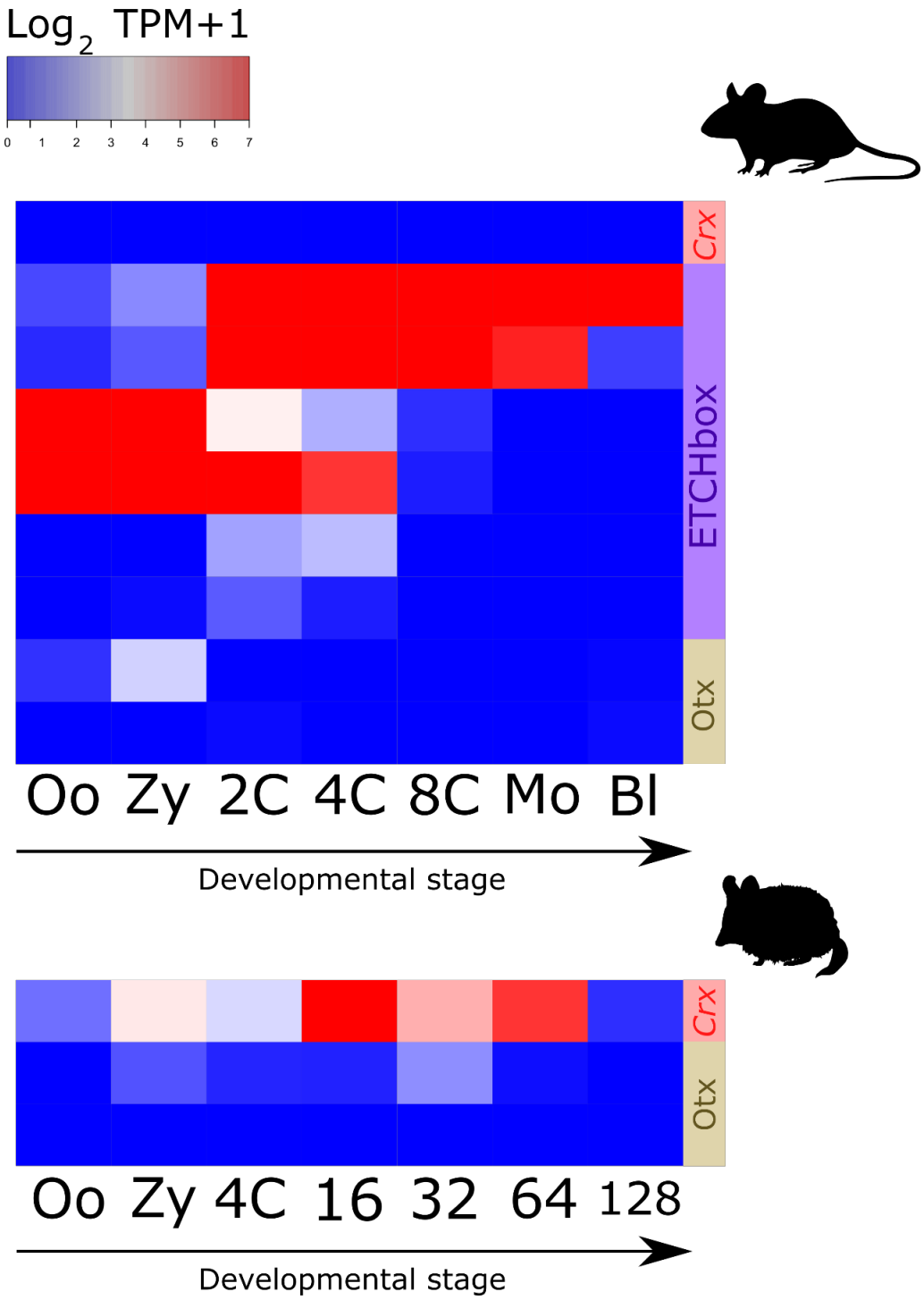


Figure 3.2 Comparison of eutherian (mouse, top) and metatherian (fat-tailed dunnart, bottom) *CRX* gene expression. Mouse ETCHbox genes, but not *Crx,* are expressed in early embryos. Dunnart *CRX* is expressed in early embryos. Expression levels represented as Log2 TPM+1 (TPM in supplementary material 3. 1B) Embryonic stages: Oo = Oocyte, Zy = Zygote, 2C = two cell, 4C = four cell, 8C = eight cell, Mo = morula, Bl = blastocyst, 16 = 16 cell, 32 = 32 cell, 64 = 64 cell, 128 = 128 cell. Embryonic stages not necessarily equivalent between species. ETCHbox genes from top to bottom*: Crxos, Oboxa1, Oboxa4, Oboxa7, Oboxb2, Oboxd2*. Otx genes from top to bottom: *Otx1*, *Otx2*.

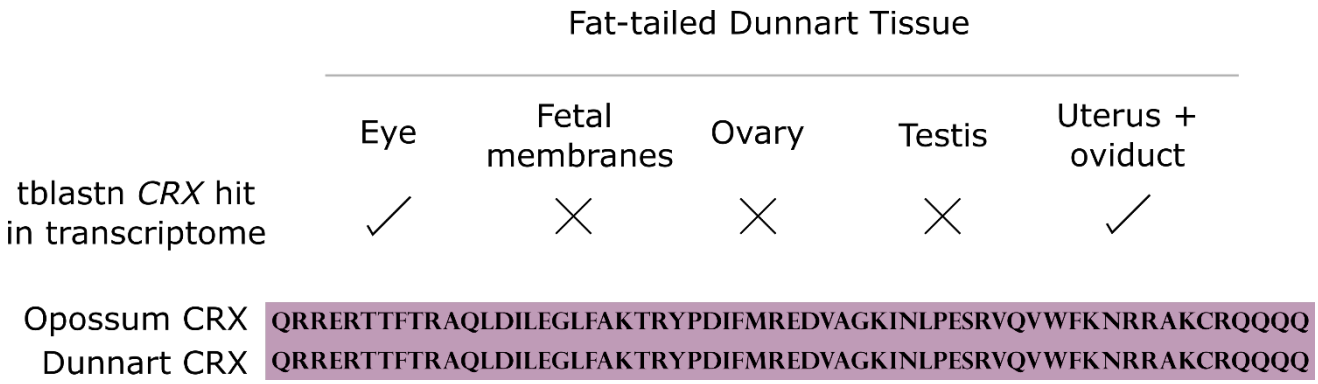


Figure 3.3 Results of *de novo* transcriptome search for *CRX* transcript. TBLASTN using opossum CRX protein sequence against fat-tailed dunnart tissue-specific transcriptomes shows that the conserved *CRX* expression in the eye is persent in the dunnart in addition to the early embryo (figure 3.2). Conversely, dunnart CRX expression is not seen in an array of transcriptomes from other tissues (Fetal membranes, ovary, or testis), however it can be identified in the uterus and oviduct mixed sample which could be explained by preganancy (top). This seach also allowed identification of the dunnart protein sequence, revealing no amino acid changes in the homeodomain when compared to the oppossum sequence (bottom).

3.3 Radical restructuring of transcriptomes by the metatherian CRX protein

We hypothesised that a homeodomain protein with diverse roles (in early embryo and eye) would be capable of regulating more target genes and influencing more regulatory networks than a homeodomain protein of highly specialised function. We expressed ectopically a *CRX* gene from a metatherian (*Monodelphis domestica*) and an eutherian mammal (*Homo sapiens*) in primary mouse embryonic fibroblast cells and used RNA-seq to determine effects on global transcriptome, relative to cells transfected with a control vector. Although this experiment does not mimic closely the in vivo situation, previous studies have found that ectopic expression of related homeobox genes in cultured cells can drive biologically-relevant transcriptomic changes (Maeso *et al.*, 2016; Royall *et al.*, 2018). Immunocytochemistry confirmed that both transfected genes are translated into protein (supplementary material 3. 2E).

Following overexpression, the two *CRX* orthologues had different global effects (figure 3.4). Ectopic expression of metatherian *CRX* resulted in differential expression of 1,496 genes (661 up, 835 down; FPKM>2, P< 0.05; supplementary material 3.1C). Ectopic expression of eutherian *CRX* resulted in differential expression of only 580 genes (118 up, 462 down; supplementary material 3. 1C). There was partial response overlap with 61 common genes up-regulated in the two experimental conditions and 278 down-regulated. The discrepancy in the number of genes differentially regulated is not due to different levels of ectopic gene expression (supplementary material 3. 2F) or variations in methodology, as samples were prepared in pairs and associated in analysis. Gene Ontology (GO) term enrichment analysis showed that genes down-regulated following eutherian CRX

expression are enriched for cell cycle functions (supplementary material 3. 1D), whereas those down-regulated following metatherian CRX expression are enriched for terms relating to extracellular matrix (supplementary material 3. 1E). We conclude that a metatherian CRX homeodomain protein elicits more downstream changes after ectopic expression than a eutherian CRX protein in a mouse fibroblast context.

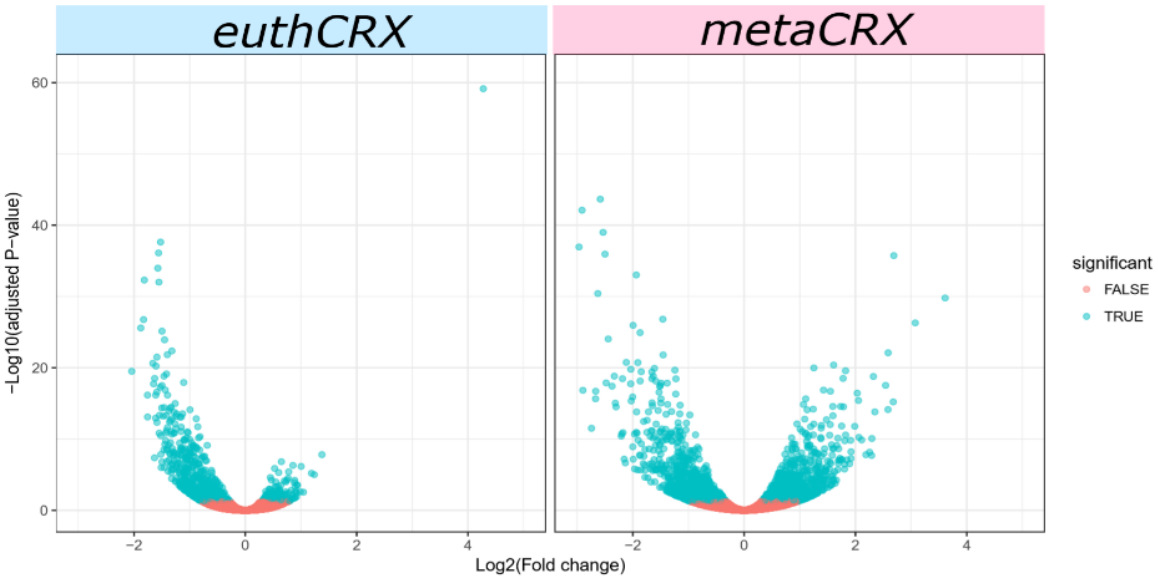
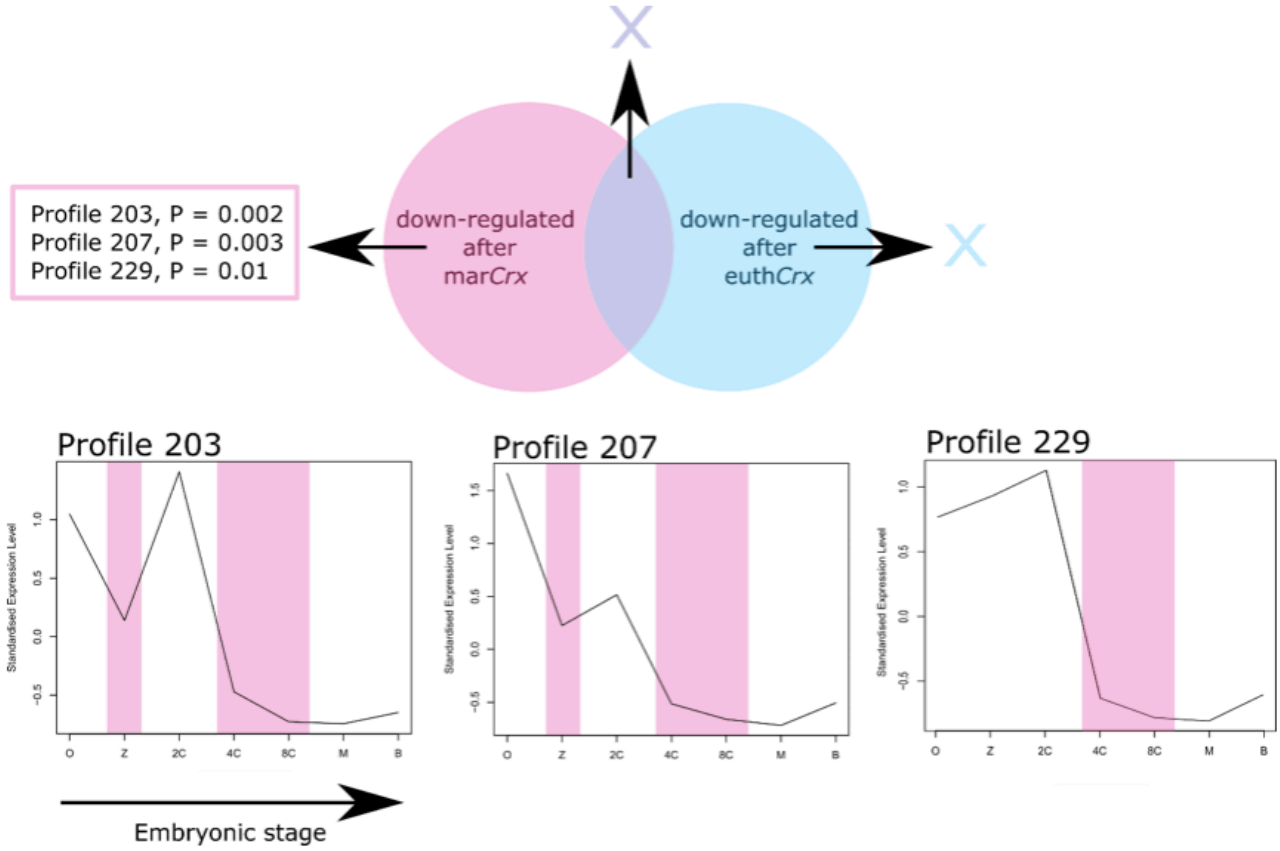


Figure 3.4 Effect on transcriptome of *CRX* ectopic gene expression. Metatherian *CRX* (*metaCRX*; right) induces more changes in gene expression than eutherian *CRX* (*euthCRX*; left) following ectopic expression. Both genes have more down-regulatory effects than up-regulatory effects. Volcano plots of –Log10(adjusted P-value) and Log2 fold-change following differential gene expression analysis.

3.4 Metatherian CRX protein regulates genes expressed in preimplantation development

To assess if transcriptomic changes elicited by CRX protein expression are associated with developmental functions, we first asked how many of the differentially regulated genes are normally expressed in mouse embryos. We compiled a database of genes expressed in preimplantation mouse embryos (Royall *et al.*, 2018) and found the transcriptomic response to metatherian *CRX* includes 732 genes that are dynamically expressed in normal preimplantation mouse development (349 up, 383 down, supplementary material 3.1F). Of the genes affected by ectopic eutherian *CRX*, only 277 are dynamically expressed in mouse preimplantation development (52 up, 225 down, supplementary material 3.1F).

Grouping developmentally-expressed genes into sets based on their temporal expression profiles reduces complexity of analysis and can help distinguish signal from noise (Madissoon *et al.*, 2016; Maeso *et al.*, 2016; Royall *et al.*, 2018). We asked if the differentially expressed genes were enriched for genes that follow particular temporal profiles in development. If profile enrichment is found, this can increase confidence in the biological relevance of the transcriptomic response and provide clues to putative *in vivo* roles (Madissoon *et al.*, 2016; Maeso *et al.*, 2016; Royall *et al.*, 2018). We used a clustering approach to assign all mouse genes to 85 different temporal expression profiles spannning early embryonic development (elctronic supplementary material 2G), and tested if the lists of genes differentially regulated after *CRX* ectopic expression were enriched for particular temporal profiles.



<u>Figure 3.5 Mouse preimplantation clusters significantly regulated by metatherian *CRX* ectopic expression</u>. Coloured strips identify common features of the clusters which we propose are related to the endogenous action of metatherian *CRX*.

We did not find profile enrichment in the lists of genes up-regulated either by metatherian CRX or by eutherian CRX proteins. We also found no profile enrichment in the set of genes down-regulated by eutherian *CRX* ectopic expression, or the common set of genes down-regulated by both eutherian and metatherian CRX proteins (figure 3.5). However, we found significant enrichment for three temporal profiles in the list of genes down-regulated by only metatherian *CRX* ectopic expression (profiles 203, 207 and 229; supplementary material 3. 2G). The temporal gene expression patterns encapsulated by these profiles differ in detail, but all three are composed of mouse genes showing reduction in gene expression level between

the 2-cell and 8-cell stages of embryonic development (figure 3.5). We conclude that the metatherian CRX protein, but not the eutherian CRX protein, is capable of down-regulating genes with common expression patterns in the early mouse embryo and is more likely to have *in vivo* functions at this period of development. Specifically, metatherian CRX can regulate a suite of genes that show declining expression in early mouse development. We propose that this activity reflects an *in vivo* role of the metatherian *CRX* gene: down-regulating expression of specific genes before blastocyst formation.

3.5 Common downstream effects of metatherian *CRX* and eutherian Obox genes

If the embryonic roles of eutherian ETCHbox genes arose initially by subfunctionalisation, a prediction is that metatherian *CRX* should have comparable embryonic roles to ETCHbox. Specifically, we hypothesized that ectopic expression of metatherian CRX protein would elicit similar downstream effects to ectopic expression of ETCHbox genes. As ETCHbox genes subsequently underwent further duplication and diversification they may have gained or lost roles, but some similarities should remain. In contrast, if the embryonic functions of ETCHbox genes are a novelty that arose by neofunctionalisation, expression of either metatherian or eutherian *CRX* would have equally dissimilar downstream effects to ETCHbox genes. To test these predictions, we took the lists of embryonic genes differentially regulated by ectopic *CRX* expression, and compared these to similarly derived lists for mouse ETCHbox genes (Royall *et al.*, 2018). Figure 3.6a displays these comparisons as the percentage of ETCHbox responsive genes (differentially expressed genes, DEGs) that are also *CRX* responsive. These lists will include direct

and indirect targets. Figure 3.6b shows the data displayed as absolute numbers of shared responsive genes. These comparisons show that, for all ETCHbox genes examined, the overlap in responsive genes is larger for metatherian *CRX* than for eutherian *CRX*. To test whether the latter conclusion is driven primarily by transcriptomic profile or is an artefact of sample size (eutherian *CRX* expression affects fewer genes globally than does metatherian *CRX*), we used Fisher's exact test. This showed that overlap between metatherian *CRX* and murine Obox genes is stronger than that for eutherian *CRX* even when sample size is accounted for (Figure 3.6a) . We conclude that metatherian *CRX* affects the expression of a similar, but not identical, set of targets as do *Crxos* and Obox genes (mouse ETCHbox genes).

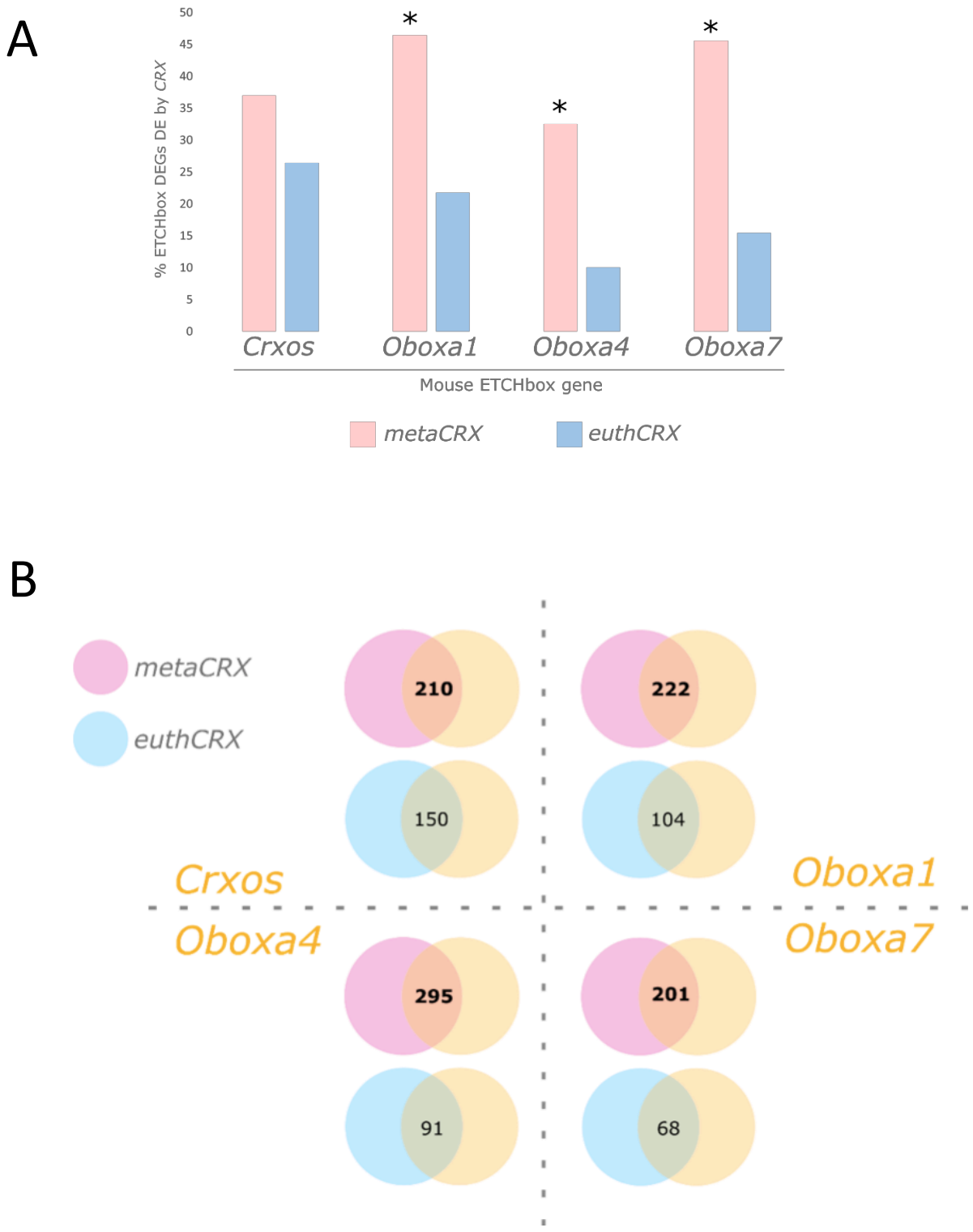


Figure 3.6 Gene sets downstream of mouse ETCHbox genes are more similar to gene sets downstream of metatherian *CRX* than eutherian *CRX*. Comparing lists of differentially regulated genes shows more genes are commonly regulated by metatherian *CRX* and each ETCHbox gene than by eutherian *CRX* and ETCHbox genes. (A) Percentage of DEGs following ETCHbox ectopic expression also regulated following *CRX* ectopic expression. Analyses show that the overlaps between Obox DEGs and *metaCRX* DEGs are stronger than the overlaps between Obox DEGs and *euthCRX* DEGs when total sample size is taken into account (indicated by *). (B) Numbers of genes in overlap between *metaCRX* and *euthCRX* with four mouse ETCHbox genes are shown by Venn diagrams.

As metatherian *CRX* and mouse ETCHbox genes are capable of eliciting similar transcriptomic effects, we asked what embryonic functions might be the basis of the similarity. We focussed on genes with well-characterised roles in early mammalian embryogenesis. Metatherian *CRX* ectopic expression leads to differential expression of genes with reported roles in early eutherian development including maternal communication (*Hbegf*), down-regulation of the maternal inflammatory response (*Ptgs2*), and chromatin remodelling (*Smarca1*) among others (supplementary material 3.1G). Each of these genes was also regulated by one of more mouse ETCHbox genes, but none are regulated following eutherian *CRX* gene over-expression. Furthermore, we repeated transfections of all eutherian *CRX*, metatherian *CRX* and *Oboxa1* and treated samples in parallel. Transcriptome analysis by qPCR shows that *Hbegf*, *Ptgs2,* and *Smarca1* are regulated by metatherian *CRX* and *Oboxa1* in the same direction, emphasizing similarity of likely embryonic roles. Again, similar regulation following eutherian *CRX* expression is not observed (figure 3.7).

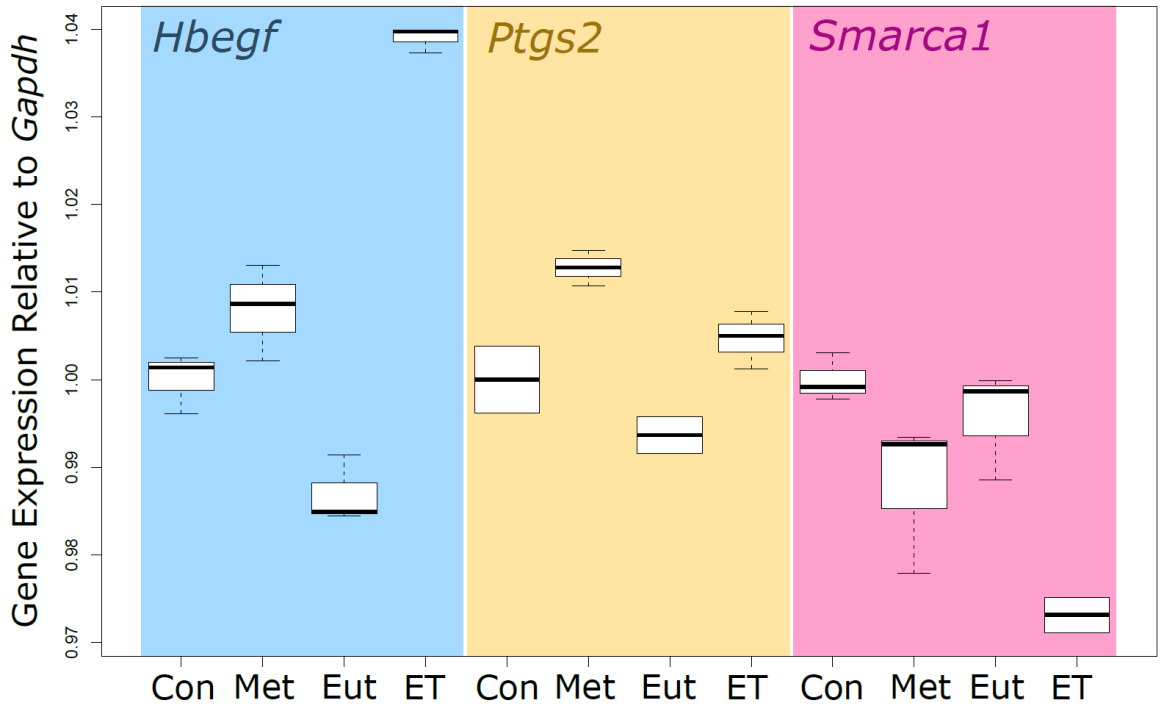


Figure 3.7 Genes with known roles in early eutherian development are similarly regulated following metatherian *CRX* and ETCHbox ectopic expression. A qPCR assay confirms that following metatherian *CRX* or *Oboxa1* ectopic expression, RNA levels of both *Hbegf* and *Ptgs2* are increased, but following eutherian *Crx* ectopic expression neither are up-regulated. Similarly following ETCHbox and metatherian *Crx*, but not eutherian *Crx*, ectopic expression Smarca1 expression decreases. Con = Control, Met = metatherian *CRX*, Eut = eutherian *CRX*, ET = ETCHbox gene (*Oboxa1*).

4. Discussion

The duplication, degeneration and complementation (DDC) model of gene evolution postulates that duplication of a pleiotropic gene leads to functional redundancy, followed by stochastic mutation of different functional elements in each duplicate gene and distribution of ancestral gene functions between the duplicates (Force *et al.*, 1999). This can be an intermediate step in the evolution of new or optimized gene functions, because if a pleiotropic gene is deployed for two or more roles in the same organism it might not be possible for natural selection to optimise the gene for each role simultaneously. This resolution from compromise has been termed 'Escape from Adaptive Conflict' (Hittinger and Carroll, 2007). Alternatively, when a new gene originates by a duplication event, it could acquire entirely new functions (or additional functions) through 'neofunctionalisation'. Distinguishing between these alternatives is crucially important if we wish to understand the origin and evolution of biological traits. However, they are not always straightforward to distinguish as this needs comparable information from outgroups, ideally close outgroups.

There is currently much interest in the processes that underpin early development of humans, from fertilization and embryonic genome activation through to the determination of cell fates, as these events have practical relevance to stem cell biology and assisted reproductive technologies. The ETCHbox genes of humans and other eutherian mammals are expressed after embryonic genome activation and before cell fate determination and may be central to these processes (Töhönen *et al.*, 2015; Maeso *et al.*, 2016; Boroviak *et al.*, 2018; Royall *et al.*, 2018). We wished to understand how these gene functions arose in evolution as,

paradoxically, their most closely related and parental gene (*CRX*) is expressed almost exclusively in the eye. Using marsupials (metatherians) as an outgroup, we found evidence that after the duplication of *CRX* in eutherians, subfunctionalisation occurred. The *CRX* gene retained an ancestral role in the eye, while the highly divergent ETCHbox genes retained ancestral functions in the early embryo; our data argue that the single *CRX* gene of marsupials has dual properties, aligning with the DDC model (figure 3.8).

Subfunctionalisation can occur through regulatory sequence evolution and/or coding sequence evolution. We find evidence for both. Changes in gene expression are indicative of regulatory evolution and it is striking that metatherian *CRX* is expressed in the very early stages of embryogenesis of the fat-tailed dunnart, as well as in adult eye, whereas in eutherian mammals *CRX* expression is limited to the eye. Coding sequence change is evident in the homeodomain of eutherian CRX, suggesting that sequence-specific DNA binding or transcription factor activity was altered in association with expression restriction to eye. Changes to the encoded protein sequence clearly have functional relevance because when each is expressed in an ectopic situation, the eutherian and metatherian CRX proteins elicit different downstream effects. These differences could only be caused by protein sequence differences. When expressed ectopically in cultured cells, we found that a metatherian CRX protein caused a significant reduction in expression of sets of genes that would also normally fall in expression in the preimplantation embryo. Expressing eutherian CRX protein did not elicit the same transcriptomic response. These experiments also showed that the transcriptional effect of metatherian *CRX* ectopic expression has similarities to the effect of ETCHbox ectopic expression. For example, the ETCHbox gene *Oboxa1* and the metatherian *CRX* cause the same effect

on the expression of candidate genes with known importance in both metatherian and eutherian embryo-maternal communication, *Hbegf* and *Ptgs2* (Griffith *et al.*, 2017), and on components of the SWI/SNF chromatin remodelling factor, *Smarca1*. We propose that metatherian *CRX*, and by extrapolation *CRX* of the common therian ancestor, functioned in the early embryo and the eye; following tandem gene duplication in eutherians, eye-associated functions were retained by one gene and embryonic functions were retained by a larger array of duplicates following DDC. Paradoxically, these latter duplicates, the ETCHbox genes, diverged greatly in sequence despite apparently retaining some ancestral roles.

Under the classical DDC model, one gene with two functions duplicates and each duplicate takes on one of the functions and become specialised. However, in the case studied here the fate of the duplicated genes has been more complex. The *CRX* gene duplication events in eutherians gave rise to many gene copies. One of the copies, referred to by the same name as the ancestral gene *CRX*, underwent a rather limited degree of sequence change (although this did alter protein function, as discussed above). The other gene copies underwent extreme sequence change, and are barely recognisable as duplicates of *CRX* (Maeso *et al.*, 2016). These are the genes that inherited the embryonic functions from *CRX* in ancestral mammal.

A key question is why eutherians deploy multiple *CRX* -derived genes in early development when metatherian development occurs with expression of a single *CRX* orthologue? And why have these *CRX* -derived genes diverged so radically in protein sequence from the ancestral sequence and from each other? We do not have clear answers to these questions, but they may lie in differences in early embryonic development. For example, eutherians have an extended gestation compared to

metatherians, in which the embryo implants into the maternal tissue through an invasive (though quite variable) placenta. This may require regulation of additional gene activities, not present in metatherians. In addition there are morphological differences between the embryos themselves; metatherian embryos do not have an inner cell mass at the blastocyst stage, instead having embryo-fated cells as a pluriblast on the surface, nor do they undergo cellular compaction (Renfree, 2010; Frankenberg *et al.*, 2013, 2016; Renfree and Shaw, 2014). These distinctions are candidate explanations for asymmetrical gene divergence of *CRX* and ETCHbox genes in eutherians, and for selective retention of multiple duplicates. The variation in copy number between eutherian species might also relate to variation in placental structure, or it may be an evolutionary consequence of partial redundancy (Royall *et al.*, 2018). This reasoning implies that the evolution of these genes in eutherians most likely involved more than the DDC mechanisms that we uncover here. In addition to ancestral pleiotropic roles being divided between the descendent genes, we propose that ETCHbox genes acquired additional functions to 'fine-tune' embryo development, and for additional embryo-maternal interactions, that were not present in the ancestral *CRX* gene. Therefore, rather than straightforward subfunctionalisation under a DDC model, we suggest that ETCHbox genes have secondarily become entwined in regulatory networks specific to eutherian embryos through specialisation (figure 3.8). We term this subfunctionalisation-specialisation mechanism DDC+.

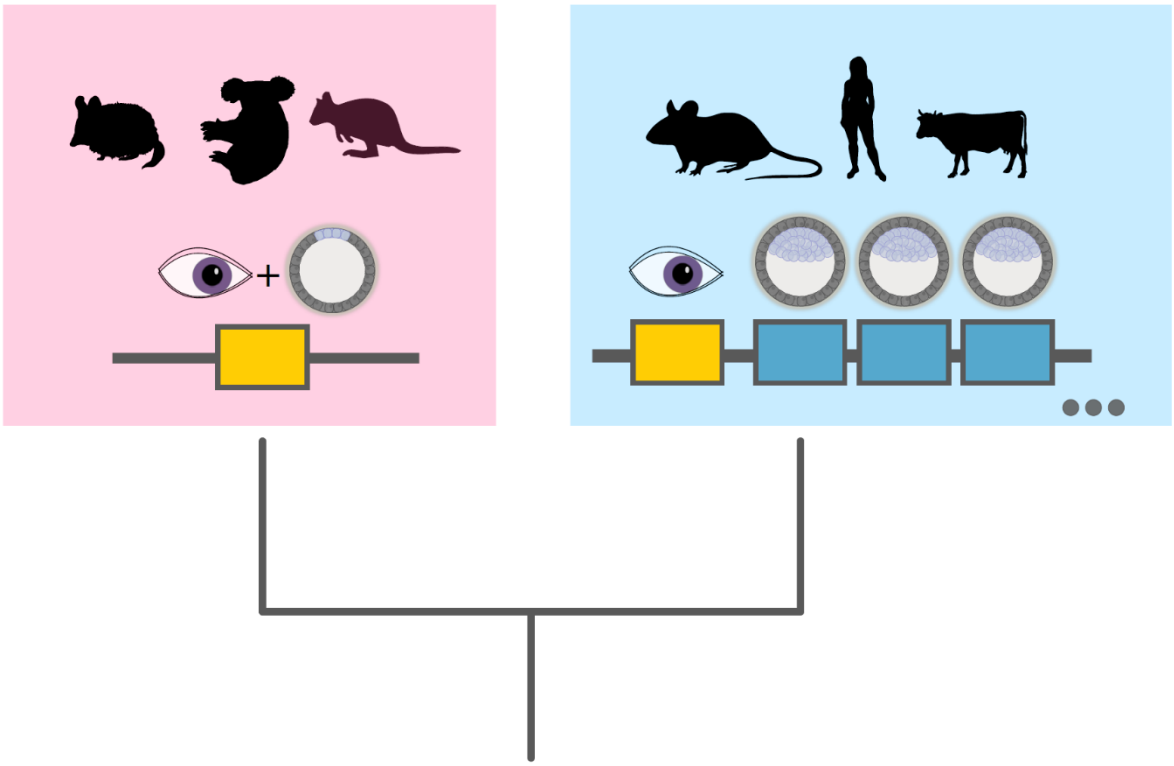


Figure 3.8 Duplication of *CRX* in eutherian lineage lead to specialisation of duplicates into embryonic or retina-specific functions. In metatherians, *CRX* (yellow) continues to carry out both functions; in eutherians, embryonic roles are retained and elaborated on by ETCHbox genes (blue) while eutherian *CRX* retains functions in the eye.

We conclude that following *CRX* duplication in the evolution of eutherian mammals, aspects of classical DDC-mediated subfunctionalisation occurred. However the influence of additional gene duplication events and extreme sequence divergence allowed incorporation of additional functions, thus adding a component of neofunctionalisation on top of subfunctionalisation and the realisation of DDC+.

5. References

Andzelm, M. M. *et al.* (2015) 'MEF2D drives photoreceptor development through a genome-wide competition for tissue-specific enhancers', *Neuron*. Elsevier Inc., 86(1), pp. 247–263. doi: 10.1016/j.neuron.2015.02.038.

Ashkenazy, H. *et al.* (2012) 'FastML: A web server for probabilistic reconstruction of ancestral sequences', *Nucleic Acids Research*, 40(W1), pp. 580–584. doi: 10.1093/nar/gks498.

Blakeley, P. *et al.* (2015) 'Defining the three cell lineages of the human blastocyst by single-cell RNA-seq', *Development*, 142(20), pp. 3613–3613. doi: 10.1242/dev.131235.

Booth, H. A. F. and Holland, P. (2004) 'Eleven daughters of NANOG', *Genomics*, 84(2), pp. 229–238. doi: 10.1016/j.ygeno.2004.02.014.

Boroviak, T. *et al.* (2018) 'Global transcriptome analysis reveals conserved and divergent features of primate preimplantation development', *Development*.

Bray, N. L. *et al.* (2016) 'Near-optimal probabilistic RNA-seq quantification', *Nature Biotechnology*, 34(5), pp. 525–527. doi: 10.1038/nbt.3519.

Dobin, A. *et al.* (2013) 'STAR: Ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Force, A. *et al.* (1999) 'Preservation of Duplicate Genes by Complementary , Degenerative Mutations'.

Frankenberg, S. *et al.* (2013) 'Early cell lineage specification in a marsupial: a case for diverse mechanisms among mammals', *Development*, 140(5), pp. 965–975. doi: 10.1242/dev.091629.

Frankenberg, S. R. *et al.* (2016) 'The mammalian blastocyst', *Wiley Interdisciplinary Reviews: Developmental Biology*, 5(2), pp. 210–232. doi: 10.1002/wdev.220.

Freund, C. L. *et al.* (1997) 'Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor', *Cell*, 91(4), pp. 543–553. doi: 10.1016/S0092-8674(00)80440-7.

Germot, A. *et al.* (2001) 'Structural evolution of Otx genes in craniates.', *Molecular biology and evolution*, 18(9), pp. 1668–1678.

Goldshmit, Y. *et al.* (2010) 'Anatomical changes in the primary visual cortex of the congenitally blind Crx-/- mouse', *Neuroscience*. Elsevier Inc., 166(3), pp. 886–898. doi: 10.1016/j.neuroscience.2009.12.039.

Grabherr, M. G. *et al.* (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome.', *Nature biotechnology*, 29(7), pp. 644–52. doi: 10.1038/nbt.1883.

Griffith, O. W. *et al.* (2017) 'Embryo implantation evolved from an ancestral inflammatory attachment reaction', *Proceedings of the National Academy of Sciences*, p. 201701129. doi: 10.1073/pnas.1701129114.

Hittinger, C. T. and Carroll, S. B. (2007) 'Gene duplication and the adaptive evolution of a classic genetic switch', *Nature*, 449(7163), pp. 677–681. doi: 10.1038/nature06151.

Holland, P. W. H. *et al.* (2017) 'New genes from old: asymmetric divergence of gene duplicates and the evolution of development', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1713), p. 20150480. doi: 10.1098/rstb.2015.0480.

Huang, D. W., Sherman, Brad T. and Lempicki, R. A. (2009) 'Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Research*, 37(1), pp. 1–13. doi: 10.1093/nar/gkn923.

Huang, D. W., Sherman, Brad T and Lempicki, R. A. (2009) 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.', *Nature protocols*, 4(1), pp. 44–57. doi: 10.1038/nprot.2008.211.

Kumar, L. and Futschik, M. (2007) 'Bioinformation Mfuzz : A software package for soft clustering of microarray data Bioinformation', pp. 5–7.

Kuroda, H. *et al.* (2000) 'Cloning a novel developmental regulating gene, Xotx5: Its potential role in anterior formation in Xenopus laevis', *Development Growth and Differentiation*, 42(2), pp. 87–93. doi: 10.1046/j.1440-169X.2000.00491.x.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', pp. 1–21. doi: 10.1186/s13059-014-0550-8.

Madissoon, E. *et al.* (2016) 'Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos', *Scientific Reports*. Nature Publishing Group, 6(July), pp. 1–14. doi: 10.1038/srep28995.

Maeso, I. *et al.* (2016) 'Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals', *BMC Biology*, 14(1), pp. 1–14. doi: 10.1186/s12915-016-0267-0.

Plouhinec, J. L. *et al.* (2003) 'The mammalian Crx genes are highly divergent representatives of the Otx5 gene family, a gnathostome orthology class of orthodenticle-related homeogenes involved in the differentiation of retinal photoreceptors and circadian entrainment', *Molecular Biology and Evolution*, 20(4), pp. 513–521. doi: 10.1093/molbev/msg085.

Plouhinec, J. L. *et al.* (2005) 'Comparative analysis of gnathostome Otx gene expression patterns in the developing eye: Implications for the functional evolution

of the multigene family', *Developmental Biology*, 278(2), pp. 560–575. doi: 10.1016/j.ydbio.2004.11.019.

Rajkovic, A. *et al.* (2002) 'Obox, a family of homeobox genes preferentially expressed in germ cells', *Genomics*, 79(5), pp. 711–717. doi: 10.1006/geno.2002.6759.

Renfree, M. B. (2010) 'Review: Marsupials: Placental Mammals with a Difference', *Placenta*. Elsevier Ltd, 31(SUPPL.), pp. S21–S26. doi: 10.1016/j.placenta.2009.12.023.

Renfree, M. B. and Shaw, G. (2014) 'Embryo-endometrial interactions during early development after embryonic diapause in the marsupial tammar wallaby', 181(July), pp. 175–181. doi: 10.1387/ijdb.140059mr.

Royall, A. H. *et al.* (2018) 'Mouse Obox and Crxos modulate preimplantation transcriptional profiles revealing similarity between paralogous mouse and human homeobox genes', *EvoDevo*. BioMed Central, 9(1), p. 2. doi: 10.1186/s13227-018-0091-4.

Saito, K. *et al.* (2011) 'Relationships between homeoprotein EGAM1C and the expression of the placental prolactin gene family in mouse placentae and trophoblast stem cells', *Reproduction*, 141(2), pp. 259–268. doi: 10.1530/REP-10-0355.

Sauka-Spengler, T. *et al.* (2001) 'Structure and expression of an Otx5-related gene in the dogfish Scyliorhinus canicula: Evidence for a conserved role of Otx5 and Crx genes in the specification of photoreceptors', *Development Genes and Evolution*, 211(11), pp. 533–544. doi: 10.1007/s00427-001-0191-2.

Shen, Y. C. and Raymond, P. A. (2004) 'Zebrafish cone-rod (crx) homeobox gene promotes retinogenesis', *Developmental Biology*, 269(1), pp. 237–251. doi: 10.1016/j.ydbio.2004.01.037.

Simeone, A. *et al.* (1992) 'Nested expression domains of four homeobox genes in developing rostral brain', *Nature*, 358(6388), pp. 687–690. doi: 10.1038/358687a0.

Töhönen, V. *et al.* (2015) 'Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development', *Nature Communications*, 6, pp. 1–9. doi: 10.1038/ncomms9207.

Viczian, A. S. (2003) 'XOtx5b and XOtx2 regulate photoreceptor and bipolar fates in the Xenopus retina', *Development*, 130(7), pp. 1281–1294. doi: 10.1242/dev.00343.

Vignali, R. *et al.* (2000) 'Xotx5b, a new member of the Otx gene family, may be involved in anterior and eye development in Xenopus laevis', *Mechanisms of Development*, 96(1), pp. 3–13. doi: 10.1016/S0925-4773(00)00367-1.

Villanueva-Cañas, J. L. *et al.* (2017) 'New genes and functional innovation in mammals', *Genome Biology and Evolution*, 9(7), pp. 1886–1900. doi:

10.1093/gbe/evx136.

Wang, L. *et al.* (2018) 'A novel CRX frameshift mutation causing cone-rod dystrophy in a Chinese family: A case report', *Medicine*, 97(32), p. e11499. doi: 10.1097/MD.0000000000011499.

Yang, Z. (2007) 'PAML 4 : Phylogenetic Analysis by Maximum Likelihood', 24(Ml), pp. 1586–1591. doi: 10.1093/molbev/msm088.

6. Supplementary files

3.1 (Excel files)

(A) FastML output for deducing ancestral vertebrate CRX sequence.

(B) RNA expression (TPM) of *CRX*, *OTX1*/2, and ETCHbox across fat-tailed dunnart and mouse early embryogenesis.

(C) Lists of genes differentially expressed following metatherian and eutherian *CRX* ectopic expression in MEFs.

(D) GO terms significantly enriched in the set of genes differentially expressed following human *CRX* ectopic expression.

(E) GO terms significantly enriched in the set of genes differentially expressed following opossum *CRX* ectopic expression.

(F) List of mouse genes dynamically expressed (FPKM variance >5) across mouse preimplantation development.

(G) Genes differentially expressed following metatherian *CRX* ectopic expression with known roles in early eutherian development.

3.2 (PDF files)

(A) SRA identifiers for mouse preimplantation transcriptome datasets.

(B) Maximum-likelihood tree of CRX protein sequences from across metazoans shows divergence of sequence in eutherian mammals.

(C) Homeodomains of mouse ETCHbox genes do not show compensatory changes for the amino acids substituted in eutherian CRX homeodomain.

(D) RT-PCR for *CRX* expression in fat-tailed dunnart adult tissues.

(E) Immunocytochemistry staining for V5-tagged CRX proteins in MEFs shows nuclear localisation and demonstrates correct translation frame.

(F) TPM for ectopically expressed genes and chemical selection gene (puromycinR) shows there is no discrepancy in the levels of ectopic expression between samples, and that there is no endogenous expression of mouse *Crx* in the MEFs used

(G) Mouse preimplantation embryo transcriptome profiles generated using MFuzz.

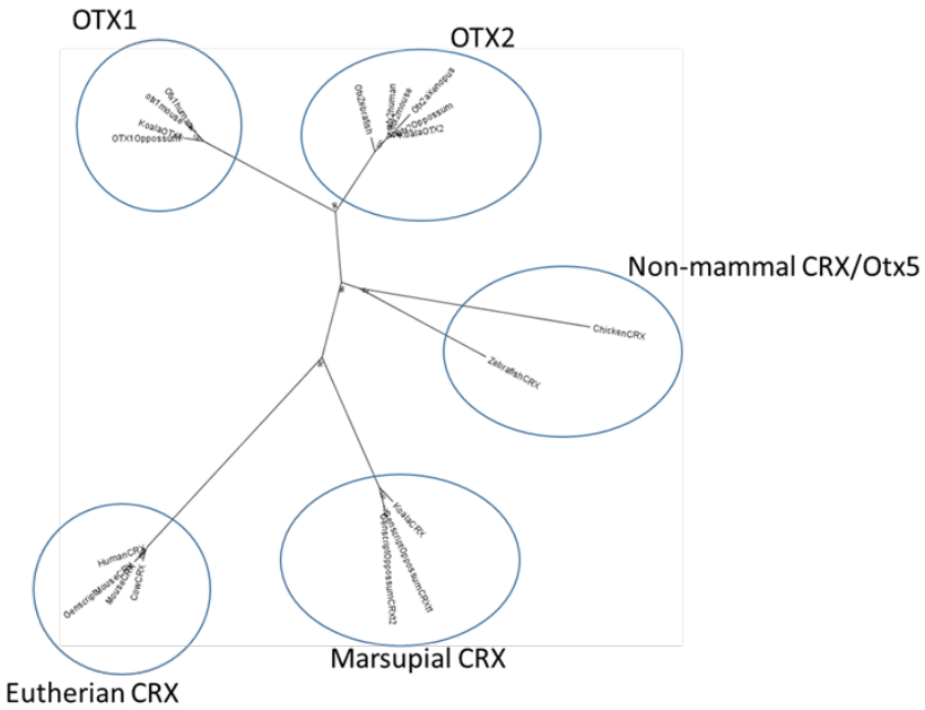
supplementary file 3.1 A-G is available online at https://royalsocietypublishing.org/doi/full/10.1098/rspb.2019.0830

supplementary file 3.2:

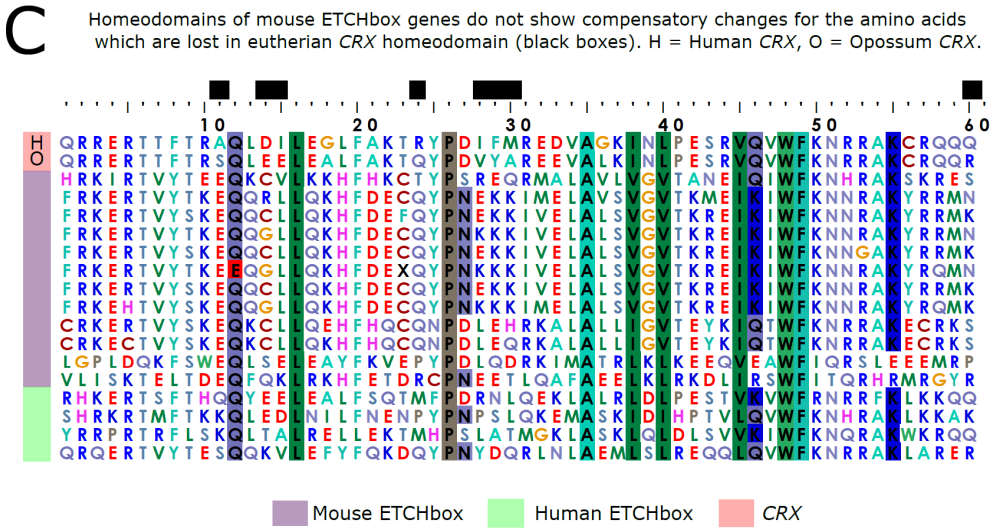
(A) SRA identifiers for mouse preimplantation transcriptome datasets.

Species	Tissue	SRA study	Run ID
Mmus	Oocyte	SRP03454	SRR1051273
Mmus	Oocyte	SRP03454	SRR1051274
Mmus	Oocyte	SRP03454	SRR1051275
Mmus	Oocyte	SRP03454	SRR1051276
Mmus	Oocyte	SRP03454	SRR1051277
Mmus	Zygote	SRP03454	SRR1567937
Mmus	Zygote	SRP03454	SRR1051279
Mmus	Zygote	SRP03454	SRR1051278
Mmus	2Cell	SRP03454	SRR1051282
Mmus	2Cell	SRP03454	SRR1051281
Mmus	2Cell	SRP03454	SRR1051280
Mmus	4Cell	SRP03454	SRR1567921
Mmus	4Cell	SRP03454	SRR1567920
Mmus	4Cell	SRP03454	SRR1567919
Mmus	8Cell	SRP03454	SRR1567924
Mmus	8Cell	SRP03454	SRR1567923
Mmus	8Cell	SRP03454	SRR1567922
Mmus	Morula	SRP03454	SRR2048257
Mmus	Morula	SRP03454	SRR2048256
Mmus	Morula	SRP03454	SRR2048255
Mmus	Blastocyst	SRP03454	SRR1567925
Mmus	Blastocyst	SRP03454	SRR1567926
Mmus	Blastocyst	SRP03454	SRR1567927
Mmus	Blastocyst	SRP03454	SRR1567928
Mmus	E7.5	SRP03776	SRR1168503

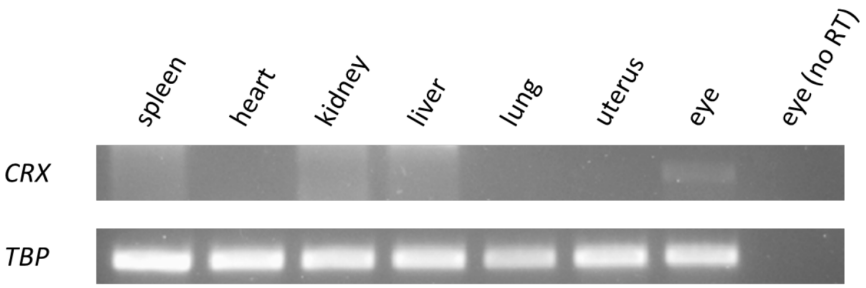
(B) Maximum-likelihood tree of CRX protein sequences from across metazoans shows divergence of sequence in eutherian mammals.



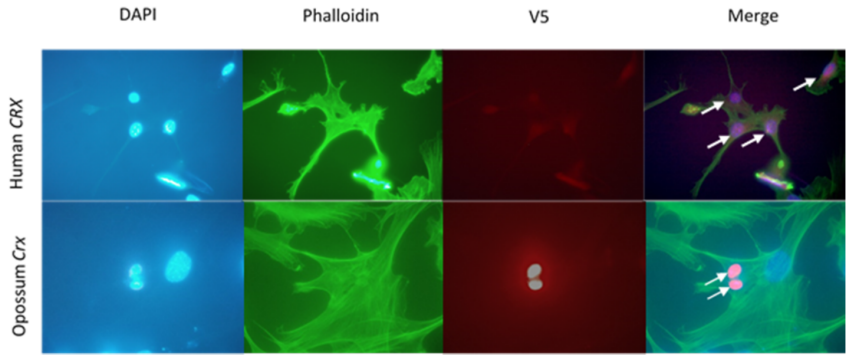
(C) Homeodomains of mouse ETCHbox genes do not show compensatory changes for the amino acids substituted in eutherian CRX homeodomain.



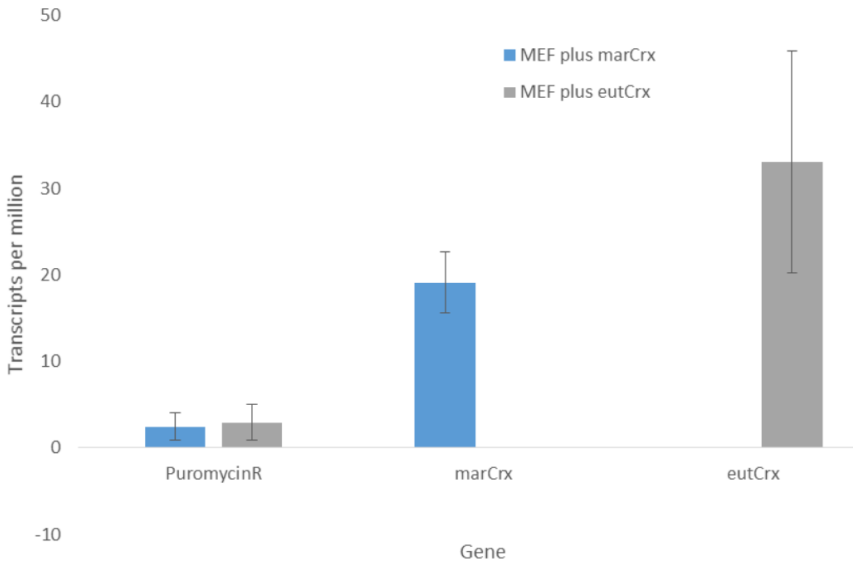
(D) RT-PCR for *CRX* expression in fat-tailed dunnart adult tissues.



(E) Immunocytochemistry staining for V5-tagged CRX proteins in MEFs shows nuclear localisation and demonstrates correct translation frame.



(F) TPM for ectopically expressed genes and chemical selection gene (puromycinR) shows there is no discrepancy in the levels of ectopic expression between samples, and that there is no endogenous expression of mouse *Crx* in the MEFs used.



(G) Mouse preimplantation embryo transcriptome profiles generated using MFuzz.

Available online at https://royalsocietypublishing.org/doi/full/10.1098/rspb.2019.0830

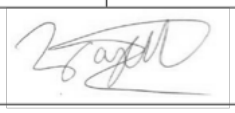
Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Of eyes and embryos; subfunctionalisation of CRX homeobox gene in mammalian evolution
Publication Status	Published
Publication Details	Royall AH, Frankenberg S, Pask AJ, Holland PWH, (2019), Of eyes and embryos: subfunctionalisation of the *CRX* homeobox genes in mammalian evolution, Proceedings of the Royal Society B 286, DOI: 10.1098/rspb.2019.0830

Student Confirmation

Student Name:	Amy H Royall		
Contribution to the Paper	Statement of contribution within the paper		
Signature		Date	08/09/2019

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Peter Holland, Linacre Professor of Zoology		
Supervisor comments Amy Royall was primary contributor to this paper. Exact details of contributions detailed in the paper		
Signature 	Date	12/9/19

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 4: ETCHbox variability and evolution in Rodentia

Author Contributions

This project and its aims were conceived by Amy H Royall and Peter WH Holland. All genomic and sequence analysis presented here was conducted by Amy H Royall. Sequences from a previous (unpublished) study were provided by Dr Ignacio Maeso (CABD, Universidad Pablo de Olavide, Seville), who also provided guidance on which genomes to analyse. Figures were created and text was written by Amy H Royall with feedback from Peter WH Holland.

Abstract

When studying the function and evolution of genes it is important to understand the wider evolutionary context in which they exist. This is the case particularly when the evolutionary history of the genes of interest is complex and dynamic because gene family evolution can directly influence gene function. The ETCHbox family is highly dynamic and complex. They are specific to eutherian mammals and in temporal expression pattern. However they are also highly variable and dynamic in gene copy number. For example, mouse has only retained two of the six known ETCHbox genes and has duplicated both. In this chapter I aim to understand this dynamic nature further through an in depth analysis of ETCHbox genes in rodents.

I show that it is not only the mouse genome that has lost most ETCHbox genes and duplicated both *Tprx1* and *Tprx2*; the rat genome has a similar gene organisation and repertoire. Furthermore, I find that the gene losses seen in murid rodents occurred stepwise through rodent evolution, and there are other species that have experienced additional independent losses. I find that the sequence divergence of *Tprx2* towards that of Obox genes did not occur specifically in the mouse lineage, but actually much deeper in the rodent evolutionary history, and that Obox genes do not necessarily have to exist in clusters as they do in mouse.

This analysis has revealed that the ETCHbox family are more dynamic than previously thought. Indeed, the mouse genome has not experienced particularly unusual levels of ETCHbox gene loss or sequence divergence. However there are differences in ETCHbox genes in even closely related species, indicating that evolution of this family is fast and ongoing.

1. Introduction

The ETCHbox genes encode homeodomain transcription factors that duplicated from the conserved Cone-Rod homeobox (*CRX*) gene in the lineage leading to eutherian mammals. The genomes of eutherian's closest outgroup, the marsupials, do not have any ETCHbox genes whereas studied members of all four eutherian subclasses have at least two, and up to five, of a maximum six known ETCHbox genes (Madissoon *et al.*, 2016; Maeso *et al.*, 2016). The six genes (*ARGFX*, *TPRX1*, *TPRX2*, *LEUTX*, *DRPX*, and *PARGFX*) reside at differing distances from *CRX,* with *TPRX1* and *TPRX2* flanking the ancestral loci, but all were firmly identified as duplicates due to conservation of a non-coding element downstream of the coding sequence (Maeso *et al.*, 2016).

The ETCHbox genes are interesting for a number of reasons, not least because they exhibit remarkable gene copy number and sequence variation across closely related species (Maeso *et al.*, 2016). The sequence changes have been so dramatic that a number of the ETCHbox genes have previously been described as species-specific (Holland, Booth and Bruford, 2007; Wilming, Boychenko and Harrow, 2015) as bioinformatics approaches fail to identify loci with high levels of similarity. Not only have two ETCHbox genes, *Tprx1* and *Tprx2*, been incorrectly described as absent in the mouse genome, their true orthologues have been officially named *Crxos* and Obox respectfully, and have themselves been described as mouse-specific (Zhong and Holland, 2011; Wilming, Boychenko and Harrow, 2015; Maeso *et al.*, 2016).

The change in sequence is not the only interesting feature of mouse ETCHbox genes. The mouse mm10 genome at the time of writing contains just two of the six

ETCHbox genes (*Crxos* and Obox), however both of these loci have independently duplicated. The *Crxos* locus has tandemly duplicated to give two adjacent transcripts containing a single homeobox each, and a third composite transcript which contains exons of each duplicate including both homeobox sequences. The Obox locus however has duplicated to give the largest homeobox cluster described to date in mouse (MacLean *et al.*, 2005; Holland, Booth and Bruford, 2007; Zhong and Holland, 2011; Wilming, Boychenko and Harrow, 2015). This cluster contains over 60 individual loci, some of which are pseudogenes, spread across a genomic region of around 1.4 mega-base-pairs (Maeso *et al.*, 2016). This is in contrast to other eutherian genomes which generally display much less gene duplication, but do still exhibit variation. For example, the human genome houses five ETCHbox genes and has lost *PARGFX,* whereas the Guinea pig has a *Leutx* cluster (Maeso *et al.*, 2016; figure 4.1). Another reason the ETCHbox genes are of interest is their very specific expression patterns. To date, expression of ETCHbox genes from any examined species is only found in the very early embryo (Töhönen *et al.*, 2015; Jouhilahti *et al.*, 2016; Madissoon *et al.*, 2016; Maeso *et al.*, 2016; Royall *et al.*, 2018).

Variation in genes utilised at this very early stage of embryogenesis is not completely novel. The 'Developmental Hourglass' theory of the evolution of development describes how the very early and very late stages of embryo development are generally variable throughout metazoans, with a highly conserved and constrained period of development partway through embryogenesis (Kuijk, Geijsen and Cuppen, 2015). Even in the light of this described variation in the earliest stages of development, within eutherians the preimplantation stages of embryogenesis appear very similar, at least visually (Frankenberg *et al.*, 2016). How

can genes with very specific inferred function at this constrained period vary so much between species?

It can be argued that the mouse genome has the most bizarre repertoire of ETCHbox genes known, with huge amounts of gene loss, duplication, and divergence. On the other hand, the guinea pig genome has retained all of the ETCHbox genes found in human, albeit with some lineage specific duplications (Maeso *et al.*, 2016; figure 4.1). This chapter aims to characterise and understand this variability by inferring evolutionary history.

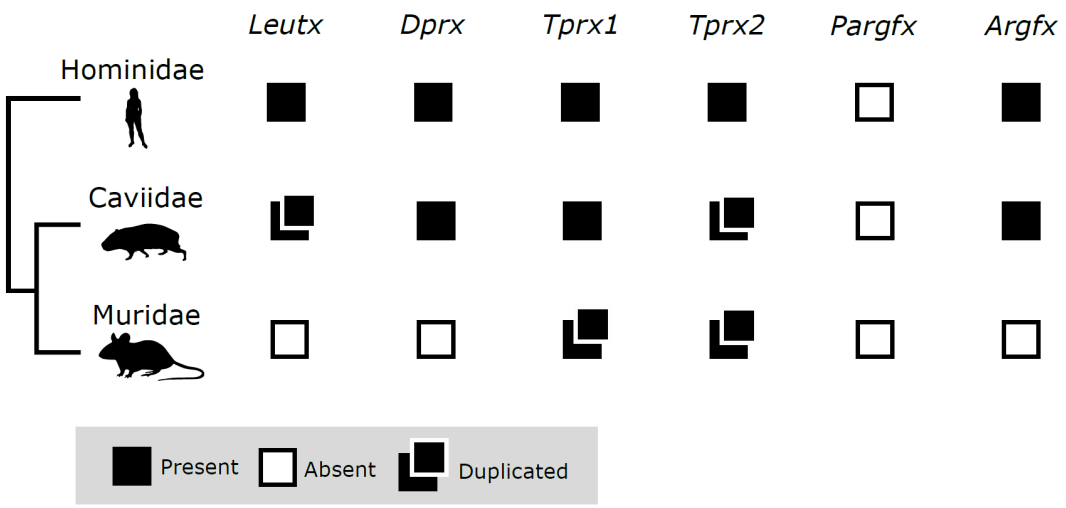


Figure 4.1 Published ETCHbox repertoire in Euarchontoglires shows that there has been numerous gene losses and additional duplications on the murid lineage. However, current taxon sampling is not detailed enough to identify when these losses and duplications occurred; is mouse weird?

Rodentia is the most diverse mammalian order and diverged ~96 mya (Nei, Xu and Glazko, 2001). Within Rodentia are 3 groupings; the "mouse-like clade" containing the suborders Myomorpha and Castorimorpha, the "squirrel-like clade"

containing the suborder Sciuromorpha, and the "guinea pig-like" clade containing the order Hystricomorpha.

Within Myomorpha (mouse-like clade, figure 4.2), the Muridae family includes mice and rats, the Cricetidae family contains hamsters, voles and lemmings, Nesomyidae contains the pouched rat and Malagasy rats, Spalacidae the sub-terrain mole-rats, and Dipodidae the jerboas, birch mice and jumping mice. The family Castoridae (Castorimorpha, mouse-like clade, figure 4.2) includes beavers, kangaroo rats and kangaroo mice.

The family Sciuridae (Sciuromorpha, squirrel-like clade, figure 4.2) contains squirrels and marmots, and the family Caviidae (Hystricomorpha, guinea pig-like clade, figure 4.2) contains guinea-pigs, degus, and chinchillas. It is contested whether Caviidae or Sciuridae is the outgroup for other rodents (Graur, Hide and Li, 1991; Blanga-Kanfi *et al.*, 2009). For the purposes of this chapter I use Caviidae as the outgroup (based on Fabre *et al.*, 2012) as this issue has little effect on the aims and findings of this study, and vice versa. The genomes of at least one representative of each family in figure 4.2 are available for interrogation.

It is unclear from our current knowledge of ETCHbox distribution whether the mouse genome alone is highly divergent from other rodents, potentially all other Euarchontoglires, or whether the path of gene loss and gain can be traced back through other more closely related species. The ETCHbox content of rodents other than the distantly related mouse and guinea pig is not known (figure 4.1).

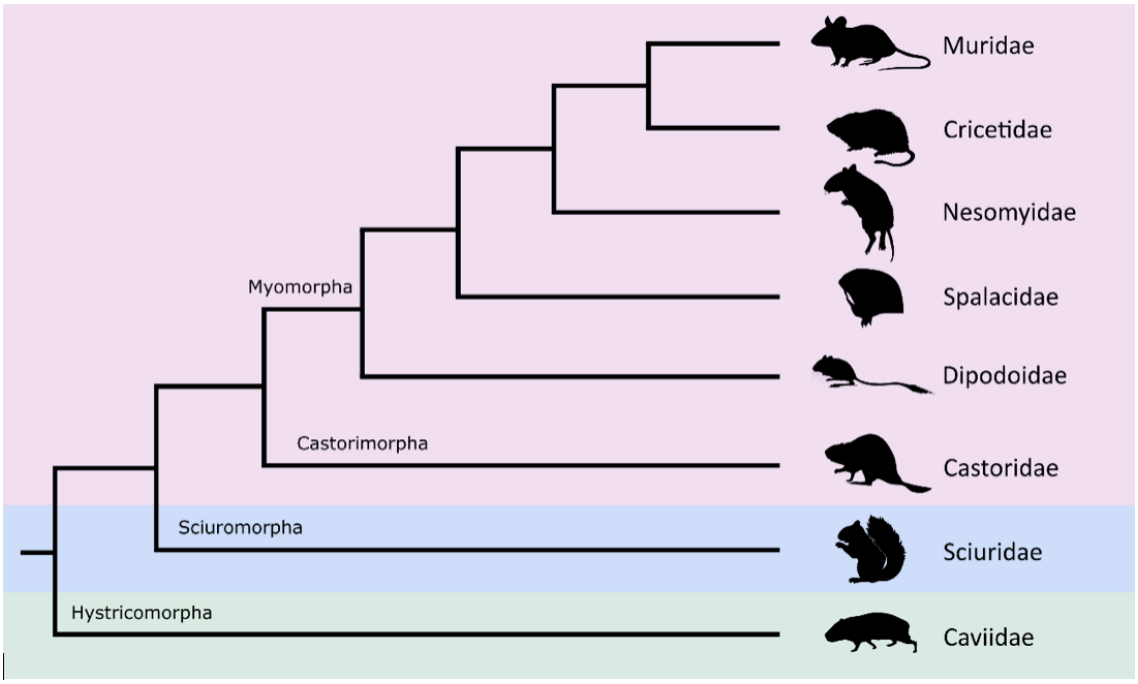


Figure 4.2 Cladogram of the rodent evolutionary history showing representative families with available genome assemblies. The mouse-like clade are in pink, the squirrel-like clade in blue, and the guinea pig-like clade in green.

I will use candidate genomes to ask when the *Tprx1* and *Tprx2* gene divergences described in mouse occurred in rodent evolutionary history. Specifically, I will uncover when *Tprx2* genes as seen in guinea pig and human became mouse 'Obox-like' in amino acid sequence, and whether this coincides with a tendency to form gene clusters. I will also ask when the extreme gene loss seen in the mouse genome occurred, through stepwise searches for loci in species phylogenetically located between mouse and guinea pig (Maeso *et al.*, 2016; figure 4.2). Overall, I aim to assess the amount of genetic change within the ETCHbox group in rodents and study whether the gain, loss, and sequence change observed in mouse occurred in a stepwise fashion. This study will reveal how dynamic ETCHbox gene evolution is through examining much more closely related species than have been studied until now.

2. Materials and Methods

2.1 Genome analyses

Candidate genomes were assessed for assembly quality (N50 value) and sequencing depth and chosen based on the most complete assembly available for each family. All genomes apart from that of *Cricetomys gamianus* are available on Ensembl. *C. gambianus* genome was analysed using the integrated genome viewer (Robinson, 2012) and BLAST (Altschul *et al.*, 1990). Assembly IDs and details are available in supplementary file 4.1. For synteny and genomic organisation analysis a combination of the UCSC genome browser (James Kent, 2002; James Kent *et al.*, 2002) and Ensembl genome browsers (Zerbino *et al.*, 2018) were used. The same browsers were also used to identify annotated predicted genes which refer to or contained the identified ETCHbox sequence. Gene and protein sequences were extracted directly from the browsers for analysis.

2.2 ETCHbox sequence identification

I define the "presence" of an ETCHbox gene as the identification of a homeodomain sequence which in associated with the respective ETCHbox gene in a reciprocal BLASTP (Altschul *et al.*, 1990) analysis using the nr protein database with a filter for eutherian mammal sequences. If a sequence could be identified as a potential ETCHbox sequence but I could not find an associated homeobox it did not meet the criteria but it is contained in supplementary file 4.2. When a predicted gene sequence (GeneScan; Burge and Karlin, 1997) was present in the loci of the identified ETCHbox amino acid sequence, it was extracted and corrected using the confirmed ETCHbox sequences of human, guinea pig and mouse. Information on all identified amino acid sequences used is available in supplementary 4.2 along with

genomic co-ordinates and any associated annotation (such as Genescan ID). Presence of a homeodomain was assessed using the NCBI Conserved Domain search tool (Marchler-Bauer *et al.*, 2015). Phylogenetic analyses used Mafft alignment (Katoh and Standley, 2013) and trimal alignment trimming software (Capella-Gutiérrez, Silla-Martínez and Gabaldón, 2009). Phylogenetic trees used RAxML-HPC-PTHREADS v8.1.1 (Stamatakis, 2014) with 500 bootstraps and PROTGAMMALG model. For *Tprx2*/Obox analysis the PROTCATJTT model was used to account for different species evolution rates.

3. Results

3.1 *Tprx1* organisation in *Mus musculus* is conserved with *Rattus norvegicus*

In order to ask whether the complex ETCHbox gene organisation and extreme duplication observed in mouse is a species-specific phenomenon, or whether it can also be found in other species, I began with the genome of the close relative of mouse, the Brown Norway Rat (*Rattus norvegicus*) with aim of describing the *Tprx* genes present.

The rat *Crx* locus is located on Chromosome 1 arm q21. To locate the region syntenic to mouse *Crxos*, I identified the rat *Sepw1* locus which, along with *Crx*, flanks *Tprx1/Crxos* (figure 4.3). Notably, there are two *Sepw1* loci in the rat genome assembly, one in the region syntenic to mouse *Sepw1*, the other lies on the other side of *Crx* (figure 4.3). Further analysis of these two regions shows that there has been a regional duplication in the rat lineage of around 28 kilobase-pairs (kbp; figure 4.3). The region Chr1: 77783017-77811088 has been duplicated to Chr1: 77509345-77538092 with 99% identity at the nucleotide level. However experimental validation of this will be necessary confirm that it is not an assembly error.

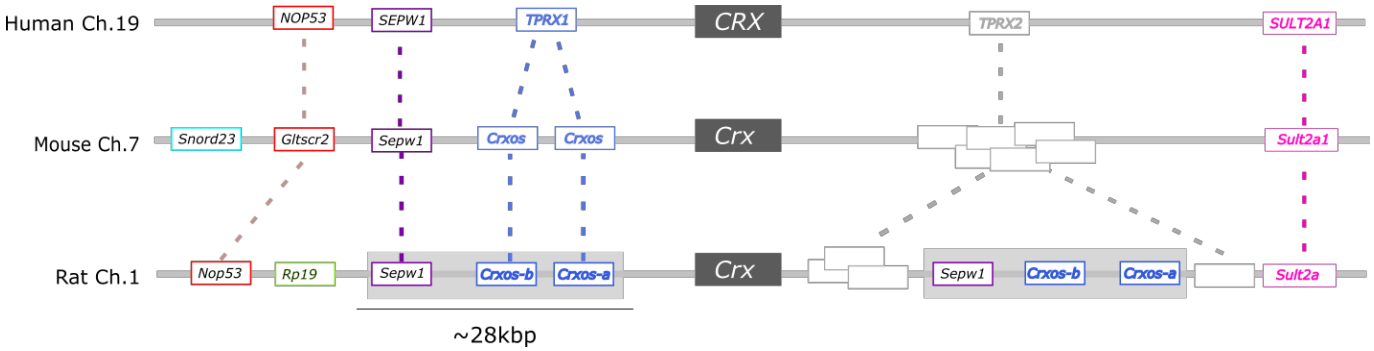


Figure 4.3 Comparison of the *Crx/Tprx* locus in rat to human and mouse. This analysis reveals a rat-specific regional duplication resulting in two identical *Crxos* "regions". The *Crxos* gene duplication in rat is comparable to that in mouse and the two distinct sequences are referred to as *Crxos*-a and *Crxos*-b.

In the rat *Crxos* region, there are two loci with the three exon organisation common to ETCHbox genes, including mouse *Crxos*. The organisation of two homeobox gene loci next to each other, between *Crx* and *Sepw1* is the same organisation as the *Crxos* duplication in mouse, giving strong support that two *Tprx1/Crxos* loci were present in the genome of the last common ancestor (LCA) of mouse and rat (supplementary file 4.2). The 28kbp duplicated region has few nucleotide changes and both *Crxos* loci are contained within this region leading to a total of four *Crxos* genes in the rat genome. The respective duplicates of each rat *Crxos* have 100% nucleotide conservation meaning it is not easily possible to decipher how the duplicates are transcribed. Herein I will refer to the two distinct loci as *Crxos-a* and *Crxos-b* (figure 4.3).

The genomic organisation of *Crxos* is conserved between mouse and rat, and there are EST sequences from adult tissues for *Crxos-b* (Sequence ID BG378186.1, (Bonaldo, Lennon and Soares, 2007)) confirming gene exon-intron structure.

However without comprehensive transcriptome data from the relevant embryonic stages I cannot confirm that the complex transcript structures utilised in mouse (Royall *et al.*, 2018) are conserved.

***3.2 Crxos* sequence evolution**

Following the result that *Crxos* was already duplicated in the LCA of mouse and rat and that it has been conserved in each species, I studied how the protein sequences of the highly divergent genes compare. In a reciprocal TBLASTN one CRXOS (CRXOS-a) protein sequence hits mouse *Crxos* using the *Mus musculus* genome as the search subject. A search for the second CRXOS sequence (CRXOS-b) using *Mus musculus* as the subject finds no hits.

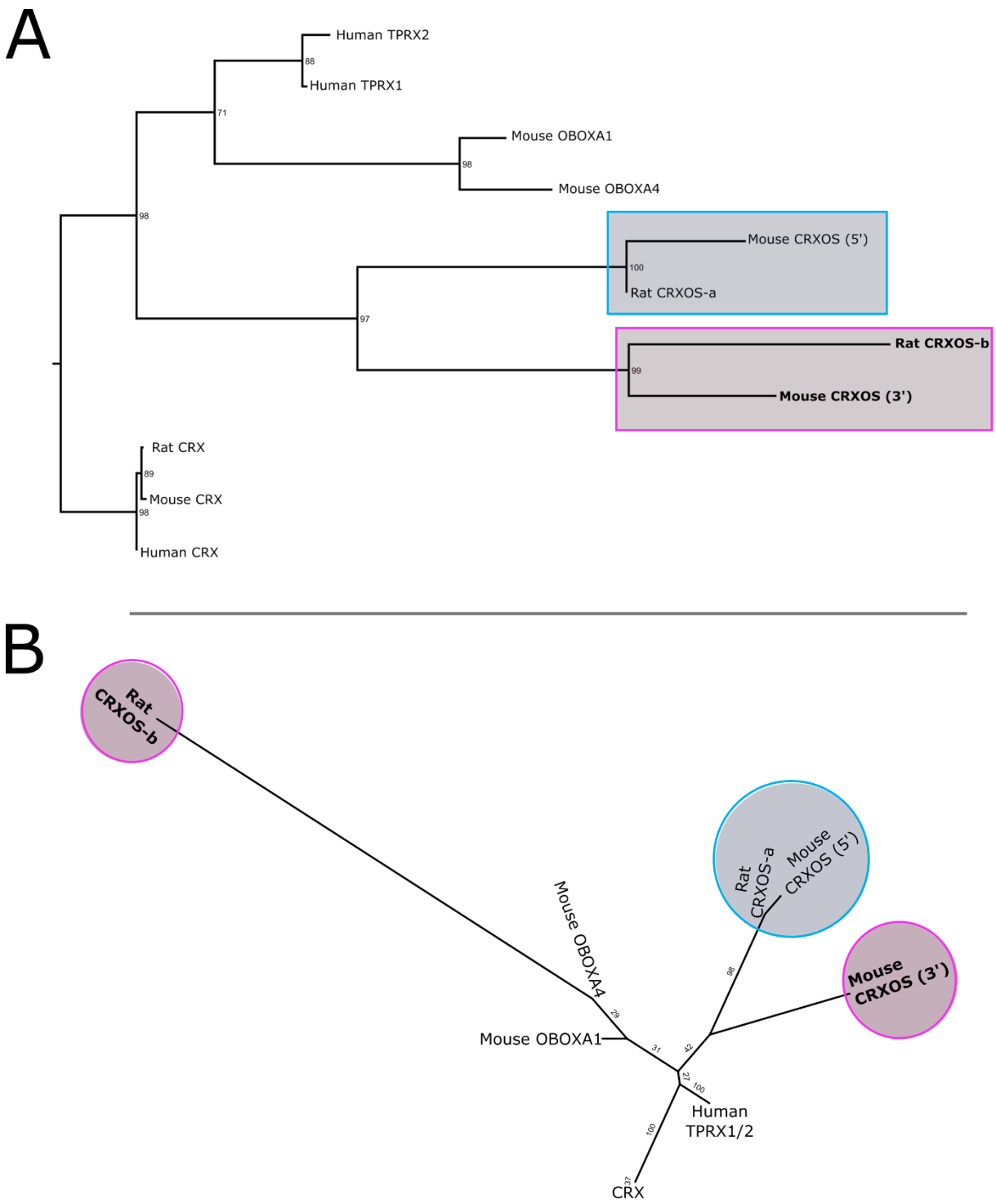


Figure 4.4 Amino acid sequence evolution of eutherian TPRX1. (A) Maximum-likelihood tree for Tprx1 whole protein sequences shows pairing of mouse and rat sequences. (B) Maximum-likelihood tree for Tprx1 homeodomain sequences however shows divergent rat Crxos sequence.

I then assessed the levels of sequence change in the rat *Crxos* loci, undertaking a phylogenetic analysis of the amino acid sequence of both the whole protein (figure 4.4a), and the homeodomain alone (figure 4.4b). The whole protein analysis reveals that each rat *Crxos* gene is paired with a mouse *Crxos* gene, as would be expected for orthologous genes. Interestingly, when I analyse the homeodomains alone, rat CRXOS-b homeodomain is highly divergent (figure 4.4b) suggesting asymmetric evolution of the rat *Crxos* genes in the rat lineage focussed on the homeodomain region. Examination of the trimmed alignment produced in the whole protein analysis shows that the homeodomain sequence has not been excluded however the homeodomain of CRXOS-b is incomplete. Genomic and EST analysis shows that this is due to a seven amino acid deletion rather than incorrect gene structure prediction.

3.3 Obox organisation in rat sheds light on the Obox content of a murine common ancestor

A search for Obox-like genes using the mouse Obox and human *Tprx2* sequences as queries revealed eleven sequences in the rat genome with full homeoboxes. I refer to these loci as Obox-like (Oboxl) (supplementary file 4.2). Five of these loci are within the 1.4mbp region syntenic to where the majority of mouse Obox genes are located, between the *Crx* locus and the *Sult2*/*Sult-like* locus. Additionally, four loci are elsewhere on chromosome 1, outside of the syntenic region, and one on each of chromosome X and 9. The locus on chromosome X (*Oboxl5*) does not contain any introns and so is likely to be a retrotransposed copy of another locus (*Oboxl2* has the most similar sequence). There are no annotated Obox genes or nucleotide sequences in the *Rattus norvegicus* genome assembly

listed on NCBI at the time of writing, however there is one annotated Obox gene in the Ensembl database. The annotated gene is *Obox2* (ENSRNOG00000032831) and in this study is referred to as *Oboxl2* (supplementary file 4.2).

Due to the divergent nature of ETCHbox genes, sequence analyses help to conclusively categorise these sequences as Obox. The amino acid sequences of all eleven rat loci, along with the protein sequences of mouse Obox and CRX, were included in a maximum likelihood phylogeny to identify groupings by sequence similarity (figure 4.5). Phylogenetic analysis including mouse OboxD can be found in supplementary material (supplementary file 4.3), it has been removed here as the complex history of OboxD group (chapter 2, Royall *et al*, 2018) reduces the resulting bootstrap values and stability of the phylogeny. This analysis indeed shows that all eleven rat Oboxl sequences are more similar to mouse Obox than the ETCHbox ancestor CRX (figure 4.5).

Moreover, the phylogeny reveals rat loci group with described mouse groups. OBOXL3 and OBOXL4 group closely with the mouse OBOXA1 sequence, suggesting that at least one *Oboxa1*-like loci existed in the LCA of mouse and rat (figure 4.4). Multiple Obox-like sequences then group more generally with the mouse OboxA group as defined in Royall *et al.*, 2018, similarly there are Obox-like sequences more closely related to mouse OboxB loci than other groups (figure 4.5). The rat OboxA-like sequences sit equally distant from mouse OboxA sequences; OboxAb and OboxAc-like sequences described in mouse (Royall *et al.*, 2018) are not found. Finally, OBOXL7 closely groups with the mouse OBOXC sequence. No Obox-like genes group with the OboxD group (supplementary file 4.3) suggesting they may have arisen in the mouse genome following the divergence from rats.

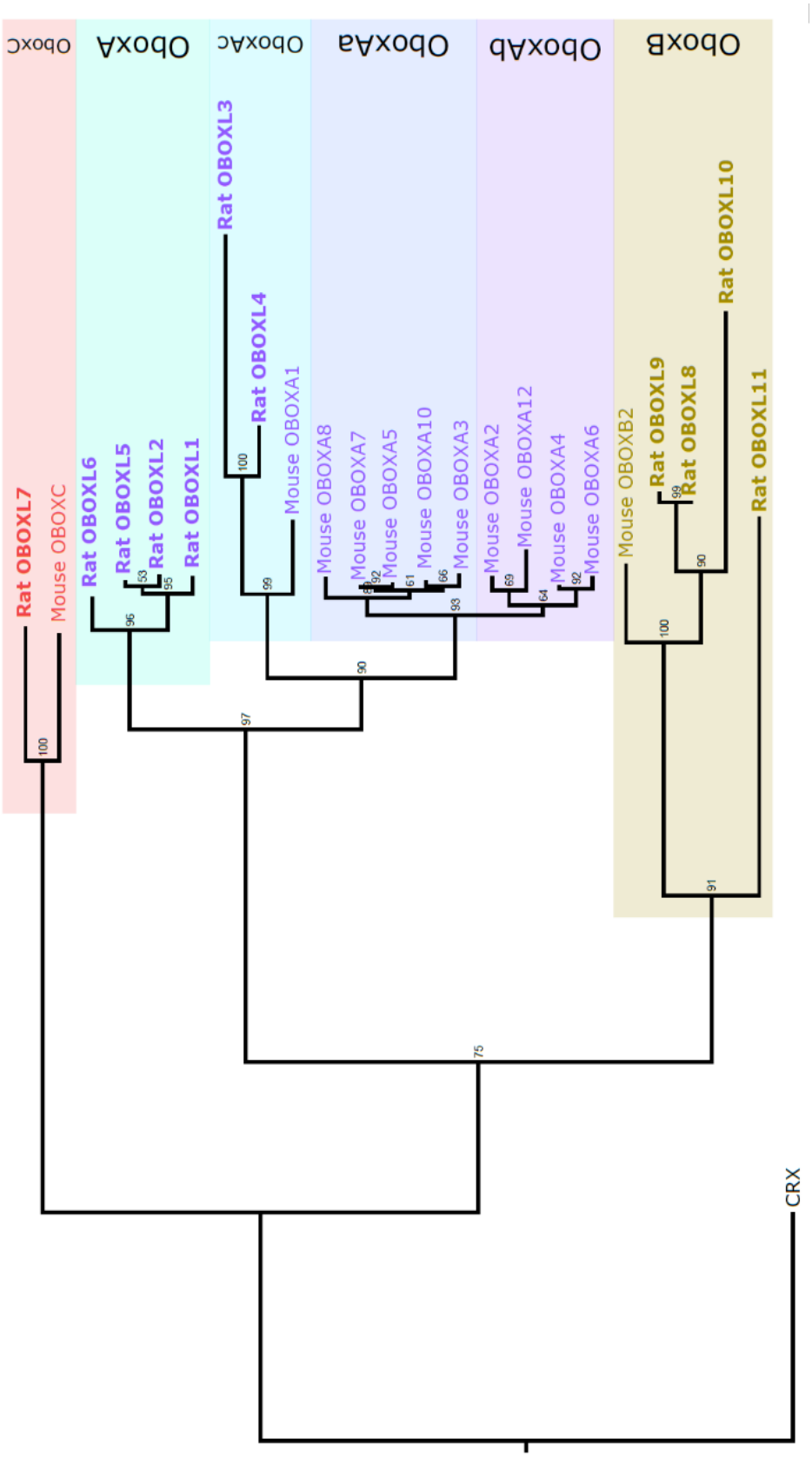


Figure 4.5 Maximum-likelihood tree of Rat and Mouse Obox amino acid sequences (no OboxD) with superimposed groupings indicated by various colours. Rat has Obox genes that group together with mouse OboxA, OboxB, and OboxC groupings, as well as two representatives of the OboxAc sub-group.

Examining the genomic location of the rat Oboxl genes reveals more about Obox evolutionary history. Specifically, the conservation of location of *Oboxc* in mouse, and *Oboxl7* in rat loci provides strong evidence for orthology (supplementary file 4.4). Both *Oboxc* and *Oboxl7* reside next to a *Crx* pseudogene located more distantly on chromosome 7 and 1 respectively. This can also be seen for mouse *Oboxb2* locus and rat *Oboxl4*, which not only exist in one phylogenetic group (figure 4.5), but are located beside the *Crx* locus at a comparative distance (supplementary file 4.4). In summary, there are rat Oboxl representatives of all mouse Obox groups (Royall *et al.*, 2018) with the exception of OboxD (figure 4.5; supplementary file 4.3).

3.4 ETCHbox distribution in Rodentia.

Following the finding that the gene duplications and ETCHbox organisation described in mouse are not *Mus*-specific but is also comparable with rat, I aimed to characterise the wider distribution of ETCHbox genes in the rodents. At the time of writing there is no publicly available gene expression data for the early embryonic stages needed for the species of interest listed in table 4.1 other than mouse. My criteria for defining the presence of an ETCHbox gene relies on the identification the full homeodomain sequence in the respective genome, and for the identified amino acid sequence to most closely hit the ETCHbox gene in question in a BLASTP analysis. This method aimed to be repeatable and objective but does come with caveats, including that I cannot be certain sequences are not pseudogenes and that non-perfect genome assemblies may result in false negatives. However, in the absence of expression data it is difficult to address these potential limitations. The

results of this ETCHbox search are displayed in figure 4.6 (and supplementary file 4.2).

Table 4.1 Family, species, and common name of genomes used to analyse rodent ETCHbox content.

Family	Species	Common Name
Muridae	*Mus musculus*	Mouse
Muridae	*Rattus novegicus*	Norway Rat
Cricetidae	*Microtus ochorgaster*	Prairie vole
Nesomyidae	*Cricetomys gambianus*	Pouched rat
Spalacidae	*Nannospalax galili*	Upper Galilee mountains blind mole rat
Dipodoidae	*Jaculus jaculus*	Lesser Egyptian jerboa
Castoridae	*Dipodomys ordii*	Kangaroo rat
Sciuridae	*Ictidomys tridecemlineatus*	13-lined Squirrel
Caviidae	*Cavia porcellus*	Guinea pig

I find evidence that the distribution of ETCHbox genes across rodents is very variable, with guinea pig retaining at least one copy of five ETCHbox genes as previously published, whereas sciurid rodents studied only have members of two ETCHbox groups, *Leutx* and *Tprx2*. I did not identify any *Pargfx* sequences in any species. The *Dprx* gene is present in three of the eight species analysed, however its distribution across the superfamily is indicative of repeated loss. Conversely there is more than one copy of *Tprx2* in all families analysed except Nesomyidae.

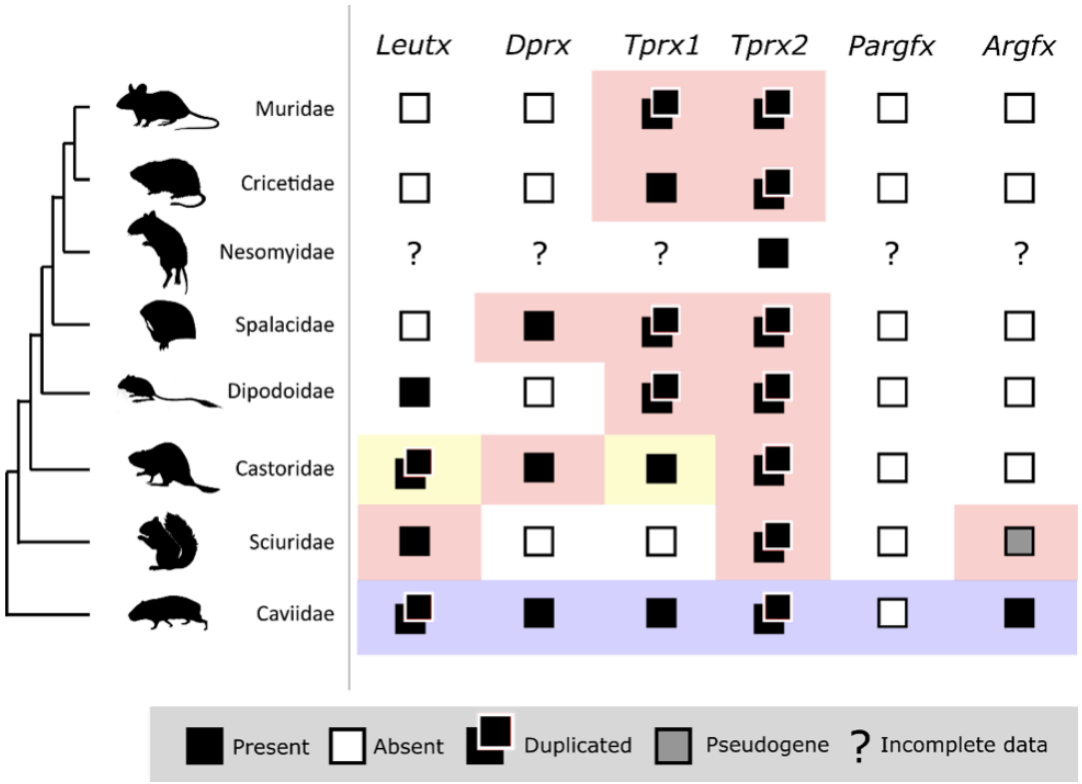


Figure 4.6 Presence of homeodomains of ETCHbox genes in the genomes of selected rodent families suggests timings of ETCHbox losses. Summary of presence and absence of ETCHbox homeodomains through the rodent order, in addition to whether the loci resides in the conserved syntenic region. For some genomes it is impossible to tell whether homeobox-containing genes lie in conserved regions or are elsewhere. Those where at least one loci is in the conserved region (ie. Not relocated in the genome compared to guinea pig) are marked by a red background. Genes where none of the sequences are in the conserved loci are represented with a yellow background, and those with assembly quality too low to assess synteny have no coloured background.

In addition to the presence of ETCHbox sequences, I also analysed whether at least one of the identified loci resides in conserved syntenic location defined as the Caviidae gene organisation (figure 4.6 - blue). This aimed to increase confidence that the identified sequences are not retrotransposed pseudogenes where identification of all exons was not possible in many cases. The vast majority of present ETCHbox genes had at least one loci in the conserved genomic region (figure 4.6 – red). In the kangaroo rat, *Leutx* and *Tprx1* sequences were identified but they are not found in the expected region (figure 4.6 - yellow). Other novelties in ETCHbox sequence and distribution that arose are listed in table 4.2.

Table 4.2 Novel findings of ETCHbox search in rodents. Other results of analysis of ETCHbox across rodents indicates a highly dynamic and changeable evolutionary history.

Family	ETCHbox novelty
Muridae, Cricetidae, and Spalacidae	*Leutx* C-teminus but no homeobox sequence can be identified in the conserved syntenic region. In rat this sequence is duplicated and amino acid sequence is 100% identical between duplicates.
Cricetidae	One *Tprx2* sequence (*Oboxld*) identified is highly divergent and to avoid interference of sequence analyses due to long branch length was removed from phylogenic trees.
Dipodoidae	The *Tprx2* loci present in the conserved genomic region has highly divergent homeodomain which has lost characteristic "WF" amino acid string in the third helix.
Castoridae	Whilst not in the conserved *Leutx* genomic region, a small cluster of 7 *Leutx* sequences are present in a different locus.
Sciuridae	No *Tprx1* loci can be identified, however the amino acid sequence of a "*Tprx2*" duplicate is grouped with *Tprx1* sequences by both BLASP and phylogenetic analyses (figure 4.7). It is possible that this is a *Tprx1* duplicate and the sequence in the conserved locus has been lost.

3.5 The arrival of the Obox-like genes - the *Tprx2* transition.

Potentially the most striking change in the evolutionary history of mouse ETCHbox genes is the "loss" of Tprx2-like amino acid sequence and appearance of Obox-like genes, or rather the transformation of the former into the latter. Current knowledge suggests that Guinea pig has *Tprx2*-like genes, whereas I have found that rat has Obox-like genes comparable to mouse (figure 4.5). Additional *Tprx* sequence analyses aimed to identify whether the Obox genes are murid rodent specific, and if not, identify where this transition to Obox-like from TPRX2-like occurred.

Taking the Tprx2 protein sequences identified above, additional analyses identified whether sequences are more similar to guinea pig and human TPRX2, or mouse and rat Obox. The combined results, along with nomenclature used herein, are displayed in Table 4.3. I find that Muridae, Cricetidae, Nesomyidae and Spalacidae all have Obox-like sequences, whereas Dipodoidae, Castoridae, and Sciuridae have Tprx2-like sequences (Table 4.3 and Figure 4.7). A combination of the BLASTP and phylogenetic tree outputs were used to allocate gene names as "*Obox*" or "*Tprx2*".

Table 4.3 The transition from *Tprx2*-like to Obox-like in rodents. Grouping of Tprx2 genes identified in this study using BLASTP against a database of known ETCHbox sequences including Tprx1 and Tprx2, and the phylogeny of the homeodomain amino acid sequences. Species with at least one sequence identified as Obox-like in both analyses are highlighted in blue. Prairie vole *Oboxld* was not included in the phylogeny as its long branch length interfered with tree resolution.

Species Common name	Family	Species	Gene name	Result using BLASTP	Result using phylogeny
Prairie vole	Cricetidae	*Microtus ochorgaster*	*Oboxla*	Obox	Obox
			Oboxlb	Obox	Obox
			Oboxlc	Obox	Obox
			Oboxld	Obox	--
Pouched rat	Nesomyidae	*Cricetomys gambianus*	*Oboxl*	Obox	Obox
Mole rat	Spalacidae	*Nannospalax galili*	*Oboxla*	Obox	Obox
			Oboxlb	Obox	Tprx1
			Oboxlc	Tprx2	Obox
			Oboxld	Obox	Obox
Jerboa	Dipodoidae	*Jaculus jaculus*	*Tprx2a*	Tprx2	Tprx2
			Tprx2b	Tprx2	Tprx2
			Tprx2c	Tprx2	Tprx2
			Tprx2d	Tprx2	Tprx2
			Tprx2e	Tprx2	Tprx2
Kangaroo rat	Castoridae	*Dipodomys ordii*	*Tprx2a*	Tprx2	Tprx2
			Tprx2b	Tprx2	Tprx1
Squirrel	Sciuridae	*Ictidomys tridecemlineatus*	*Tprx2a*	Tprx2	Tprx2
			Tprx2b	Tprx1	Tprx1
Guinea pig	Caviidae	*Cavia porcellus*	*Tprx2a*	Tprx2	Tprx2
			Tprx2b	Tprx2	Tprx2
			Tprx2c	Tprx2	Tprx2

The phylogenetic analysis, whilst not to be used to infer evolutionary history, is able to direct on more general groupings based on homeodomain sequence. I used this to identify which species (of those with Obox genes) have sequences that group with previously defined mouse and rat Obox groups and subgroups (figure 4.7). As

previously shown, rat has Obox-like sequences that group within OboxA, OboxB. and OboxC but not OboxD. This analysis shows that Cricetidae also have Obox-like sequences that group with OboxA and OboxB, with one sequence specifically grouping with OboxAc (figure 4.7), but not OboxD or OboxC. Both Spalacidae and Cricetidae also have sequences which lie as outgroups to the OboxA/B/D clade in this analysis. The Nesomyidae sequence groups with the murid OboxC-like sequences.

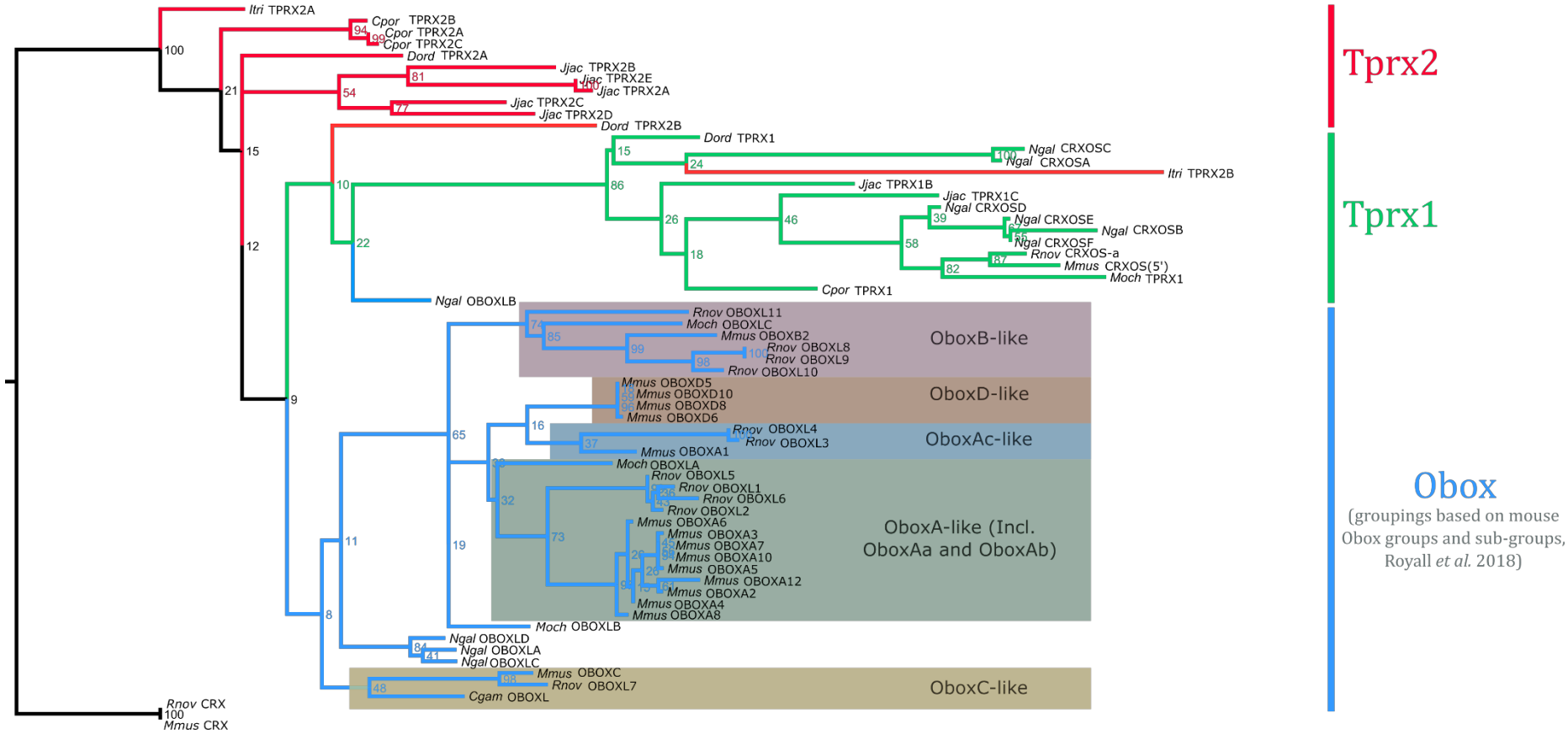


Figure 4.7 Phylogenetic tree of homeodomain amino acid sequences of rodent TPRX. This shows Tprx2 sequences of *Cricetidae*, *Nesomyidae*, and *Spalacidae*, fall in the Obox-clade (blue). *Dipodoidae*, *Castoridae*, or *Sciuridae* sequences fall in the Tprx2-like clade (red). Itri Tprx2b is placed within the Tprx1-clade (green). Species names are contractions of scientific names in table 3, with the addition of *Mmus* = *Mus musculus*, *Rnov* = *Rattus norvegicus*. **NB: The short protein sequences coupled with high levels of divergence means that this tree cannot be used to interpret evolutionary history, but instead to group sequences by similarity more generally.**.

4. Discussion

4.1 ETCHbox gene evolution in murid rodents

My analysis of the rat gene repertoire suggests that in the murid LCA, there was at least one representative of the OboxA group, the OboxAc subgroup, the OboxB group and the OboxC group. Genome analysis of specific sequences give strong evidence for orthology of mouse *Oboxc* and rat *Oboxl7*, and mouse *Oboxb2* and rat *Oboxl4*. There is also strong evidence for conserved *Crxos* gene structure and sequence between mouse and rat, contrary to the results of a much simpler earlier analysis (Royall *et al.*, 2018; chapter 2). No ETCHbox genes absent in the mouse genome were found in rat.

Taken together, this provides multiple lines of evidence that the striking ETCHbox gene organisation in mouse is not particularly unique to this one species, and that the gene loss, duplication, and sequence change had already occurred in the murid ancestor. However there are changes in both Obox and *Crxos* genes within murid rodents. One of the two homeodomains in rat is highly divergent (figure 4.4b), which is likely why BLAST analysis against mouse revealed no hits. The long branch length of the rat CRXOS sequence in phylogenetic analysis suggests that it is the rat locus that has become divergent, more so than mouse *Crxos*. This divergence is potentially due to the presence of a double-homeobox transcript as in mouse (Royall *et al.*, 2018). In this case, evolutionary constraint to keep two functional homeodomains could be lost if the divergent homeobox is not contained in a separate single homeobox transcript, allowing for sequence divergence. I also report a regional duplication of the *Crxos* and *Sepw1* loci in the rat genome (figure

4.3), although experimental analysis is needed to verify this is not an assembly error.

Changes in the Obox gene family includes the mouse-specific OboxD group, for which I find no rat representatives. This is suggestive of either OboxD genes evolving in the mouse lineage, or loss in the rat lineage. The rat genome also contains two *Oboxa1*-like genes, and just six OboxA-like genes compared to at least ten in mouse (Royall *et al.*, 2018), evidence that lineage specific duplications have occurred even in species as closely related as mouse and rat.

4.2 Small scale ETCHbox gene duplication and loss is common in Rodentia

Following the finding that the described gene losses and duplications are not mouse-specific, I widened the search for ETCHbox genes to members of each family within Rodentia. One of the most striking findings from this analysis is that small scale ETCHbox gene duplications and gene losses are common throughout the rodents (figure 4.6), and that I have found a novelty for almost every lineage examined (table 4.2). I have used this analysis to hypothesise when gene losses occurred in evolutionary history (figure 4.8). Not only are there branch-specific losses, such as independent loss of *Dprx* in Sciuridae and Dipodoidae, the ETCHbox gene loss seen in murids appears to have occurred in a stepwise fashion with *Leutx* being lost elsewhere in Myomorpha evolution, and *Argfx* being lost in the LCA of Castors (Castoridae) and Murids (Muridae; figure 4.8). Additionally, this analysis is perhaps an under representation of the true rate of ETCHbox gene loss. The use of a single representative of each family in this study risks underestimating the variation in ETCHbox gene repertoire within each family by assuming each event is

family-specific rather than species-specific. Comparisons of additional species could help decipher if a gene loss or duplication is family or species specific. It is also hard to definitively know whether a gene has been lost in any given species due to the quality of the genome assembly. To tackle this issue I have aimed to use the best quality genomes available for each family. I emphasise that the deduced timings of gene losses and duplications in evolutionary history (figure 4.8) are hypothesised estimates based on the data gathered here.

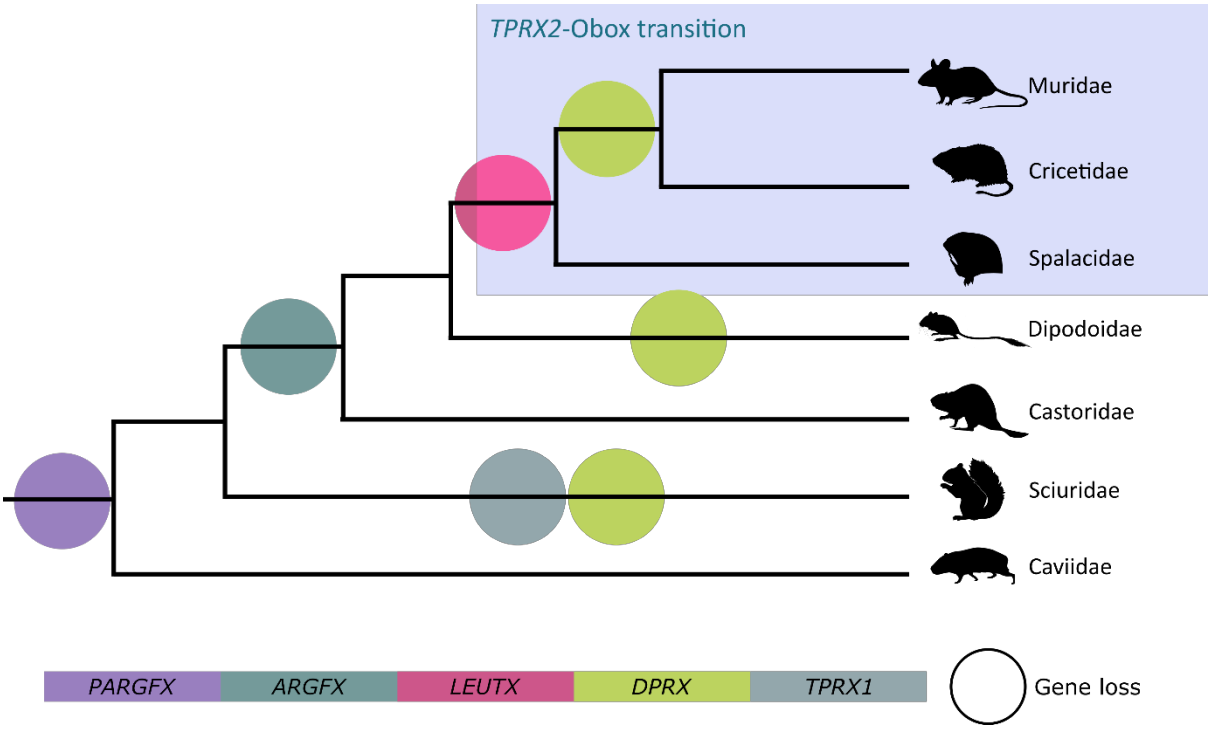


Figure 4.8 Summary of ETCHbox gene loss in *Rodentia* suggests that the ETCHbox repertoire in mouse is not due to exceptional divergence. I find stepwise loss of *Argfx*, *Leutx*, and *Dprx* in murid evolution but these losses are not murid-specific. Additionally, we find further gene loss in other lineages including Sciuridae and Dipodoidae history. The Nesomyidae have been removed from this cladogram as so few sequences were identified inclusion adds no resolution to these hypotheses.

4.3 *Tprx2* became Obox-like during Myomorpha evolution.

Another aim of this study was to identify the time point where the *Tprx2* gene sequences became Obox-like, and whether the extreme duplications seen in the mouse Obox cluster coincide with this transition. This is important as Obox genes have much more varied expression patterns than has thus far been identified for *Tprx2* genes, presumably reflecting different functions. Specifically, both my research and that of others has described four Obox expression patterns, all contained within the oocyte and preimplantation embryo (Rajkovic *et al.*, 2002; Cheng *et al.*, 2007; Lee *et al.*, 2016; Ge, 2017; Royall *et al.*, 2018). Conversely, all *Tprx2* expression described has been in a sharp peak at the eight cell and morula stages (Madissoon *et al.*, 2016; Maeso *et al.*, 2016).

My analysis of additional rodent genomes shows that *Tprx2* loci are Obox-like in amino acid sequence in Spalacidae, Cricetidae, and Muridae, but *Tprx2*-like in all other genomes examined (table 4.3). I suggest that *Tprx2* became Obox-like in Myomorpha evolution, specifically that Obox genes were present in the LCA of Spalacidae, Cricetidae, and Muridae (figure 4.8). Interestingly, in all rodent species examined other than mouse, the presence of Obox-like genes does not seem to be coupled with extreme duplications and the level of gene duplication in "Obox-like" genes is comparable to that of *Tprx2*-like genes (figure 4.6, table 4.3). This is evidence for the high levels of gene duplication described in mouse as being the only element in this story which is mouse-specific.

4.4 Implications

My results have separated the extreme divergence described in mouse ETCHbox genes into four distinct events; ETCHbox gene losses occurring stepwise through rodent evolution, *Tprx2* becoming Obox-like in sequence in the Myomorpha (figure 4.8), most Obox groups being present in the murid LCA (figure 4.5), and extreme Obox duplication occurring in mouse lineage specifically (figure 4.6, table 4.3). Whilst I have focussed on one eutherian order, this is strong evidence that ETCHbox genes are even more dynamic than previously thought with gene duplications and loss distributed throughout many lineages, rather than specifically in the mouse. To verify this, a comparable study should be carried out in other eutherian families, such as primates.

With collaborators, I have demonstrated that ETCHbox genes may have had a role in the evolution of eutherian novelties in the embryo, in the development of traits such as the inner cell mass and compact morula (chapter 3, Royall *et al.*, 2019). It is possible that there is a biological link between a species' ETCHbox repertoire and their preimplantation development, such as the extent of placental invasion or litter size. Detailed descriptions of the preimplantation development of additional rodent species are needed in order to develop this hypothesis further.

Whilst this study has revealed a great deal of information about ETCHbox evolutionary history, there are still questions relating to the evolutionary mechanisms underlying this variability and rapid change. I have previously suggested that partial functional overlap of ETCHbox genes can relieve evolutionary constraint leaving gene sequences free to evolve, as long as partial overlap is restored (chapter 2, Royall *et al.*, 2018). In this case the variability of ETCHbox genes

is not itself adaptive, but a consequence of selection for the robustness conferred by partially overlapping function. Another possible mechanism is that the variability of ETCHbox genes is itself adaptive. It has been suggested by Katayama *et al.* (2018) that ETCHbox genes could act like "speciation genes". In this case the fast evolving genes reduce the viability of embryos of hybrids between closely related species and hence can contribute to speciation. Whatever the mechanism, I have here shown that the highly variable nature of these genes is seen within Rodentia and so is tolerated (or selected for) in eutherian evolution, despite their potent transcriptomic effects and expression at a vital period of development (Maeso *et al.*, 2016; Royall *et al.*, 2018).

5. References

Altschul, S. F., Gish, W., Pennsylvania, T., Park, U., Miller, W., Myers, E. W. and Lipman, D. J. (1990) 'Basic Local Alignment Search Tool', *Journal of molecular biology*, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Blanga-Kanfi, S., Miranda, H., Penn, O., Pupko, T., Debry, R. W. and Huchon, D. (2009) 'Rodent phylogeny revised: Analysis of six nuclear genes from all major rodent clades', *BMC Evolutionary Biology*, 9(1), pp. 1–12. doi: 10.1186/1471-2148-9-71.

Bonaldo, M. F., Lennon, G. and Soares, M. B. (2007) 'Normalization and subtraction: two approaches to facilitate gene discovery.', *Genome Research*, 6(9), pp. 791–806. doi: 10.1101/gr.6.9.791.

Burge, C. and Karlin, S. (1997) 'Prediction of complete gene structures in human genomic DNA', *Journal of Molecular Biology*, 268(1), pp. 78–94. http://www.sciencedirect.com/science/article/pii/S0022283697909517.

Capella-Gutiérrez, S., Silla-Martínez, J. M. and Gabaldón, T. (2009) 'trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses', *Bioinformatics*, 25(15), pp. 1972–1973. doi: 10.1093/bioinformatics/btp348.

Cheng, W. C., Hsiu, M. H. L., Yeh, Y. J. and Li, H. (2007) 'Mice lacking the Obox6 homeobox gene undergo normal early embryonic development and are fertile', *Developmental Dynamics*, 236(9), pp. 2636–2642. doi: 10.1002/dvdy.21261.

Fabre, P. H., Hautier, L., Dimitrov, D. and P Douzery, E. J. (2012) 'A glimpse on the pattern of rodent diversification: A phylogenetic approach', *BMC Evolutionary Biology*, 12(1). doi: 10.1186/1471-2148-12-88.

Frankenberg, S. R., de Barros, F. R. O., Rossant, J. and Renfree, M. B. (2016) 'The mammalian blastocyst', *Wiley Interdisciplinary Reviews: Developmental Biology*, 5(2), pp. 210–232. doi: 10.1002/wdev.220.

Ge, S. X. (2017) 'Exploratory bioinformatics investigation reveals importance of " junk " DNA in early embryo development'. BMC Genomics, pp. 1–19. doi: 10.1186/s12864-017-3566-0.

Graur, D., Hide, W. and Li, W.-H. (1991) 'Is the Guinea-Pig a Rodent?', *Nature*, 351, pp. 649–652. doi: doi.org/10.1038/351649a0.

Holland, P. W. H., Booth, H. A. F. and Bruford, E. A. (2007) 'Classification and nomenclature of all human homeobox genes', *BMC Biology*, 5, pp. 1–28. doi: 10.1186/1741-7007-5-47.

James Kent, W. (2002) 'BLAT - The BLAST-like alignment tool', *Genome Research*, 12(4), pp. 656–664. doi: 10.1101/gr.229202. Article published online before March 2002.

James Kent, W., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002) 'The human genome browser at UCSC', *Genome Research*, 12(6), pp. 996–1006. doi: 10.1101/gr.229102. Article published online before print in May 2002.

Jouhilahti, E.-M., Madissoon, E., Vesterlund, L., Töhönen, V., Krjutškov, K., Plaza Reyes, A., Petropoulos, S., Månsson, R., Linnarsson, S., Bürglin, T., Lanner, F., Hovatta, O., Katayama, S. and Kere, J. (2016) 'The human PRD-like homeobox gene *LEUTX* has a central role in embryo genome activation', *Development*, 143(19), pp. 3459–3469. doi: 10.1242/dev.134510.

Katayama, S., Ranga, V., Jouhilahti, E.-M., Airenne, T. T., Johnson, M. S., Mukherjee, K., Bürglin, T. R. and Kere, J. (2018) 'Phylogenetic and mutational analyses of human LEUTX, a homeobox gene implicated in embryogenesis', *Scientific Reports*, 8(1), p. 17421. doi: 10.1038/s41598-018-35547-5.

Katoh, K. and Standley, D. M. (2013) 'MAFFT multiple sequence alignment software version 7: Improvements in performance and usability', *Molecular Biology and Evolution*, 30(4), pp. 772–780. doi: 10.1093/molbev/mst010.

Kuijk, E., Geijsen, N. and Cuppen, E. (2015) 'Pluripotency in the light of the developmental hourglass', *Biological Reviews*, 90(2), pp. 428–443. doi: 10.1111/brv.12117.

Lee, H. S., Kim, K. H., Kim, E. Y., Lee, S. Y., Ko, J. J. and Lee, K. A. (2016) 'Obox4-silencing-activated STAT3 and MPF/MAPK signaling accelerate nuclear membrane breakdown in mouse oocytes', *Reproduction*, 151(4), pp. 369–378. doi: 10.1530/REP-15-0020.

MacLean, J. A., Chen, M. A., Wayne, C. M., Bruce, S. R., Rao, M., Meistrich, M. L., Macleod, C. and Wilkinson, M. F. (2005) 'Rhox: A new homeobox gene cluster', *Cell*, 120(3), pp. 369–382. doi: 10.1016/j.cell.2004.12.022.

Madissoon, E., Jouhilahti, E. M., Vesterlund, L., Töhönen, V., Krjutškov, K., Petropoulous, S., Einarsdottir, E., Linnarsson, S., Lanner, F., Månsson, R., Hovatta, O., Bürglin, T. R., Katayama, S. and Kere, J. (2016) 'Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos', *Scientific Reports*. Nature Publishing Group, 6(July), pp. 1–14. doi: 10.1038/srep28995.

Maeso, I., Dunwell, T. L., Wyatt, C. D. R., Marlétaz, F., Veto, B., Bernal, J. A., Quah, S., Irimia, M. and Holland, P. W. H. (2016) 'Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals', *BMC Biology*, 14(1), pp. 1–14. doi: 10.1186/s12915-016-0267-0.

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Marchler, G. H., Song,

J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C. and Bryant, S. H. (2015) 'CDD: NCBI's conserved domain database', *Nucleic Acids Research*, 43(D1), pp. D222–D226. doi: 10.1093/nar/gku1221.

Nei, M., Xu, P. and Glazko, G. (2001) 'Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms', *Proceedings of the National Academy of Sciences*, 98(5), pp. 2497–2502. doi: 10.1073/pnas.051611498.

Rajkovic, A., Yan, C., Yan, W., Klysik, M. and Matzuk, M. M. (2002) 'Obox, a family of homeobox genes preferentially expressed in germ cells', *Genomics*, 79(5), pp. 711–717. doi: 10.1006/geno.2002.6759.

Robinson, J. T. (2012) 'Integrated genomics viewer', *Nature biotechnology*, 29(1), pp. 24–26. doi: 10.1038/nbt.1754.Integrative.

Royall, A. H., Frankenberg, S., Pask, A. J. and Holland, P. W. H. (2019) 'Of eyes and embryos : subfunctionalization of the CRX homeobox gene in mammalian evolution', *Proceedings of the Royal Society B: Biological Sciences*.

Royall, A. H., Maeso, I., Dunwell, T. L. and Holland, P. W. H. (2018) 'Mouse Obox and Crxos modulate preimplantation transcriptional profiles revealing similarity between paralogous mouse and human homeobox genes', *EvoDevo*. BioMed Central, 9(1), p. 2. doi: 10.1186/s13227-018-0091-4.

Stamatakis, A. (2014) 'RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies', 30(9), pp. 1312–1313. doi: 10.1093/bioinformatics/btu033.

Töhönen, V., Katayama, S., Vesterlund, L., Jouhilahti, E. M., Sheikhi, M., Madissoon, E., Filippini-Cattaneo, G., Jaconi, M., Johnsson, A., Bürglin, T. R., Linnarsson, S., Hovatta, O. and Kere, J. (2015) 'Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development', *Nature Communications*, 6, pp. 1–9. doi: 10.1038/ncomms9207.

Wilming, L. G., Boychenko, V. and Harrow, J. L. (2015) 'Comprehensive comparative homeobox gene annotation in human and mouse', *Database*, 2015(2), pp. 1–12. doi: 10.1093/database/bav091.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Cunningham, F., Yates, A.

and Flicek, P. (2018) ‘Ensembl 2018’, *Nucleic Acids Research*, 46(D1), pp. D754–D761. doi: 10.1093/nar/gkx1098.

Zhong, Y. F. and Holland, P. W. H. (2011) ‘The dynamics of vertebrate homeobox gene evolution: Gain and loss of genes in mouse and human lineages’, *BMC Evolutionary Biology*, 11(1). doi: 10.1186/1471-2148-11-169.

6. Supplementary Files

4.1 Information on Genome Assemblies

4.2 Rodent ETCHbox sequence loci and information, **Bold** text indicates scaffold or chromosome is the same as *Crx* in that species, ie. ETCHbox gene on conserved chromosome.

4.3 Phylogenetic analysis including mouse OboxD (noted as OboxAB).

4.4 Linked Obox loci in mouse and rat is good evidence for orthology.

4.1 Information on Genome Assemblies

Species	Common name	Family	Scaffold N50	Coverage	Seq. Platform	Assembly ID	Year
Mus musculus	Mouse	*Muridae*	52,589,046	mixed	mixed	GCA_000001635.8	2012, 2017
Rattus norvegicus	Rat	*Muridae*	14.99Mb	mixed	Sanger, SOLiD, PacBio	GCA_000001895.4	2014
Microtus ochorgaster	Prairie vole	*Cricetidae*	17,270,019	94x	HiSeq	GCA_000317375.1	2012
Cricetomys gambianus	Pouched rat	*Nesomyidae*	110,049	36.7x	HiSeq	GCA_004027575.1	2019
Nannospalaxm galili	Mole rat	*Spalacidae*	3,618,479	86x	HiSeq	GCA_000622305.1	2014
Jaculus jaculus	Egyptian jerboa	*Dipodoidea*	22,080,993	78x	HiSeq	GCA_000280705.1	2012
Dipodomys ordii	Kangaroo rat	*Castoridae*	11,931,245	181x	Illumina Sanger	GCA_000151885.2	2014
Marmota marmota	Alpine Marmot	*Sciuridae*	31,340,621	30x	?	GCA_001458135.1	2015
Ictidomys tridecemlineatus	13 lined squirrel	*Sciuridae*	8,192,786	495x	HiSeq	GCA_000236235.1	2011
Cavia Porcellus	Guinea pig	*Cavidae*	27,942,054	6.8x	Sanger	GCA_000151735.1	2008

4.2 Rodent ETCHbox sequence loci and information, Bold text indicates scaffold or chromosome is the same as *Crx* in that species, ie. ETCHbox gene on conserved chromosome.

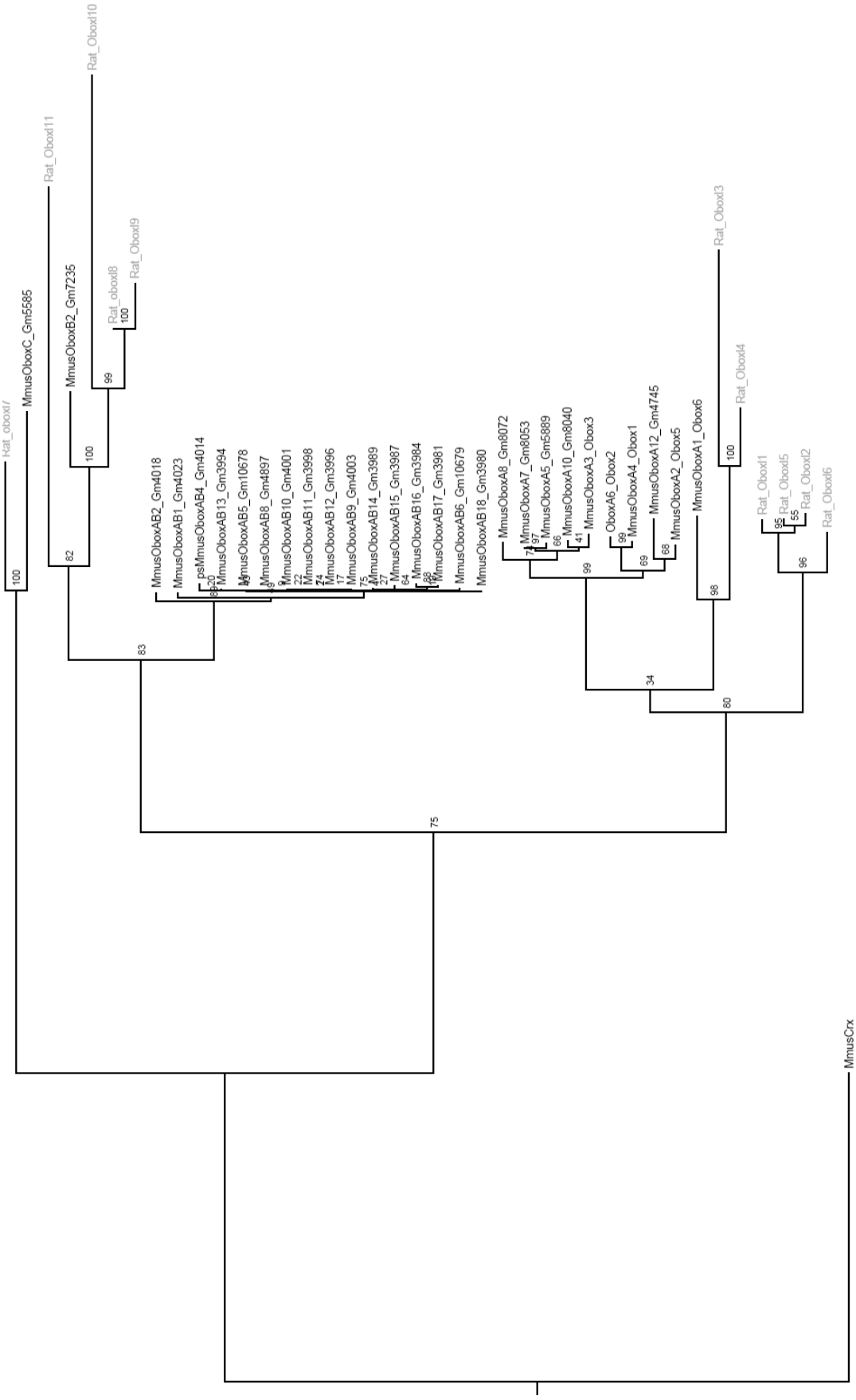
Species	Gene	Present?	Duplicated	HD?	Ch	Start	Stop	name	Predicted ID
Rat	*ARGFX*	n	--		--	--	--	--	--
	LEUTX	y	2	No	**1**	84785968	84786759	--	Chr1.2361
					1	85531709	85532500	--	Chr1.2409
	DPRX	n	--		--	--	--	--	--
	PARGFX	n	--		--	--	--	--	--
	TPRX2/OBOX	y	Y	11	**1**	71766007	71767465	Rat_Oboxl1	XP_008771600.1
					1	71785750	71787208	Rat_Oboxl2	XP_008771600.1
					1	77615145	77617701	Rat_Oboxl3	XP_003753303.2
					1	77621919	77623102	Rat_Oboxl4	XP_006228422.1
					X	93064620	93065639	Rat_Oboxl5	XP_008771600.1
					9	1.01E+08	1.01E+08	Rat_Oboxl6	XP_006223134.1
					1	75927706	75928821	Rat_Oboxl7	XP_008757318.1
					1	77702767	77709433	Rat_Oboxl8	GenScan00000011631.1
					1	77599286	77603937	Rat_Oboxl9	GenScan00000011629.1
					1	77476432	77482117	Rat_Oboxl10	--
					1	1.03E+08	1.03E+08	Rat_Oboxl11	--
	TPRX1	y	Y		**1**	77784013	77789400	--	--
		y			**1**	77510340	77516619	--	--
Prairie vole	*ARGFX*	n	--	--	--	--	--	--	--
	TPRX1	y	n	Y	**LG4**	18598520	18602978	Crxos	LOC10198746, xm_005361074.2
	TPRX2_OBOX	y	4	y	JH996434	11541184	11541645	Oboxla	
				y	**lg4**	18669221	18669727	Oboxlb	end of Genscan0000008207
				y	lg1	40980816	40981163	Oboxlc	
				Y	JH996705.1	30210	33676	Oboxld	GENSCAN00000012294.1
				y w stop	lg9	26022667	26023305	psOboxl4	
				n	LG9	26022586	26023206	Oboxl4partial	
				n	**LG4**	18670138	18670563	Oboxl6partial	Genscan0000008207
	LEUTX	y	n	N	1	50309935	50311782	Just c-term	Genscan00000029540.1
	DPRX	n	--	--	--	--	--	--	--
	PARGFX	n	--		--	--	--	--	--
Pouched rat	*ARGFX*	?							--
	TPRX1	?							--
	?RX2_OBO	y			scaffold_28809	1527	1946	Oboxl	--
	LEUTX	?							--
	DPRX	?							--
	PARGFX	?							--
Blind mole rat	*ARGFX*	n	--		--	--	--	--	--
	TPRX1	y	6 w HD, 8 total	Y	**KL199454**	16290	19180	Crxos_a	GenScan1903 EXONS 1,2,3
				Y	**KL199454**	3626	11650	Crxos_b_partial	Genscan1903 exons 4, 5
				Y	KL201849	1494732	1495382	Crxos_c	--
				N	axcs01350981.1			Crxos_b_ps	--
				n	KL203044	649687	650302	Crxos_a_ps	--
				Y	KL201001	9435977	9436339	Crxos_d	GenScan16040
				Y	KL202943	1036425	1036769	Crxos_e	--
				Y	KL203188	1545691	1546023	Crxos_f	--
	TPRX2_OBOX	y	4	y	KL204779	8488808	8489272	Oboxla	--
				y	**KL199454**	74308	74481	Oboxlb	GenScan1899
				y	**KL199454**	111104	111283	Oboxlc	GenScan1900
				y	**KL199454**	87984	88130	Oboxld	GenScan1901
	LEUTX	y	2	n	KL204554	2474865	2475113	Leutx	--
				n	KL205461	631478	631798	Leutxps	--
	DPRX	Y	no	Y	KL199583	296570	307026	dprx	Genscan00000013501.1
	PARGFX	n	--	--	--	--	--	--	--

Egyptian Jerboa	*ARGFX*	n	--	--	--	--	--	--	--
	TPRX1	y	2	y	JH725461	14281139	14281747	Tprx1b	Genscan00000021716
									--
				y	**JH725653**	198285	209540	Tprx1c	Genscan00000019440
	TPRX2_OBOX	y	5	y	JH725464	2519712	25197212	Tprx2b	--
				y	jh725475	7654524	7654700	Tprx2c	--
				y	**JH725653**	144564	145430	Tprx2d	--
				y	JH725523	517393	517716	Tprx2a	**Genscan00000039315**
				y	JH725473	16915549	16915842	Tprx2e	Part of Genscan00000021275
	LEUTX	y		y	JH726981	3303	3737	Leutx	--
	DPRX	n	--		--	--	--	--	--
	PARGFX	n	--		--	--	--	--	--
Kangaroo rat	*ARGFX*	n	--		--	--	--	--	--
	TPRX1	y	n	y	KN672561	1124566	1137260	KanRat_Tprxb	--
	TPRX2_OBOX	Y	y	y	KN672391	8685813	8688467	Tprx2b	Genscan00000101552
				Y	**KN672502**	7856	10100	Tprx2a	Genscan00000100392
	LEUTX	Y	7	Y	KN672414	2019234	2019803	Leutxa	Genscan00000103825 (part of)
				y	KN672414	2030896	2031399	Leutxb	--
				y	KN672414	2028834	2029331	Leutxc	--
				y	KN672414	2041665	2042162	Leutxd	--
				y	KN672414	2044158	2044640	Leutxe	Starts at HD start
				Y	KN672356	22567901	22568539	Leutxf	Genscan00000091903
				y	KN672414	1997678	2004562	Leutxg	Genscan00000103825 (part of)
	DPRX	y		y	KN672451	3282134	3284773	KanRat_Dprx	Genscan00000104837.1
	PARGFX	n	--		--	--	--	--	--
13-lined squirrel	*ARGFX*	ps	--	n	JH393346	8025833	8040420	prem stops	XM_005342965.1
	TPRX1	n	--	--	--	--	--	--	--
	TPRX2_OBOX	y	2	y	**JH393474**	1427335	1429266	Tprx2a	XM_013363464.1
				y	**JH393474**	1520467	1528053	Tprx2b	GENSCAN00000019212
	LEUTX	y	n	y	JH393471	2403970	2406044	Squ_Leutx	Part of Genscan000031224
	DPRX	y			JH393638	560676	650834	Squ_Dprx_partial	--
	PARGFX	n	--	--	--	--	--	--	--
Guinea pig	*ARGFX*	y		y	DS562890	22887521	22904021	Gpig_Argfx	Part of Genscan0000041105
	TPRX1	y	n	y	**DS562935**	7508234	7511237	Gpig_Tprx1	Genscan00000009053
	TPRX2_OBOX	y	3 w HD	y	**DS562935**	7545739	7548278	Tprx2a	Genscan0000009044
				y	**DS562935**	7537426	7539307	Tprx2b	Genscan0000009042
				y	DS563064	818288	853638	Tprx2c	Genscan0000014214
				n	DS563064	913094	913579	Gpig_Tprxd_ps	Part of Genscan0000014216
	LEUTX	y	maeso et al. 2016						
	DPRX	y	n	y	DS562925	3110777	3109845	Gpig_Dprx	Part of Genscan0000004708
	PARGFX	n	--	--	--	--	--	--	--

Key:

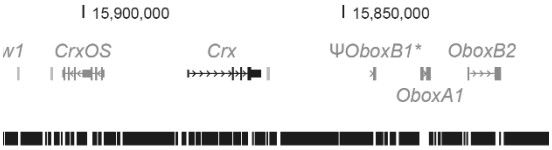
Y = yes, n = no, ps = pseudogene, -- = not applicable, HD?= Homeodomain present?

4.3 Phylogenetic analysis including mouse OboxD (noted as OboxAB).

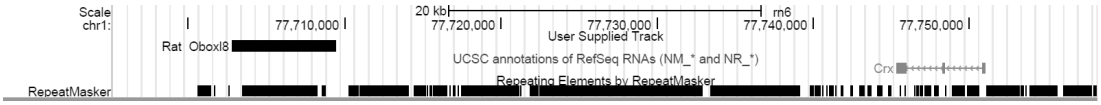


4.4 Linked Obox loci in mouse and rat is good evidence for orthology.

(a) Mouse *Oboxb1*…



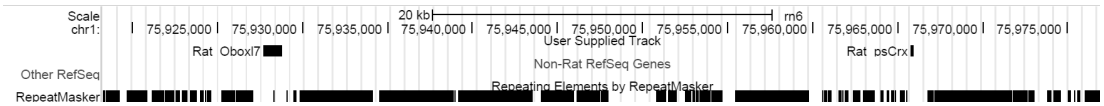
… and rat *Oboxl8* adjacent to *Crx*.



(b) Mouse *Oboxc* …



…and rat *Oboxl7* adjacent to *psCrx.*



Chapter 5: ETCHbox Genes and Pluripotency.

Author Contributions

The somatic cell reprogramming and counting of resulting colonies within this chapter was carried out by Timothy Davies, MSc (Sir William Dunn School of Pathology, University of Oxford). The methodology and approach taken in this chapter was developed by Amy Royall and Peter Holland with input and advice from Timothy Davies and Paul Fairchild (Sir William Dunn School of Pathology, University of Oxford). Amy Royall conducted bioinformatic analyses of public data, maintained cell cultures and assessed ETCHbox expression in stem cell colonies, produced differentiated cells exposed to ETCHbox genes prior to reprogramming, assessed marker gene expression in resulting colonies, and conducted statistical analysis. The text within this chapter was written with input from Tim Davies on relevant methodologies.

Abstract

The evolutionary timing of the arrival of ETCHbox genes corresponds to the arrival of other novelties arriving on the lineage to eutherian mammals, including a morphologically distinct pluripotent cell population; the Inner Cell Mass (ICM). In this chapter I probe any relationship between these co-occurring eutherian novelties and reveal a potential interplay of ETCHbox genes and the derivation of the pluripotent state.

Firstly, I describe the expression of ETCHbox genes in mouse pluripotent stem cells, both embryo-derived (ESC) and experimentally induced (iPSC). Secondly, I find evidence for direct regulation of *Crxos* by the Yamanaka factors; factors used to induce pluripotency in differentiated cells. Using somatic cell reprogramming after exposure cells to ETCHbox genes I demonstrate the ability of ETCHbox genes to increase reprogramming efficiency. I then show that following *Crxos* exposure before reprogramming, the resulting stem cell colonies differentially express markers of increased differentiation potential compared to control stem cells reprogrammed using current commercial methods. Finally, I find additional evidence for the involvement of "distributed robustness" in the evolution of ETCHbox genes and find additional support for opposing effects of human ETCHbox genes (*ARGFX* and *TPRX1*) on the progression to the first cell lineage decision.

I conclude that following the adoption of ancestral embryonic function by the ETCHbox genes, they have developed it (DDC+) and gained additional roles in regulatory networks which confer proper timing of pluripotency from totipotency. This transition is associated with the progression of the embryo towards eutherian-specific traits such as the inner cell mass.

1. Introduction

The eutherian mammal-specific ETCHbox genes are variable in sequence and copy number across Eutheria. Current knowledge suggests that all are united in their expression during the earliest stages of embryonic development as cells are making the first cell lineage decisions (figure 5.1) (Maeso *et al.*, 2016; Royall *et al.*, 2018). This divergence in repertoire and sequence, but constraint in temporal expression, suggests an involvement of ETCHbox in a conserved gene-regulatory network vital for embryo survival but with variable elements. The coding region of genes is able to evolve alongside its targets, however there is reduced sequence change in regulatory elements, retaining the overarching expression pattern.

The progression of eutherian mammal preimplantation development has many features common across distantly related species. A hollowed blastocyst structure with a cavity, physical separation of the two cell populations which go on to form the embryo proper and extraembryonic lineages, and the invasion of the blastocyst into maternal tissue are some of these features and are well described processes regulated by characterised genetic pathways in mouse (Frankenberg *et al.*, 2016; Rossant, 2016; Griffith *et al.*, 2017). However, with the practical and ethical difficulty of studying the equivalent stages of human development, only now are researchers beginning to unravel the precise genetic mechanisms that control our own early development (Fogarty *et al.*, 2017; Li *et al.*, 2018). Differences between the gene regulatory networks (GRNs) conducting development of mouse and human have already been described (Niakan and Eggan, 2013; Xue *et al.*, 2013; Boroviak *et al.*, 2018; Wu *et al.*, 2018), but mouse remains a commonly used model

system for studying these earliest stages (Ménézo and Hérubel, 2002). This renders it interesting and relevant to uncover further genetic differences between species.

While it is difficult to study these earliest stages of development *in vivo*, it is possible to study them indirectly through manipulation of differentiated cells to produce cells with increased differential potential, akin to those of the embryo. This process is referred to as "somatic cell reprogramming" and produces cells which are morphologically and phenotypically similar to embryonic stem cells, named induced pluripotent stem cells (iPSCs) (Takahashi and Yamanaka, 2006; Polo *et al.*, 2010; Park *et al.*, 2017). These iPSCs are produced using specific transcription factors known as the 'Yamanaka factors' or OKSM factors (*Oct4*, *Klf4*, *Sox2,* C-*myc*) (Takahashi and Yamanaka, 2006) which 'reprogram' the transcriptome and chromatin state of differentiated cells.

The differentiation potential of iPSCs, which aim to recapitulate the potential of embryonic stem cells (ESCs) in the early embryo, derived from mouse and human adult cells exemplifies one of the well-described differences between human and mouse development (Tesar *et al.*, 2007; Chin *et al.*, 2009; Kuijk, Geijsen and Cuppen, 2015; Theunissen *et al.*, 2016; Kojima *et al.*, 2017; Kilens *et al.*, 2018). Through somatic cell reprogramming, it is possible to manipulate adult mouse cells to form stem cell colonies which show properties similar to those seen in the mouse inner cell mass population, with the potential to differentiate into all three germ layers of the embryo proper. This state is known as naïve pluripotency and is recapitulated by cultured embryonic stem cells derived directly from the inner cell mass of mouse embryos (Karagiannis and Takashima, 2017; Du *et al.*, 2018; Hassani *et al.*, 2018). Using the same factors and inducing genetic networks in human somatic cells, it is

also possible to gain pluripotent stem cells, but these populations are more similar to the ICM derived epiblast of later stage embryos with diminished developmental potential (EpiSCs)(Weinberger *et al.*, 2016). This state is known as 'primed' pluripotency and is also observed for human ESC populations (Tesar *et al.*, 2007; Weinberger *et al.*, 2016). Although naïve human stem cells have been derived from preimplantation embryos the fundamental difference between naïve and primed pluripotency in human cells is currently a large area of study (Takashima *et al.*, 2015; Guo *et al.*, 2016; Theunissen *et al.*, 2016; Pastor *et al.*, 2018; Giulitti *et al.*, 2019; Messmer *et al.*, 2019; Rostovskaya, Stirparo and Smith, 2019).

In addition to naïve and primed pluripotency, there is a third state which is of interest to developmental biologists studying the earliest stages of development; totipotency (Baumann, 2017; Choi *et al.*, 2017; Ladstätter and Tachibana, 2018). A totipotent cell could be compared to a zygotic or two-cell stage mouse embryo (figure 5.1), with the potential to differentiate not only into all cell types of the embryo proper, but also those of extraembryonic lineages (Suwińska *et al.*, 2008; Casser *et al.*, 2017; Baker and Pera, 2018). Although the specific definition of totipotency varies, some strict definitions only recognise the zygote as truly "totipotent" (Condic, 2013; Boiani *et al.*, 2019), whereas others define cleavage stage embryos as such (Ziomek, Johnson and Handyside, 1982). For the purpose of this chapter I will refer to markers of pre-morula embryo as "totipotency markers" for convenience but acknowledge that this may not adhere with the strictest definitions of totipotency.

The derivation of these cell types, and manipulation to form them *in vitro*, is of interest medically with the aim of reducing the risk of immunological rejection

following organ transplantation, but also for understanding how embryos remove themselves from this state and begin the journey towards differentiation. Additionally, there is a medical interest in replacing *C-Myc* in the genetic cocktail administered to adult cells due to its oncogenic properties (Wernig *et al.*, 2008). For these reasons, there is a wide interest in finding suitable candidates which may either reinstate pluripotency more efficiently or create a more potent state such as totipotency.

My co-authors and I have previously identified that genes differentially regulated downstream of ETCHbox genes are enriched for functions in the preimplantation embryo at both pluripotent and totipotent stages (Maeso *et al.*, 2016; Royall *et al.*, 2018, chapter 2). Some of the Yamanaka factors themselves are regulated downstream of ETCHbox in our datasets (Royall *et al.*, 2018). If ETCHbox genes exist in a hierarchy of genes responsible for totipotency in the early embryo, it is possible that they are involved in global genome organisational changes which are recapitulated *in vitro* through cellular reprogramming.

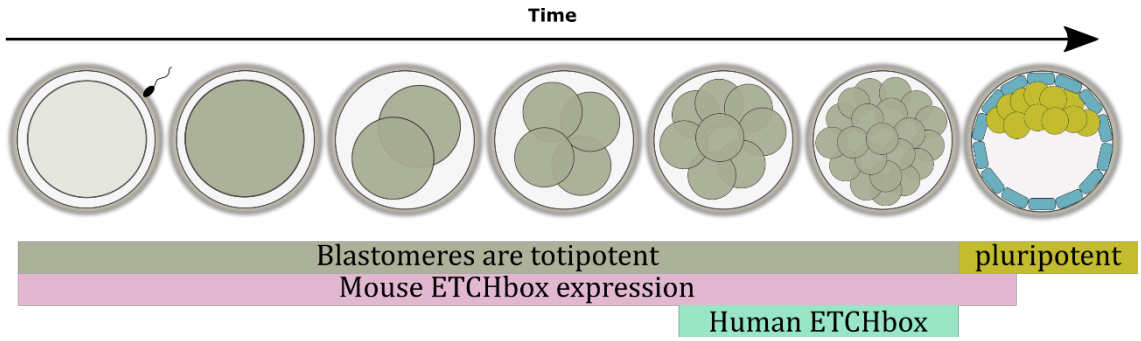


Figure 5.1 Development of the preimplantation embryo. The progression of eutherian preimplantation development coincides with the loss of totipotency (grey), the initiation of pluripotency (yellow), and expression of both human (green) and mouse (pink) ETCHbox genes. Throughout this period of development, gene transcription and translation begin, and the first cell lineage decision is made.

Co-authors and myself have also presented evidence that after the duplication of the *CRX* gene in a eutherian ancestor to give rise to the ETCHbox genes, they have not only adopted an embryonic function once held by *CRX*, but elaborated it (Royall *et al.*, 2019, chapter 3), potentially allowing for eutherian novelties such as an internal ICM to arise. Recent studies of reprogramming of adult cells to the pluripotent state have found a role for mouse ETCHbox genes in improving the efficiency of successful derivation of naïve pluripotent stem cells (Schiebinger *et al.*, 2017; Wu *et al.*, 2017). Additionally, it has been presented that another *Crxos* transcript (aka EGAM1N, not tested here) can partially suppress differentiation (Sato *et al.*, 2015). It has been reported that human ETCHbox genes are not expressed in human ESCs (Maeso *et al.*, 2016) which display primed pluripotency, but it is unknown if they are expressed in mouse stem cells.

Here, using knowledge of pluripotency, somatic cell reprogramming, and ETCHbox function collected by us and others, I hypothesise that the arrival of the ETCHbox genes on the same lineage as characteristic blastocyst morphology, is related to the evolution of the pluripotent inner cell mass of eutherian mammals. Within this, I explore whether mouse ETCHbox genes are expressed in stem cell populations and if so, if they are involved in potency of the stem cells. The answers to these questions have the potential to further our knowledge of the function of ETCHbox genes *in vivo*, and the genetic networks involved in inaccessible stages of early embryogenesis.

2. Materials and Methods

2.1 Bioinformatic analysis of pluripotent cells.

Gene expression was assessed in public RNA-sequencing data sets using STAR aligner to align reads to the mm10 Mus musculus genome (Dobin *et al.*, 2013). Cufflinks software was used to extract the FPKM values. For ATAC-seq and ChIP-seq datasets, MACS2 (Zhang *et al.*, 2008) was deployed using default settings and bed files were viewed in integrative genomics viewer (Robinson, 2012; Thorvaldsdóttir, Robinson and Mesirov, 2013).

2.2 Induced Pluripotent Stem Cell (iPSC) culture.

Mouse CBA x CBA iPSCs (XY, normal karyotype, form all three germ layers in teratoma assay) were grown in Knock-out DMEM (Gibco #10829018 with 1X Glutamax, 1X Non-essential Amino Acids, 1X Sodium Pyruvate, 10% FBS, 1% Penstrep) on a confluent feeder layer of CBA MEFs treated with mitomycin-C for an hour to prevent proliferation. Cells were passaged at 10% twice a week onto feeders to collect different ages for RNA-extraction. For feeder-free iPSCs, to avoid the possibility that gene expression was from feeder cells, harvested iPSCs were separated from feeders by three 30 minute incubations in new plasticware to allow feeders to attach and leave iPSCs in solution. The enriched iPSCs were then plated feeder free in plasticware coated in 0.1% gelatin:PBS.

2.3 RNA-extraction and Reverse Transcriptase (RT)-PCR.

RNA was extracted using the RNeasy Micro Plus RNA extraction kit (Qiagen) using manufacturer instructions. cDNA was synthesised using iScript Advanced cDNA synthesis for RT-PCR (BIO-RAD) to manufacturer's instructions. RT-PCR was conducted using BioMix with 1ul cDNA and 1uM primer mix (Table 5.1, Table 5.2).

Table 5.1. Primers used for RT-PCR analysis of ETCHbox expression

Target Gene (mouse)	Forward Primer	Reverse Primer	Fragment Size (bp)
Crxos (5')	CGTTCCCTGACCTCCTGTA	GGTGGCCATGATCTTTCGAT	129
Crxos (3')	TGGAAGAAATCGACGGGATG	TTTCCTCATTAGGGCAACGG	141
Oboxa1	GGAAGCACCGGAAAATACGC	ACTTAGCCCGGTGGTTCTTG	171
Oboxa4	AGAAAATTGTGGAGCTGGCA	ATATTCTGGAGGTTCATCCGC	106
Oboxa7	GACCCCAGGAAGTACAATGC	GCCATAGGGCCAGTTTGTTT	141
Oboxc	CACATGTTACCGGAAAGGCT	CCCACGAGCTTTTAGACGTT	236
Oboxb2	ATGCAAGTCACTGGCTCAAA	GTGGGCCTACAATGAGGTTT	141

Table 5.2: PCR protocol used for RT-PCR. Steps marked with * repeated for 35 cycles

Step	
1	3 minute denature at 94 degrees
2	15 second denature at 94 degrees*
3	30 second anneal at 55 degrees*
4	1 minute extension at 69 degrees*

2.4 Fibroblast Culture, electroporation, and reprogramming.

Mouse embryonic fibroblasts for reprogramming were grown to 70-80% confluency in DMEM supplemented with 10% FCS, 1% Penstrep. Cells were then harvested and electroporated with constructs containing the CDS of a human ETCHbox gene (*ARGFX, TPRX1,* or *TPRX2*), a mouse ETCHbox gene (*Oboxa1, Oboxa4,* or *Crxos*), or no ETCHbox gene control and puromycin resistance sequence used for selection. One day after electroporation the media was supplemented with 0.5ug/mL puromycin. Cells were cultured under puromycin selection for 82hours before a 24 hour recovery period. Cells were then harvested and re-plated for reprogramming.

The following steps were conducted blindly, with no knowledge of which sample had been exposed to an ETCHbox gene or a no-ETCHbox gene control. Cytotune 2.0 (Lot number 2010028) was added 24 hours after re-plating and left overnight. The following day the media was replaced to remove dead cells, this was repeated on day 2 and 3. On day 5 the media was changes and iPSC-like colonies were counted before being plated onto feeder cells (Mitomycin C-treated CBA MEFs, produced by Tim Davies, Paul Fairchild Lab, Sir William Dunn School of Pathology, University of Oxford). Cells on feeders were cultured in full mouse stem cell media. On day 10 number of colonies per 10,000 cytotune 2.0 treated cells was calculated. To maximise reliability and repeatability of results, two experimental repeats were conducted, each containing triplicate repeats for each sample.

2.5 Analysis of marker gene expression in ETCHbox+ iPSCs.

Quantitative PCR (qPCR) analysis of iPSCs pre-exposed to ETCHbox genes used the New England Biolabs Luna One-Step qPCR kit to manufacturer's instructions and Techne qPCR equipment. We tested for transcripts of three marker genes; *MuERVL*, *Nanog*, and *ZScan4c*. To ensure maximum reliability of the results I used three housekeeping genes, iPSC colonies of different ages (those marked as "late" had RNA extracted after an additional passaging and have an associated control sample) and from different experimental repeats. I only consider results consistent between two out of three housekeeping genes to be reliable.

Table 5.3: Primers used for qPCR of potency marker genes

Target Gene	Forward Primer	Reverse Primer	Fragment Size (bp)
Gapdh	CGGGTTCCTATAAATACGGACTG	AATACGGCCAAATCCGTTCA	145
ActinB	CTGCGTGTGGCCCCT	TGAAGGTCTCAAACATGATCTGG	103
Pgk1	TCCATGGTGGGTGTGAATCT	CTTTAGCGCCTCCCAAGATAG	119
Mu-ERVL	TCTCAGCCTCAGAGTGTGTC	GGGGCAAACCCTTGAGAATC	143
ZScan4c	TGGTCTTGAATGTCTTTTACAGT	CTGTGAAGCCATTGTGGTGA	122
Nanog	AGATGCAAGAACTCTCCTCCA	GATTCAGAAGGGCTCAGCAC	144

Gene expression of marker genes relative to each housekeeping gene was calculated for each sample and a no-homeobox gene control which had been prepared alongside ETCHbox experimental samples. Anomalous qPCR readings were discounted (defined as readings that took the variance of all readings for that sample >2) and statistical analysis comparing ETCHbox+ colonies to the control used the Welches unequal variances two-sample t-test.

3.Results

3.1 Mouse ETCHbox genes are expressed in pluripotent stem cells

I screened publicly available RNA-sequencing data from a variety of mouse tissues and cell lines to investigate systems which show endogenous ETCHbox expression outside of the preimplantation embryo (supplementary file 5.1). For the vast majority of mouse ETCHbox genes tested, no endogenous expression was recorded (figure 5.2). However, analysis of mouse iPS cells formed from MEFs indicated that *Oboxa1* and *Crxos* transcripts are expressed, although expression varies with cell age (passage) and is much lower in older iPS cell colonies (figure 5.2). A low level of expression of *Crxos* can be seen in the Embryonic Stem Cell (ESCs) transcriptomic data.

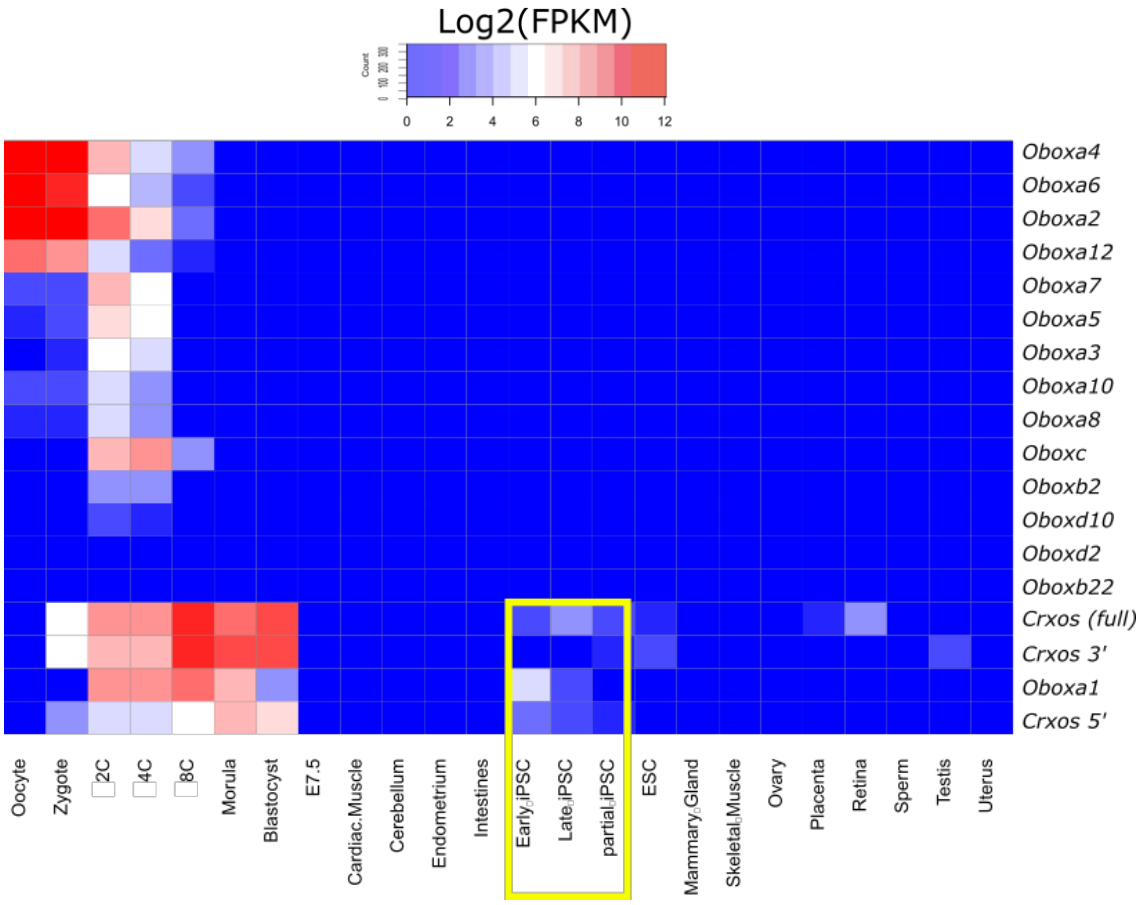


Figure 5.2. Expression of mouse ETCHbox genes is almost exclusively restricted to that observed in the preimplantation embryo. However, *Oboxa1* has RNA expression in early passage mouse iPSCs which declines in older passage iPSCs (yellow box) (supplementary file 5.2). 2C = two-cell embryo, 4C = four-cell embryo, 8C = eight-cell embryo, E7.6 = Embryonic day 7.5 embryo.

To check the repeatability of my findings in sequencing data, and to investigate the decline in expression with age further, I screened mouse iPS cells for ETCHbox expression across a variety of passage numbers spanning recently reprogrammed cells (passage 7) to older cells (passage 21) (figure 5.3). I also screened cells in the presence of feeders and at different stages of feeder removal to ask whether cell-cell interaction can influence ETCHbox expression (supplementary file 5.3), although it is difficult to definitively prove that expression is from stem cells and not feeders in this case. While other ETCHbox genes are somewhat inconsistently expressed in different iPS cell populations (eg. *Crxos*, *Oboxa4*),

Oboxa1 seems to be consistently expressed in all stem cell populations tested and not in feeder cells (figure 5.3). No mouse ETCHbox gene expression can be seen in the feeder cell population (figure 5.3).

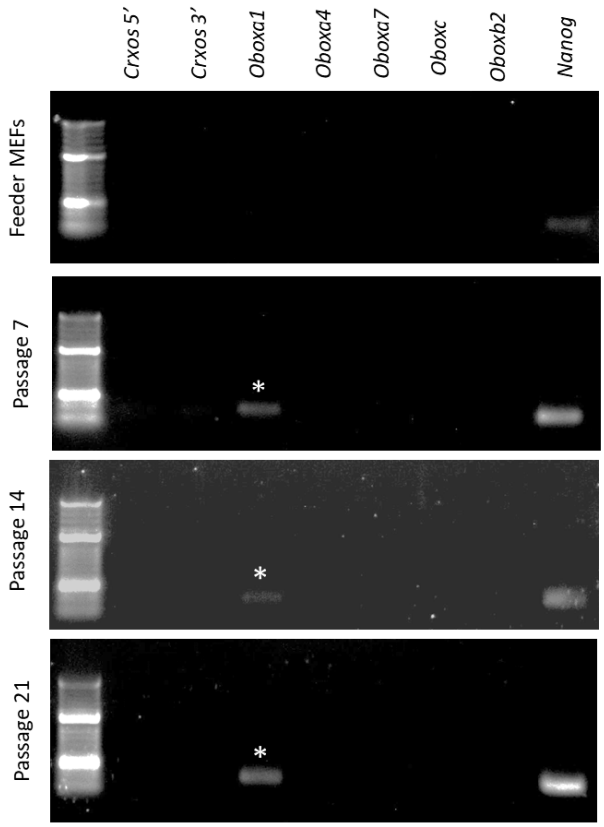


Figure 5.3. Expression of ETCHbox genes in iPSCs by RT-PCR. Polymerase chain reaction (PCR) of cDNA collected from mouse iPSCs at different passages shows that *Oboxa1* (*) is consistently expressed in mouse iPSCs whereas expression of other mouse ETCHbox genes cannot be identified here. *Nanog* expression was used as a positive control for pluripotency. No ETCHbox genes are expressed in the feeder cells (Feeder MEFs).

3.2 *Crxos* may directly interact with the pluripotency network.

To investigate further the expression of ETCHbox in pluripotent cells, and due to access to public large datasets, I examined ETCHbox expression by proxy using ATAC-seq data from iPSCs (Maza *et al.*, 2015) (supplementary file 5.1). This analysis revealed two peaks of open chromatin in the *Crxos* locus in mouse iPSCs, one of which located just upstream of the TSS of the transcript (supplementary file 5.4). When I analysed this peak for the presence of transcription factor (TF) binding motifs computationally, I find enrichment for the binding motif of the Yamanaka factor *Klf4* (supplementary file 5.5). I then examined public chromatin-immunoprecipitation-sequencing (ChIP-Seq) data probing the target DNA-binding sites of the Yamanaka factor *Klf4* in embryonic stem cells (Chronis *et al.*, 2017)(supplementary file 5.1). This analysis revealed a peak of *Klf4* binding within the first intron of the *Crxos* transcript tested here (supplementary file 5.4). Together with previous RNA-seq data following ectopic expression, this provides evidence that ETCHbox are intertwined in the pluripotency network, both regulated by and regulating OKSM factors (Royall et al 2018, Chapter 2).

3.3 Effect of ETCHbox ectopic expression on reprogramming efficiency.

As *Oboxa1* is expressed in iPSCs, I hypothesised that it functions somehow in the pluripotency gene network. Due to the overlapping expression patterns of ETCHbox genes in human and mouse, I investigated whether presence of ETCHbox genes from either species could influence the production or characteristics of iPSCs. In order to assess whether the presence of ETCHbox genes prior to the

reprogramming protocol can prime the differentiated cell to allow more efficient reprogramming, I ectopically expressed a selection of human and mouse ETCHbox genes in primary mouse embryonic fibroblasts (MEFs) prior to reprogramming and the number of established colonies were counted after 10 days (figure 5.4).

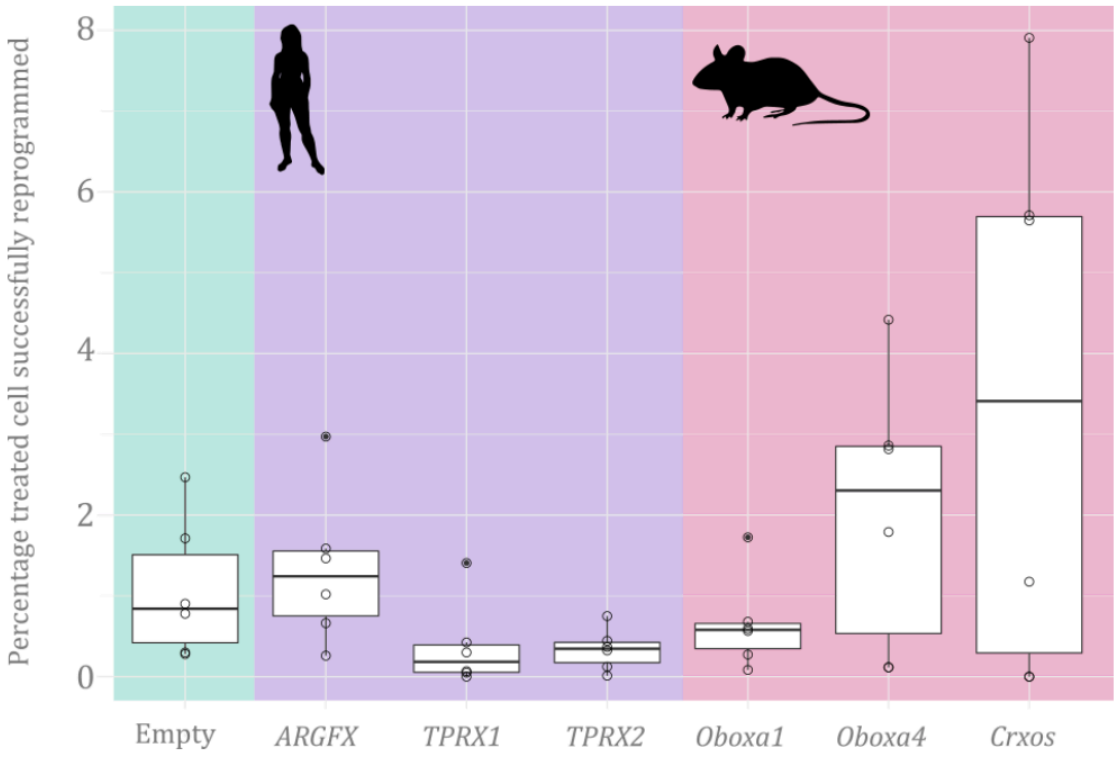


Figure 5.4 The effect of ETCHbox pre-exposure on the efficiency of reprogramming. The number of colonies visible 10 days post-treatment were converted into a percentage of the total number of treated cells (supplementary 5.6). Results shown here are combined results from two anonymous replicate studies. The results of six individual replicates per experimental sample show that there is no repeatable significant difference in reprogramming efficiency.

When any of the six ETCHbox genes experimentally over-expressed in this study were present pre-reprogramming, I find no reliable difference in the efficiency of reprogramming compared to a control (figure 5.4). The control consisted of differentiated cells transfected with only the puromycin-resistance gene prior to reprogramming treatment. Visually there is an indication that the

presence of mouse *Crxos, Oboxa4*, and human *ARGFX* prior to reprogramming can lead to increased number of derived colonies (figure 5.4, supplementary file 5.6), but with the data obtained this effect does not reach significance.

3.4 Regulation of two-cell embryo markers in iPSCs exposed to *Crxos* before reprogramming.

After the reprogramming of cells which were exposed to ectopically expressed ETCHbox genes, I tested whether the resulting colonies were different from those not pre-exposed to ETCHbox. Specifically, as ETCHbox expression precedes expression of OKSM in the ICM, would ETCHbox+ iPSCs have increased differentiation potential? I asked whether any differences in the expression of pluripotent or totipotent cell marker genes (Oh, 2017) could be assessed by qPCR (supplementary file 5.7) alongside a selection of housekeeping genes (Willems *et al.*, 2006).

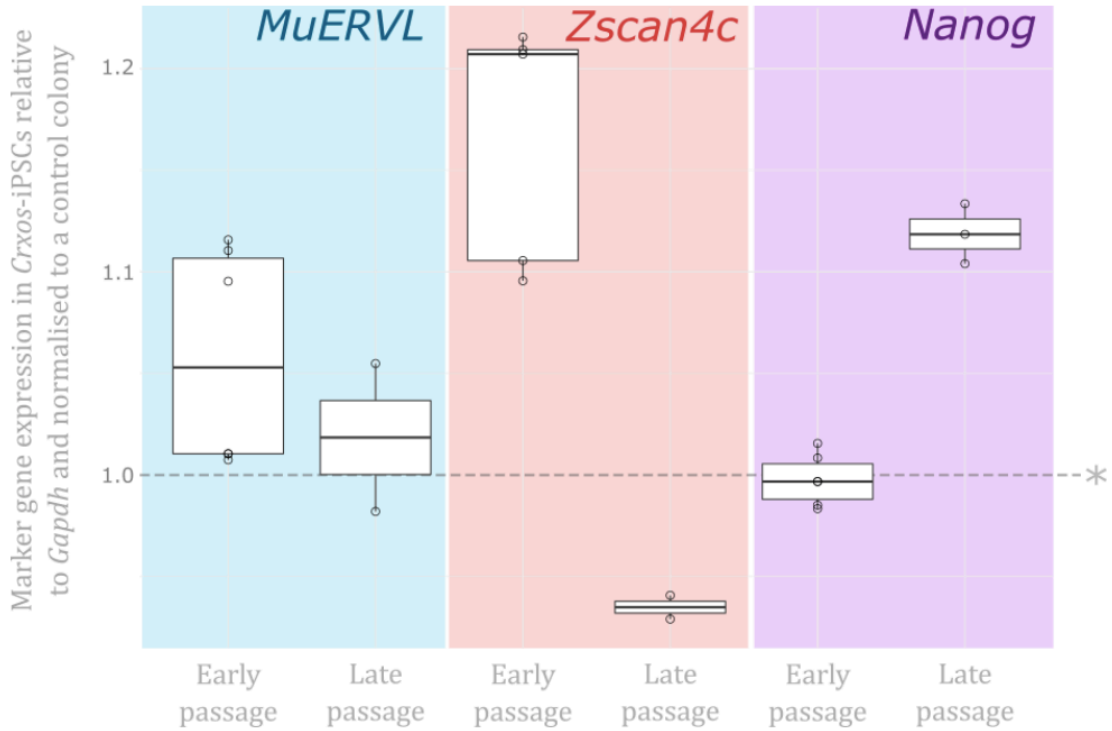


Figure 5.5. The effect of *Crxos* exposure before reprogramming alters the expression of totipotency markers. Normalised gene expression relative to *Gapdh* expression of two-cell embryo markers *MuERVL* and *Zscan4c*, and pluripotency marker *Nanog* in iPSC colonies exposed to *Crxos* and a control iPSC colony. *MuERVL* and *Zscan4c* are up-regulated in early (younger) colonies exposed to *Crxos* pre-reprogramming. *Nanog* is up-regulated in an older colony exposed to *Crxos* pre-reprogramming compared to a matched control iPSC colony. Early = Passage 3 iPSCs, Late = Passage 4 iPSCs. * indicates the control level of gene expression.

Colonies transfected with *Crxos* significantly up-regulated two-cell embryo (Totipotent) markers *MuERVL* and *Zscan4c* compared to a control and relative to at least two housekeeping genes (figure 5.5, table 5.4). Older colonies up-regulated *Nanog,* but not *MuERVL* or *Zscan4c,* when compared to control iPSCs and normalised using three housekeeping genes (figure 5.5, table 5.4).

Table 5.4 Statistical output following exposure of *Crxos* to fibroblasts before reprogramming.
Corrected P-Values from Welch Two-Sample T-Test comparing marker gene expression normalised to three housekeeping genes from three iPSC colonies exposed to *Crxos* before reprogramming, and a matched control iPSC. Control colonies were matched by day of transfections, puromycin treatment and reprogramming treatment. * indicated significantly reduced gene expression. qPCR data summary in supplementary file 5.7.

	P-value								
Crxos	*MuERVL*			*Zscan4c*			*Nanog*		
Colony	*Gapdh*	*ActB*	*Pgk1*	*Gapdh*	*ActB*	*Pgk1*	*Gapdh*	*ActB*	*Pgk1*
Early 1	0.006	0.0129	0.0006	0.02	0.005	ns	ns	ns	ns
Early 2	0.0003	0.0027	0.0009	0.00002	0.00005	0.00003	ns	0.04*	ns
Late	ns	ns	ns	ns	ns	ns	0.0009	0.0057	0.000165

3.5 Regulation of *Zscan4c* in colonies exposed to human ETCHbox genes.

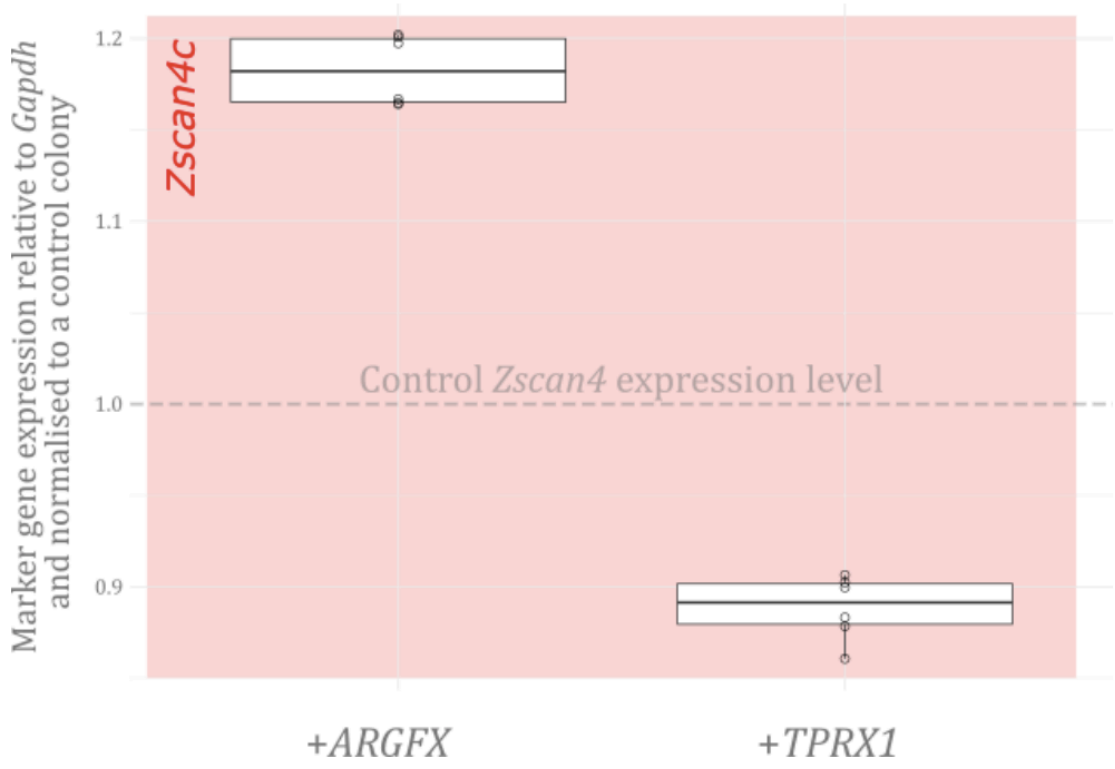


Figure 5.6. Exposure to human ETCHbox genes *ARGFX* and *TPRX1* prior to reprogramming alters the expression of *Zscan4c* in resulting stem cell colonies. qPCR results show that exposure to *ARGFX* prior to reprogramming can increase the subsequent expression of *Zscan4c* in established colonies. In contrast, exposure to *TPRX1* reduces *Zscan4c* expression. Gene expression relative to *Gapdh* and matched control iPSCs. +*ARGFX* = MEFs ectopically expressing *ARGFX* prior to reprogramming, +*TPRX1* = MEFs ectopically expressing *TPRX1* prior to reprogramming. P-values from Welch Two-sample t-test show significant differences from the control when normalised using three different housekeeping genes (supplementary file 5.8) qPCR data summary in supplementary file 5.6.

Interestingly, I also found an effect of human ETCHbox gene over-expression on gene expression in mouse iPSCs. Colonies that had *ARGFX* or *TPRX1* over-expressed before reprogramming showed altered expression levels of the early embryonic gene *Zscan4c* (figure 5.6*)*. *ARGFX* pre-exposure resulted in an increased level of the totipotency marker, in contrast to *TPRX1* which caused a down-regulation of *Zscan4c* (figure 5.6). Both these results translate into statistically significant differences (supplementary file 5.8).

4. Discussion

4.1 ETCHbox genes and the pluripotency network

The period of embryogenesis encompassing ETCHbox expression is a difficult system to manipulate in a laboratory context, even for common models such as mouse. The embryos need complex combinations of nutrients and growth substrates to survive, as well as researchers requiring ethical approval. To overcome this we have used primary embryonic cells from later embryonic stages to investigate the role of ETCHbox genes (Maeso *et al.*, 2016; Royall *et al.*, 2018, 2019), however this brings another caveat in that these systems generally lack the endogenous expression needed for knock-out and/or knock-down experimentation. I have in this study identified and verified a population of cells which endogenously express some mouse ETCHbox genes (figure 5.3). Moving forward, a variety of techniques are now available for investigating the function of Obox genes in their natural context; knock-out experiments followed by subsequent analysis of cell/embryo morphology and transcriptome are able to complement over-expression experiments. Here, the endogenous expression of ETCHbox identified by RNA-seq has allowed me to use other large datasets to increase understanding *Crxos'* place in the pluripotency network, potentially as a direct target of the Yamanaka factor *Klf4* (Sridharan *et al.*, 2009).

4.2 Exposure to ETCHbox genes might increase reprogramming efficiency.

After revealing the endogenous expression of mouse ETCHbox in iPSCs, I asked whether they could alter the efficiency of reprogramming if ectopically provided before treatment. I have found that three of six ETCHbox genes tested (human *ARGFX*, mouse *Oboxa4*, and *Crxos*) have the potential to increase efficiency when compared to a control. *Oboxa4* is a maternally deposited factor in the mouse Oocyte which remains expressed until mid-preimplantation development. Its effect on the reprogramming efficiency here is suggestive of *Oboxa4* expression causing a shift in the transcriptome to appear more like a zygotic or oocyte transcriptome, which *in vivo* naturally precedes native expression of the Yamanaka factors. The presence of *Oboxa4* prior to reprogramming may have led to the initiation of Yamanaka factor misexpression prior to their ectopic expression during reprogramming, allowing increased efficiency.

We have previously shown that *ARGFX* and *Crxos* have comparable downstream effects and both are expressed prior to EGA in their respective species. My results here may be related to an active role in promoting the transition to a pluripotent state which coincides with EGA in the embryo. However these results are not statistically significant, possibly reflecting high iPSC variability or some other factor, such as sensitivity to culturing media or environmental factors, which I was unable to control for. Additional study in this area is needed to test the repeatability of my results presented here.

During this research other researchers published findings indicating that *Oboxa1* and *Oboxa4* were indeed able to increase efficiency of reprogramming (Schiebinger *et al.*, 2017; Wu *et al.*, 2017). As noted above, I also obtained an

increase in reprogramming efficiency with prior exposure to *Oboxa4*, but how to correctly interpret my results is not entirely clear. The discrepancy of the results of this reprogramming efficiency experiment compared to the literature regarding Obox genes could be explained by differing experimental design (Schiebinger *et al.*, 2017; Wu *et al.*, 2017). In this experiment cells receive a pulse of gene expression which is not prolonged through the reprogramming process, whereas published studies maintain Obox expression through reprogramming using an inducible gene-expression system. This could mean that the increase in reprogramming efficiency noted in these studies is due to a role of Obox within the reprogramming network, rather than in the initiating "priming" stages which would be reflected in this experiment. Additionally, the complex gene organisation of the ETCHbox genes, with many exhibiting similar expression patterns *in vivo,* raises the possibility that any activity priming cells for reprogramming could be distributed across multiple genes.

4.3 Influence of *Crxos* exposure on totipotency markers in iPSCs.

Multiple studies of the earliest period of eutherian development have identified the transposable element *MuERV-L* and the *Zscan4c* gene as being some of the first transcripts identified during and following mouse embryonic genome activation at the two cell stage (Kigami *et al.*, 2003; Falco *et al.*, 2007; Schoorlemmer *et al.*, 2014; Choi *et al.*, 2017; Ishiguro *et al.*, 2017; Godini and Fallahi, 2018; Zhang *et al.*, 2019). At this time cells of the embryo are broadly described as totipotent; they can differentiate into all cells of the embryo and the extra-embryonic tissues. Expression of both *MuERV-L* and *Zscan4c* decreases as the blastomeres gradually

define their lineage and so I use them here as 'markers' of increased differential potential, akin to totipotency. Additionally, *Zscan4* expression is only seen in a small percentage of mouse ESCs at any given timepoint (Zalzman *et al.*, 2010) and can itself increase the reprogramming efficiency in iPSC generation (Hirata *et al.*, 2012), further suggesting that its expression *in vivo* precedes the pluripotent ESC state and can therefore be used here as a marker of iPSCs with increased potency. Conversely, *Nanog* expression is limited to the ICM of the blastocyst, when cell are pluripotent, and has been described as a pluripotency sustaining factor (Chambers *et al.*, 2003; Kaoru Mitsui *et al.*, 2013; Bates and Silva, 2017).

I find that iPSC colonies pre-exposed to *Crxos* up-regulate markers of totipotency (figure 5.5) (Kigami *et al.*, 2003; Schoorlemmer *et al.*, 2014; Ishiguro *et al.*, 2017; Godini and Fallahi, 2018). This regulation is not found in older colonies however in these I find up-regulation of the pluripotency marker *Nanog* (Chambers *et al.*, 2003; Kaoru Mitsui *et al.*, 2013; Blakeley *et al.*, 2015), a potential consequence of increased potency in a younger stem cell population (figure 5.5). We have previously shown that ectopic *Crxos* can alter the expression of genes which have dynamic expression as the embryo becomes a blastocyst. In the light of the findings reported here, I hypothesise that *Crxos* has an active involvement in the GRN defining changes in pluripotency of blastomeres at this embryonic stage, which is consistent with knockdown experiments (Cui *et al.*, 2016).Notably, I also find regulation of totipotency markers in mouse iPSC colonies pre-exposed to human ETCHbox genes *ARGFX* and *TPRX1* (figure 5.6A), suggesting that ETCHbox genes may have a conserved role, across eutherian species, in regulating the persistence and exit of blastomeres from totipotency.

It is notable that I do not find any reproducible effect of ectopic expression of *Oboxa1* on the regulation of the marker genes tested here, given that it is endogenously expressed in iPSCs (figure 5.3; supplementary file 5.6). This may indicate that *Oboxa1* sits at a lower level in a pluripotent network; its expression initiated by pluripotency master regulators and not vice versa.

4.4 Regulation of two-cell embryo markers by paralogous ETCHbox genes supports 'Distributed Robustness' theory of ETCHbox evolution.

We have previously suggested that the theory of distributed robustness may explain the complex and dynamic nature of ETCHbox gene repertoire and sequence across eutherian mammals (Chapter 2). Generally, distributed robustness refers to the robustness of a gene network being divided between its component parts by partial redundancy (Wagner, 2005). In this hypothesis, partially overlapping functions of a group genes in a context with high levels of gene duplication can allow for large amounts of gene sequence change and gene gain whilst preserving the underlying function of the group (Wagner, 2005; Royall *et al.*, 2018). I find here that exposure to each of the mouse ETCHbox genes *Crxos* and human ETCHbox gene *ARGFX* can lead to increased expression of the mouse totipotency marker *Zscan4c* in iPSCs. In addition to this, exposure to *Crxos* also lead to up-regulation of another totipotent cell marker, *Mu-ERVL*. This is a strong line of evidence in support of the distributed robustness theory as it exhibits an overlapping function of non-orthologous genes leading to up-regulation of the two-cell embryo marker *Zscan4c*. Not only this, but I find *Crxos* exposure is able to elicit up-regulation of a second

marker not found after *ARGFX* exposure. This would be expected under the theory of distributed robustness as only partial overlapping function is necessary.

Combining 1) that both can cause differential expression to genes involved in the totipotent state, with 2) previous findings that the effect of ectopic expression of *ARGFX* and *Crxos* have very similar but non-identical consequences on the global transcriptome, I suggest there is a vital and general ETCHbox function conserved between species and shared between genes. This function could define the timing of exit from the totipotent state, with some genes promoting totipotency and some diminishing it. This provides the context for distributed robustness; partially overlapping roles of ETCHbox genes prevent failure of the embryo to correctly exit totipotency, but allow for dramatic sequence and copy number variability.

4.5 Regulation of *Zscan4c* in colonies exposed to *ARGFX* and *TPRX1* recapitulates ETCHbox functional antagonism in human cells.

I find that iPSC colonies pre-exposed to the human ETCHbox genes *ARGFX* and *TPRX1* reliably and significantly differentially express, in opposing directions, the mouse totipotency marker *Zscan4c,* which accumulates in the 8-cell human embryo (Töhönen *et al.*, 2015). *ARGFX* and *TPRX1* have recently been shown by us and others to have opposing regulatory effects in other cell contexts (Madissoon *et al.*, 2016; Maeso *et al.*, 2016), a finding that is supported by my results. The up-regulation of *Zscan4c* in iPSCs pre-exposed to *ARGFX*, and down-regulation of those exposed to *TPRX1* suggests that whilst *ARGFX* may be acting to preserve totipotency in the human embryo, *TPRX1* may be acting to push the blastomeres away from totipotency. Both ETCHbox genes are expressed at a stage of human embryo

development when the embryonic genome is becoming active and the cells initiate their first lineage decision. Remarkably, the finding that *ARGFX* exposure alone could shift the transcriptome of iPSCs towards a totipotent state suggests that it may sit high up in the GRN which defines totipotency. Similarly, my results suggest that *TPRX1* can shift iPSCs to transcribe fewer genes in the totipotency network, and so may be a significant contributor to initiating the departure from totipotency. This is an important finding which may have not been identified by other studies exploring the human totipotent state as ETCHbox genes are not expressed in human ESCs/iPSCs, presumably because they display an epiblast 'primed' transcriptomic signature rather than an ICM 'naïve' one.

To develop these findings further differentiation assays of ETCHbox+ iPSCs will reveal their true state (De Los Angeles *et al.*, 2015), as well as repeating the assay described here with increased replicates. However, this current study has exemplified that using a mouse cell system to measure differentiation potential can reveal new insights into both the roles of genes at the earliest stages of development, and how these stages of development evolve and vary between species. I have also recapitulated published findings relating to ETCHbox function, which increases my confidence that the changes in marker gene expression and reprogramming efficiency are not artefacts of a variable experimental system.

It is notable that ETCHbox genes with a significant effect on marker gene expression (figures 5.5 and 5.6) also seem to be among the most likely to be able to increase reprogramming efficiency (figure 5.4). Moving forward, *ARGFX* and *Crxos* should remain a focus of research on the influence of ETCHbox on pluripotency.

4.6 Conclusions

The ETCHbox genes and the ICM are both eutherian novelties. The interaction of ETCHbox with the pluripotency network described here suggests that the emergence of ETCHbox genes may be intertwined with the evolution of a morphologically and transcriptionally distinct pluripotent ICM encased by trophectoderm, a defining feature of Eutherian embryogenesis and consistent with the hypothesis of DDC+ (Royall *et al.*, 2019, chapter 3). This chapter has also added to our understanding of how early embryonic genes can be used in applications in the medical field, such as increasing the yield and altering the potency of iPSCs.

5.References

Baker, C. L. and Pera, M. F. (2018) 'Capturing Totipotent Stem Cells', *Cell Stem Cell*. Elsevier Inc., 22(1), pp. 25–34. doi: 10.1016/j.stem.2017.12.011.

Bates, L. E. and Silva, J. C. (2017) 'Reprogramming human cells to naïve pluripotency: how close are we?', *Current Opinion in Genetics and Development*, 46, pp. 58–65. doi: 10.1016/j.gde.2017.06.009.

Baumann, K. (2017) 'Stem cells: A key to totipotency', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 18(3), p. 137. doi: 10.1038/nrm.2017.9.

Blakeley, P. *et al.* (2015) 'Defining the three cell lineages of the human blastocyst by single-cell RNA-seq', *Development*, 142(20), pp. 3613–3613. doi: 10.1242/dev.131235.

Boiani, M. *et al.* (2019) 'Totipotency continuity from zygote to early blastomeres – a model under revision', *Reproduction*, 10249, pp. 1–42. doi: doi.org/10.1530/REP-18-0462.

Boroviak, T. *et al.* (2018) 'Global transcriptome analysis reveals conserved and divergent features of primate preimplantation development', *Development, in revision*.

Casser, E. *et al.* (2017) 'Totipotency segregates between the sister blastomeres of two-cell stage mouse embryos', *Scientific Reports*. Springer US, 7(1), pp. 1–15. doi: 10.1038/s41598-017-08266-6.

Chambers, I. *et al.* (2003) 'Functional Expression Cloning of Nanog, a Pluripotency Sustaining Factor in Embryonic Stem Cells', 113, pp. 643–655. doi: 10.1016/S0092-8674(03)00392-1.

Chin, M. H. *et al.* (2009) 'Induced Pluripotent Stem Cells and Embryonic Stem Cells Are Distinguished by Gene Expression Signatures', *Cell Stem Cell*. Elsevier Ltd, 5(1), pp. 111–123. doi: 10.1016/j.stem.2009.06.008.

Choi, Y. J. *et al.* (2017) 'Deficiency of microRNA miR-34a expands cell fate potential in pluripotent stem cells', *Science*, 355(6325). doi: 10.1126/science.aag1927.

Chronis, C. *et al.* (2017) 'Cooperative Binding of Transcription Factors Orchestrates Reprogramming', *Cell*, pp. 442–459. doi: 10.1016/j.cell.2016.12.016.

Condic, M. L. (2013) 'Totipotency: What It Is and What It Is Not', *Stem Cells and Development*, 23(8), pp. 796–812. doi: 10.1089/scd.2013.0364.

Cui, W. *et al.* (2016) 'Towards Functional Annotation of the Preimplantation Transcriptome: An RNAi Screen in Mammalian Embryos', *Scientific Reports*, 6. doi: 10.1038/srep37396.

Dobin, A. *et al.* (2013) 'STAR: Ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Du, P. *et al.* (2018) 'An Intermediate Pluripotent State Controlled by MicroRNAs Is Required for the Naive-to-Primed Stem Cell Transition', *Cell Stem Cell*. Elsevier Inc., 22(6), pp. 851-864.e5. doi: 10.1016/j.stem.2018.04.021.

Falco, G. *et al.* (2007) 'Zscan4: A novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells', *Developmental Biology*, 307(2), pp. 539–550. doi: 10.1016/j.ydbio.2007.05.003.

Fogarty, N. M. E. *et al.* (2017) 'Genome editing reveals a role for OCT4 in human embryogenesis', *Nature*. Nature Publishing Group, 550(7674), pp. 67–73. doi: 10.1038/nature24033.

Frankenberg, S. R. *et al.* (2016) 'The mammalian blastocyst', *Wiley Interdisciplinary Reviews: Developmental Biology*, 5(2), pp. 210–232. doi: 10.1002/wdev.220.

Giulitti, S. *et al.* (2019) 'Direct generation of human naive induced pluripotent stem cells from somatic cells in microfluidics', *Nature Cell Biology*. Springer US, 21(2), pp. 275–286. doi: 10.1038/s41556-018-0254-5.

Godini, R. and Fallahi, H. (2018) 'Dynamics changes in the transcription factors during early human embryonic development', *Journal of Cellular Physiology*. doi: 10.1002/jcp.27386.

Griffith, O. W. *et al.* (2017) 'Embryo implantation evolved from an ancestral inflammatory attachment reaction', *Proceedings of the National Academy of Sciences*, p. 201701129. doi: 10.1073/pnas.1701129114.

Guo, G. *et al.* (2016) 'Naive Pluripotent Stem Cells Derived Directly from Isolated Cells of the Human Inner Cell Mass', *Stem Cell Reports*. The Authors, 6(4), pp. 437–446. doi: 10.1016/j.stemcr.2016.02.005.

Hassani, S.-N. *et al.* (2018) 'Transition of inner cell mass to embryonic stem cells: mechanisms, facts, and hypotheses', *Cellular and Molecular Life Sciences*. Springer International Publishing, (0123456789), pp. 1–20. doi: 10.1007/s00018-018-2965-y.

Hirata, T. *et al.* (2012) 'Zscan4 transiently reactivates early embryonic genes during the generation of induced pluripotent stem cells', *Scientific Reports*, 2, pp. 1–11. doi: 10.1038/srep00208.

Ishiguro, K. ichiro *et al.* (2017) 'Expression analysis of the endogenous Zscan4 locus and its coding proteins in mouse ES cells and preimplantation embryos', *In Vitro Cellular and Developmental Biology - Animal*. In Vitro Cellular & Developmental Biology - Animal, 53(2), pp. 179–190. doi: 10.1007/s11626-016-0097-y.

Kaoru Mitsui *et al.* (2013) 'The Homeoprotein Nanog Is Required for Maintenance

of Pluripotency in Mouse Epiblast and ES Cells', 113, pp. 631–642. doi: 10.1016/S0092-8674(03)00393-3.

Karagiannis, P. and Takashima, Y. (2017) 'Surface Markers Guide the Journey toward Naive Pluripotency', *Cell Stem Cell*. Elsevier Inc., 20(6), pp. 737–738. doi: 10.1016/j.stem.2017.05.004.

Kigami, D. *et al.* (2003) 'MuERV-L Is One of the Earliest Transcribed Genes in Mouse One-Cell Embryos1', *Biology of Reproduction*, 68(2), pp. 651–654. doi: 10.1095/biolreprod.102.007906.

Kilens, S. *et al.* (2018) 'Parallel derivation of isogenic human primed and naive induced pluripotent stem cells', *Nature Communications*, 9(1), pp. 1–13. doi: 10.1038/s41467-017-02107-w.

Kojima, Y. *et al.* (2017) 'Evolutionarily Distinctive Transcriptional and Signaling Programs Drive Human Germ Cell Lineage Specification from Pluripotent Stem Cells', *Cell Stem Cell*. Elsevier Inc., 21(4), pp. 517-532.e5. doi: 10.1016/j.stem.2017.09.005.

Kuijk, E., Geijsen, N. and Cuppen, E. (2015) 'Pluripotency in the light of the developmental hourglass', *Biological Reviews*, 90(2), pp. 428–443. doi: 10.1111/brv.12117.

Ladstätter, S. and Tachibana, K. (2018) 'Genomic insights into chromatin reprogramming to totipotency in embryos', *The Journal of Cell Biology*, p. jcb.201807044. doi: 10.1083/jcb.201807044.

Li, L. *et al.* (2018) 'Single-cell multi-omics sequencing of human early embryos', *Nature Cell Biology*. Springer US, 20(July). doi: 10.1038/s41556-018-0123-2.

De Los Angeles, A. *et al.* (2015) 'Hallmarks of pluripotency', *Nature*, 525(7570), pp. 469–478. doi: 10.1038/nature15515.

Madissoon, E. *et al.* (2016) 'Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos', *Scientific Reports*. Nature Publishing Group, 6(July), pp. 1–14. doi: 10.1038/srep28995.

Maeso, I. *et al.* (2016) 'Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals', *BMC Biology*, 14(1), pp. 1–14. doi: 10.1186/s12915-016-0267-0.

Maza, I. *et al.* (2015) 'Transient acquisition of pluripotency during somatic cell transdifferentiation with iPSC reprogramming factors', *Nature Biotechnology*. Nature Publishing Group, 33(7), pp. 769–774. doi: 10.1038/nbt.3270.

Ménézo, Y. J. R. and Hérubel, F. (2002) 'Mouse and bovine models for human IVF.', *Reproductive biomedicine online*, 4(2), pp. 170–175. doi: 10.1016/S1472-6483(10)61936-0.

Messmer, T. *et al.* (2019) 'Transcriptional Heterogeneity in Naive and Primed Human Pluripotent Stem Cells at Single-Cell Resolution', *Cell Reports*, 26(4), pp. 815-824.e4. doi: 10.1016/j.celrep.2018.12.099.

Niakan, K. K. and Eggan, K. (2013) 'Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse', *Developmental Biology*. Elsevier, 375(1), pp. 54–64. doi: 10.1016/j.ydbio.2012.12.008.

Oh, S. (2017) 'Reprogramming Pluripotent Stem Cell Towards Totipotency', *Northwestern University Interdepartmental Neuroscience Program*, (Thesis submitted December 2017).

Park, J. *et al.* (2017) 'Identification of the early and late responder genes during the generation of induced pluripotent stem cells from mouse fibroblasts', *Plos One*, 12(2), p. e0171300. doi: 10.1371/journal.pone.0171300.

Pastor, W. A. *et al.* (2018) 'TFAP2C regulates transcription in human naive pluripotency by opening enhancers', *Nature Cell Biology*. Springer US, 20(5), pp. 553–564. doi: 10.1038/s41556-018-0089-0.

Polo, J. M. *et al.* (2010) 'Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells', *Nature Biotechnology*. Nature Publishing Group, 28(8), pp. 848–855. doi: 10.1038/nbt.1667.

Robinson, J. T. (2012) 'Integrated genomics viewer', *Nature biotechnology*, 29(1), pp. 24–26. doi: 10.1038/nbt.1754.Integrative.

Rossant, J. (2016) 'Genetic Control of Early Lineages in Mammalian Embryos', *Annu Rev Genet*, 51(1), p. annurev-genet-120116-024544. doi: 10.1146/annurev-genet-120116.

Rostovskaya, M., Stirparo, G. G. and Smith, A. (2019) 'Capacitation of human naïve pluripotent stem cells for multi-lineage differentiation', *Development*, 146(7), p. dev172916. doi: 10.1242/dev.172916.

Royall, A. H. *et al.* (2018) 'Mouse Obox and Crxos modulate preimplantation transcriptional profiles revealing similarity between paralogous mouse and human homeobox genes', *EvoDevo*. BioMed Central, 9(1), p. 2. doi: 10.1186/s13227-018-0091-4.

Royall, A. H. *et al.* (2019) 'Of eyes and embryos : subfunctionalization of the CRX homeobox gene in mammalian evolution', *Proceedings of the Royal Society B: Biological Sciences*.

Sato, S. *et al.* (2015) 'Partial inhibition of differentiation associated with elevated protein levels of pluripotency factors in mouse embryonic stem cells expressing exogenous EGAM1N homeoprotein', *Journal of Bioscience and Bioengineering*,

120(5), pp. 562–569. doi: http://dx.doi.org/10.1016/j.jbiosc.2015.03.001.

Schiebinger, G. *et al.* (2017) 'Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming.', *bioRxiv (preprint)*, p. 191056. doi: 10.1101/191056.

Schoorlemmer, J. *et al.* (2014) 'Regulation of Mouse Retroelement MuERV-L/MERVL Expression by REX1 and Epigenetic Control of Stem Cell Potency', *Frontiers in Oncology*, 4(February), pp. 1–18. doi: 10.3389/fonc.2014.00014.

Sridharan, R. *et al.* (2009) 'Role of the Murine Reprogramming Factors in the Induction of Pluripotency', *Cell*. Elsevier Inc., 136(2), pp. 364–377. doi: 10.1016/j.cell.2009.01.001.

Suwińska, A. *et al.* (2008) 'Blastomeres of the mouse embryo lose totipotency after the fifth cleavage division: Expression of Cdx2 and Oct4 and developmental potential of inner and outer blastomeres of 16- and 32-cell embryos', *Developmental Biology*, 322(1), pp. 133–144. doi: 10.1016/j.ydbio.2008.07.019.

Takahashi, K. and Yamanaka, S. (2006) 'Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors', *Cell*, 126(4), pp. 663–676. doi: 10.1016/j.cell.2006.07.024.

Takashima, Y. *et al.* (2015) 'Erratum: Resetting Transcription Factor Control Circuitry toward Ground-State Pluripotency in Human (Cell (2014) 158 (1254-1269))', *Cell*. Elsevier, 162(2), pp. 452–453. doi: 10.1016/j.cell.2015.06.052.

Tesar, P. J. *et al.* (2007) 'New cell lines from mouse epiblast share defining features with human embryonic stem cells', *Nature*, 448(7150), pp. 196–199. doi: 10.1038/nature05972.

Theunissen, T. W. *et al.* (2016) 'Molecular Criteria for Defining the Naive Human Pluripotent State', *Cell Stem Cell*, 19(4), pp. 502–515. doi: 10.1016/j.stem.2016.06.011.

Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. (2013) 'Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration', *Briefings in Bioinformatics*, 14(2), pp. 178–192. doi: 10.1093/bib/bbs017.

Töhönen, V. *et al.* (2015) 'Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development', *Nature Communications*, 6, pp. 1–9. doi: 10.1038/ncomms9207.

Wagner, A. (2005) 'Distributed robustness versus redundancy as causes of mutational robustness', *BioEssays*, 27(2), pp. 176–188. doi: 10.1002/bies.20170.

Weinberger, L. *et al.* (2016) 'Dynamic stem cell states: Naive to primed pluripotency in rodents and humans', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 17(3), pp. 155–169. doi: 10.1038/nrm.2015.28.

Wernig, M. *et al.* (2008) 'c-Myc Is Dispensable for Direct Reprogramming of Mouse Fibroblasts', *Cell Stem Cell*, 2(1), pp. 10–12. doi: 10.1016/j.stem.2007.12.001.

Willems, E. *et al.* (2006) 'Selection of reference genes in mouse embryos and in differentiating human and mouse ES cells', *International Journal of Developmental Biology*, 50(7), pp. 627–635. doi: 10.1387/ijdb.052130ew.

Wu, J. *et al.* (2018) 'Chromatin analysis in human early development reveals epigenetic transition during ZGA', *Nature*, 557(7704), pp. 256–260. doi: 10.1038/s41586-018-0080-8.

Wu, L. *et al.* (2017) 'Oocyte-Specific Homeobox 1, Obox1, Facilitates Reprogramming by Promoting Mesenchymal-to-Epithelial Transition and Mitigating Cell Hyperproliferation', *Stem Cell Reports*, 9(5), pp. 1692–1705. doi: 10.1016/j.stemcr.2017.09.012.

Xue, Z. *et al.* (2013) 'Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing', *Nature*. Nature Publishing Group, 500(7464), pp. 593–597. doi: 10.1038/nature12364.

Zalzman, M. *et al.* (2010) 'Zscan4 regulates telomere elongation and genomic stability in ES cells', *Nature*. Nature Publishing Group, 464(7290), pp. 858–863. doi: 10.1038/nature08882.

Zhang, W. *et al.* (2019) 'Zscan4c activates endogenous retrovirus MERVL and cleavage embryo genes', *Nucleic Acids Research*, 47(16), pp. 8485–8501. doi: 10.1093/nar/gkz594.

Zhang, Y. *et al.* (2008) 'Model-based analysis of ChIP-Seq (MACS)', *Genome Biology*, 9(9). doi: 10.1186/gb-2008-9-9-r137.

Ziomek, C. A., Johnson, M. H. and Handyside, A. H. (1982) 'The developmental potential of mouse 16-cell blastomeres', *Journal of Experimental Zoology*, 221(3), pp. 345–355. doi: 10.1002/jez.1402210310.

6. Supplementary files

5.1 Sequence Read Archive (SRA) identification numbers for public data used in this chapter

5.2 FPKM extracted from public RNA-sequencing data used for heatmap figure

5.3 RT-PCR gel image showing expression of mouse ETCHbox genes in mouse iPSCs throughout the feeder-removal process. iPSC colonies with a decimal identifier have had feeders gradually removed

5.4 Integrative Genome Viewer screenshot of iPSC ATAC-seq and Klf4-ChIP-seq peaks aligned to mouse *Crxos* locus.

5.5 Extracted ATAC peak 65643 annotated for transcription factor binding sites

5.6 Reprogramming efficiencies following ETCHbox ectopic expression in fibroblasts prior to reprogramming

5.7 Summary of qPCR results for potency markers in iPSCs exposed to ETCHbox genes

5.8 P-values of welches unequal vairances two sample t-test testing whether expression of ZScan4c is different in iPSCs pre-exposed to empty control plasmid, ARGFX, or TPRX1.

5.1: Public SRA data IDs

Tissue	Type	SRA number	Tissue	Type	SRA number
Retina	RNA	SRR3957210	Skeletal Muscle	RNA	SRR3744935
Retina	RNA	SRR3957211	Skeletal Muscle	RNA	SRR3744936
Retina	RNA	SRR3957212	Skeletal Muscle	RNA	SRR3744937
Retina	RNA	SRR3957213	Skeletal Muscle	RNA	SRR3744938
Testis	RNA	SRR5224362	Cardiac Myocytes	RNA	SRR5363117
Testis	RNA	SRR5224368	Cardiac Myocytes	RNA	SRR5363118
Testis	RNA	SRR5224369	Cardiac Myocytes	RNA	SRR5363119
Testis	RNA	SRR5224376	Cardiac Myocytes	RNA	SRR5363120
Testis	RNA	SRR5224377	Cardiac Myocytes	RNA	SRR5363121
Ovary	RNA	SRR2065460	Intestine	RNA	SRR5047870
Ovary	RNA	SRR2065461	Intestine	RNA	SRR5047871
Ovary	RNA	SRR2065462	Intestine	RNA	SRR5047872
Ovary	RNA	SRR2065463	Uterus	RNA	SRR4017306
Ovary	RNA	SRR2065464	Uterus	RNA	SRR4017307
Ovary	RNA	SRR2065465	Uterus	RNA	SRR4017308
Ovary	RNA	SRR2065466	Uterus	RNA	SRR4017309
Cerebellum	RNA	SRR3821733	Uterus	RNA	SRR4017310
Cerebellum	RNA	SRR3821734	Uterus	RNA	SRR4017311
Cerebellum	RNA	SRR3821735	Sperm	RNA	SRR4238840
Cerebellum	RNA	SRR3821736	Sperm	RNA	SRR4238841
J1 ESCs	RNA	SRR5027802	Sperm	RNA	SRR4238842
J1 ESCs	RNA	SRR5027803	Early iPSCs	RNA	SRR5188289
Placenta	RNA	SRR5312420	Early iPSCs	RNA	SRR5188288
Placenta	RNA	SRR5312421	Late iPSCs	RNA	SRR5188292
Placenta	RNA	SRR5312425	Late iPSCs	RNA	SRR5188293
Endometrium	RNA	SRR5041731	Part reprogrammed iPSCs	RNA	SRR1047505
Endometrium	RNA	SRR5041732	Part reprogrammed iPSCs	RNA	SRR1047506
Endometrium	RNA	SRR5041733	iPSC	ATAC	SRR1930171
Mammary Stromal Fibroblast	RNA	SRR3724460	iPSC	ATAC	SRR1930172
Mammary Stromal Fibroblast	RNA	SRR3724461	ESC	Klf4 ChIP	SRR5077692
Mammary Stromal Fibroblast	RNA	SRR3724462			
Skeletal Muscle	RNA	SRR3744933			
Skeletal Muscle	RNA	SRR3744934			

5.2 FPKM table used for heatmap

Tissue Key:

Oocy: Oocyte

Zyg: Zygote

2C: Two-cell embryo

4C: Four-cell embryo

8C: Eight-cell embryo

Mor: Morula

Blast: Blastocyst

Cmusc: Cardiac Muscle

Cere: Cerebellum

Endo: Endometrium

Inte: Intestine

EiPSC: Early iPSC

LiPSC: Late iPSC

piPSC: Partial iPSC

ESC: Embryonic Stem Cell

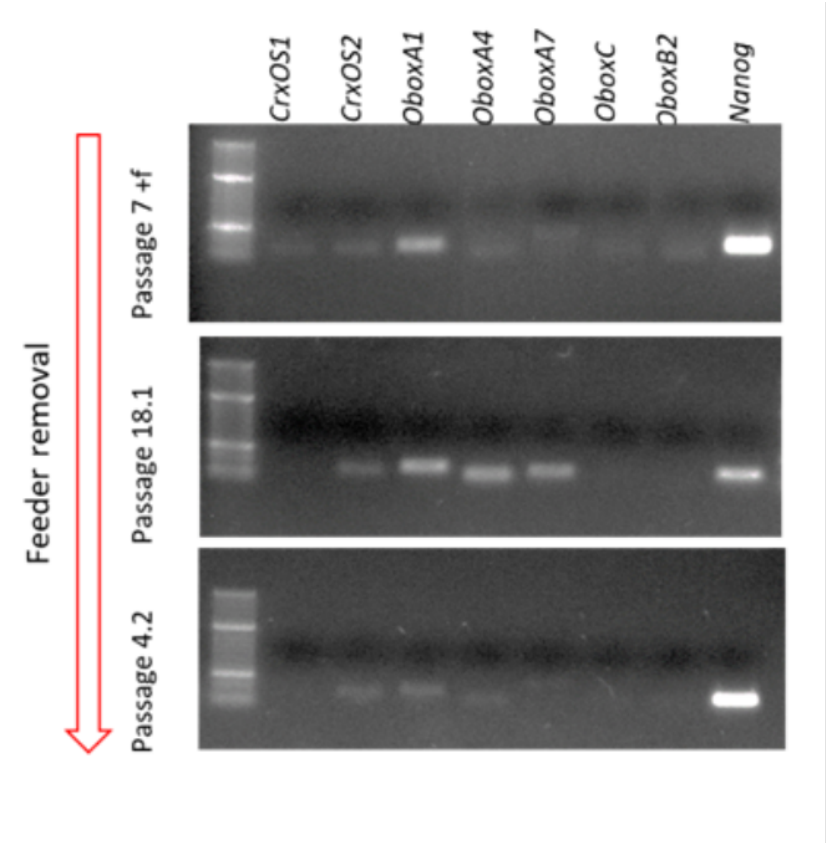
Mam: Mammary Gland

Smusc: Skeletal Muscle

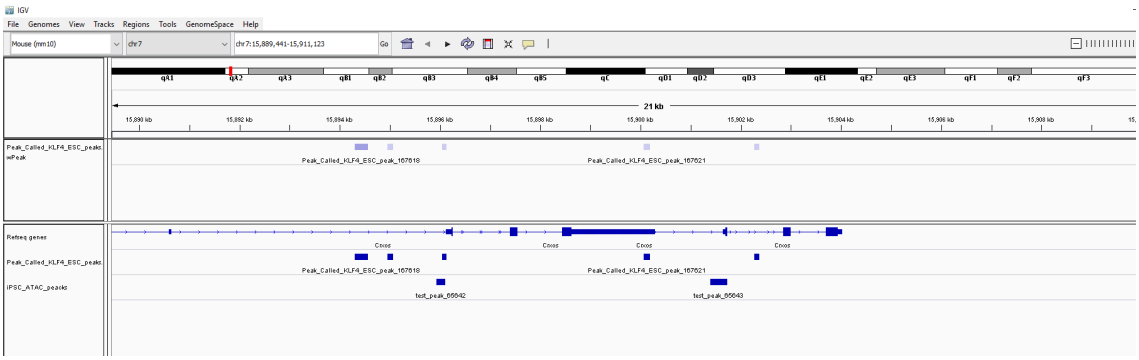
Plac: Placenta

Gene	Oocy	Zyg	2C	4C	8C	Mor	Blast	E7.5	Cmusc	Cere	Endo	Inte	EiPSC	LiPSC	piPSC	ESC	Mam	Smusc	Ovary	Plac	Retina	Sperm	Testis	Uterus
CrxOS Full	0.7	84.8	398.0	386.0	2043.0	650.0	1290.0	0.0	0.0	0.0	0.0	0.0	2.2	14.9	2.3	1.0	0.0	0.0	0.0	1.5	13.9	0.0	0.0	0.0
CrxOS 5'	0.3	8.7	35.7	46.0	74.0	177.0	135.0	0.0	0.0	0.0	0.0	0.0	7.3	2.9	1.0	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
CrxOS 3'	0.0	64.3	239.0	236.0	1688.0	952.0	1148.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4	2.4	0.0	0.0	0.1	0.0	0.0	0.0	3.1	0.0
ObоxA1	0.1	0.7	394.0	406.0	749.0	228.0	13.6	0.0	0.0	0.0	0.1	0.2	36.1	2.9	0.1	0.1	0.0	0.0	0.0	0.0	0.4	0.0	0.3	0.0
ObоxA2	3882.0	2688.0	529.0	134.0	4.9	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
ObоxA3	0.5	1.5	75.0	33.4	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ObоxA4	4317.2	2993.0	174.0	46.8	9.2	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ObоxA10	3.0	3.2	36.0	13.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ObоxA12	506.0	426.0	30.0	4.4	1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ObоxA5	2.0	2.7	107.0	59.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ObоxA7	2.6	3.1	174.0	76.0	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ObоxA8	0.9	2.0	43.0	11.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ObоxD10	0.1	0.0	4.1	1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ObоxAB2	0.0	0.1	0.5	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ObоxB2	0.0	0.0	8.8	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ObоxB22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0
ObоxC	0.2	0.0	255.0	366.0	14.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
psOboxA6	3606.7	1415.8	78.7	25.2	3.6	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5.3 ETCHbox iPSC RT-PCR



5.4 ATAC-seq and KLF4 ChIP-seq peaks



5.5 Crxos ATAC peak (65643) Transcription factor binding sites

CTGGAATCAAGATCCTAGTGAGGCCATGCTGCCCTCTGGGCTCTGGA
GGATTGAGGAGGCCAGTGTGGCCAAGTCAGTTTTTTTCCCCTCTCAG
AGTAGCTACTCAGGAAGACCTTGAAACCTTGGAAGGGAGTGTCTGG
AGGTTCCAGGGGGGAGTGGCAGAAAACAAATGAGTGGGAGGGGCT
CCTGGTGGGAGGGTCCTAAATAGGAGGTTTCAATATGGAAGTGGGA
TTTAAGCAAAAAAGTCCTAGAGAAGAGCCTCAAACTGAAGGAGGGA
ACTCAAGTGGGCAGCAGCCTCGTGATGGAAGAAATCGACGGGATGA
TAGTTCAA**ATGCGTG**A

Klf4, Foxo1, Tcfl5, Nr5a2, Hif1a, Hand1, Arntl, **Arnt**

5.6 Reprogramming efficiencies following ETCHbox ectopic expression in fibroblasts prior to reprogramming.

Experiment 1 was carried out inadequately blinded, the analysis by Tim Davies (Sir William Dunn School) in repeats 2 and 3 were fully blind and experiment ordering was randomised.

Gene	Experiment repeat	% cells on day -1 reprogrammed	Average	SD
Empty	2	0.28	1.073977	0.784957
Empty	2	0.78		
Empty	2	0.3		
Empty	3	**0.903158**		
Empty	3	**1.712281**		
Empty	3	**2.468421**		
ARGFX	2	0.26	1.327636	0.86253
ARGFX	2	1.02		
ARGFX	2	2.971429		
ARGFX	3	**1.465263**		
ARGFX	3	**1.585965**		
ARGFX	3	**0.663158**		
TPRX1	2	0.05	0.375	0.485373
TPRX1	2	0		
TPRX1	2	0.3		
TPRX1	3	**0.424286**		
TPRX1	3	**0.068571**		
TPRX1	3	**1.407143**		
TPRX2	2	0.444444	0.337199	0.235925
TPRX2	2	0.75		
TPRX2	2	0.325		
TPRX2	3	**0.12375**		
TPRX2	3	**0.3675**		
TPRX2	3	**0.0125**		
Oboxa1	2	0.68	0.654001	0.520927
Oboxa1	2	0.273684		
Oboxa1	2	0.084211		
Oboxa1	3	**0.566667**		
Oboxa1	3	**0.594444**		
Oboxa1	3	**1.725**		
Oboxa4	2	2.861538	2.020182	1.549179
Oboxa4	2	4.418182		
Oboxa4	2	2.818182		
Oboxa4	3	**0.115942**		
Oboxa4	3	**0.115942**		
Oboxa4	3	**1.791304**		
Crxos	2	5.711111	3.406786	3.129769
Crxos	2	7.907692		
Crxos	2	5.646154		
Crxos	3	**1.175758**		
Crxos	3	**0**		
Crxos	3	**0**		

5.7 qPCR results summary

Green = up regulated, Blue = down-regulated, Red = not significant

ETCHbox	Colony	*MuERVL*			*Nanog*			*ZScan4c*		
		HK1	HK2	HK3	HK1	HK2	HK3	HK1	HK2	HK3
Oboxa1	1	ns	u	ns	ns	u	ns	ns	u	ns
	2	ns	ns	ns	ns	ns	ns	ns	ns	ns
	3 - late	ns	d	d	u	u	ns	u	ns	ns
Oboxa4	1									
	2									
	3 - late									
Crxos	1									
	2									
	3 - late									
ARGFX	1									
	2 -late									
	3 - late									
TPRX1	1									
	2									
	3 - late									
TPRX2	1									
	2									
	3 - late									

5.8 P-values for ZScan4 expression levels in iPSCs exposed to *ARGFX* and *TPRX1*

(* = down-regulated, X = data not available, Early = passage 3 iPSC colony, Late = passage 4 colony).

P-values:

ARGFX	*Zscan4c*		
	Gapdh	*ActB*	*Pgk1*
Early	0.039	X	X
Late	0.002	0.001	0.003

TPRX1	*Zscan4c*		
	Gapdh	*ActB*	*Pgk1*
Early	4.4×10^{-5}*	5.7×10^{-5}*	5.1×10^{-5}*
Late	0.0009*	0.00015*	0.00033*

Chapter 6: General Discussion and Perspectives

1.Synopsis

This thesis contributes to our understanding of the highly dynamic ETCHbox genes, their distribution, function, evolution, and potential applications in bioscience and medicine. I have conducted the first comprehensive analysis of all known Obox loci, including elucidating multiple expression patterns indicating both *cis*-regulatory and protein coding sequence evolution. I have revealed striking functional overlap between non-orthologous genes, challenging our understanding of inheritance of gene function. I have also shown that after their evolution, the ETCHbox genes have both subfunctionalised and neofunctionalised, potentially influencing lineage specific traits such as the hollowed blastocyst and the inner cell mass, and blurring the lines of the classical DDC model of gene evolution after duplication (Force *et al.*, 1999).

Further evolutionary analysis has allowed me to show that the variability and dynamic nature of ETCHbox gene distribution are higher than previously thought (Maeso *et al.*, 2016), and that the mouse genome is not particularly unique in its dramatic sequence and gene copy number divergence. Finally, I have tested the potential of ETCHbox genes to influence how we reprogram differentiated cells into iPSCs. This area of my work is interesting as it could open future lines of enquiry into their effect on increasing the potency of the resulting stem cells, and allowing for the possibility that ETCHbox genes could increase the efficiency, and widen the potential applications, of the gold standard reprogramming protocol.

2. Protein coding sequence changes and evolution

Throughout the experimental work in this thesis I have tested the effect of coding sequence change on gene evolution. When we think about mechanisms that facilitate the evolution of regulatory networks, the most common explanation is that of modularity conferred by *cis*-regulatory elements (Ludwig *et al.*, 2005). The idea is that evolution can change the profile or level of expression of a given gene, allowing for the reorganisation and evolution of a network without the risk of damaging conserved protein domains; this view is common in the field of evolutionary developmental biology (Wray, 2007; Carroll, 2008). There have been multiple descriptions of how pleiotropic enhancer elements can influence evolution by allowing for the deployment of a gene in multiple contexts (Villar *et al.*, 2015; Buffry, Mendes and McGregor, 2016). However, another mechanism for changing gene function is to change the amino acid sequence potentially by alternative splicing to alter transcript sequence (Chen, Tovar-Corona and Urrutia, 2012), or alternatively by fixation of nonsynonymous mutations in the coding genomic sequence.

Protein sequence change risks causing loss-of-function mutations in conserved functional domains, or protein mis-folding, and so may be imagined to be more constrained than *cis*-regulatory element evolution (Drummond and Wilke, 2008). The potentially dramatic consequences of this mean one may expect protein sequence change to be selected against and hence not contribute to larger scale network evolution. However, in transcription factor evolution, changes to the DNA-binding motifs can alter the target genes downstream, or the level to which target genes are expressed and repressed (for *bicoid* example in Drosophila see Shaw *et*

al., 2002). This allows for more subtle changes in the gene regulatory network which are less likely to be highly detrimental to an organism.

My work in this thesis has demonstrated that protein sequence change in the ETCHbox genes is very rapid, has not been purged by evolutionary mechanisms, and has occurred alongside regulatory element evolution. All my ectopic expression experiments are testing only the protein sequence changes between ETCHbox genes, as the regulatory elements are not present in my system. This work provides an example of coding sequence and transcription factor target evolution occurring very rapidly in homeobox genes, which are commonly very highly conserved throughout the animal kingdom (Booth and Holland, 2007; Bürglin, 2011). In summary, my work has demonstrated that the effect of coding sequence change should not be overlooked when studying the evolution of gene regulatory networks.

3. Functional overlaps and innovation in ETCHbox genes; robustness and evolution

Multiple times throughout this research I have found evidence for functional overlaps between ETCHbox genes. Overlaps have been described in various forms; overlapping functions of ETCHbox between species, overlapping effects of ectopic expression of ETCHbox gene of the same species (Royall *et al.*, 2018), and overlapping involvement in wider regulatory networks.

Specifically, I have demonstrated that following ectopic expression there is a comparable effect on the transcriptome of the non-orthologous human *ARGFX* and mouse *Crxos* ETCHbox genes, highlighting overlap between species. I also show that

the genes differentially regulated following all three mouse Obox genes tested are enriched for the same temporal expression profile, and there are 611 genes that are differentially regulated also by all Obox genes tested. This demonstrates functional overlap between different ETCHbox genes in the same species. Finally, I have shown overlap in the general pathways affected by ectopic expression of these genes. In particular, through my iPSC experiments I have demonstrated additional evidence that human *ARGFX* and *TPRX1* regulate the same suite of genes, although in opposing directions (ARGFX causes up-regulation of totipotency markers whereas TPRX1 causes downregulation in iPSCS, Chapter 5). Additionally I have shown that mouse *Crxos* and *Oboxa1* likely overlap in that they are both involved in the pluripotency network, with ectopic *Crxos* altering the expression of potency markers within stem cells *in vitro*, and endogenous *Oboxa1* being expressed in pluripotent stem cells *in vivo*.

These results together suggest that there is a large amount of functional overlap between the ETCHbox genes, both within and between species, which could shed light on their functional evolution. I have suggested distributed robustness as a possible underlying property which affects ETCHbox evolution and the high levels of sequence and copy number divergence observed (Wagner, 2005; Royall *et al.*, 2018, Chapters 2 and 5). Building on this, I suggest that the ETCHbox genes function at some level as a group, sharing some functions and a complex integrative network downstream. Functional overlap in this context confers robustness to a network implicated in the earliest stages of embryo development as suggested in chapter 2, but also under certain circumstances could facilitate evolutionary change. As all of the overlaps described are partial (in that two genes do not function identically in my experiments) robustness here could release evolutionary constraint.

Reiterating the Distributed Robustness hypothesis in chapters 2 and 5, I suggest these partially overlapping functions not only allow a group of genes to retain a vital function, under some circumstances this could allow an ETCHbox gene in one species to become dynamic and diverge whilst leaving no one "function" unaccounted for, keeping this integrative regulatory network on track (figure 6.1). This divergence may not occur, or may occur in a different gene or with a different function, in another species. Under this model, genes are constrained by a complex "group function", but not the precise mechanism by which this function is carried out. Thus allowing functional reorganisation, and functions to be redistributed between the ETCHbox genes, in different ways in different species (figure 6.1) which may then lead to expression or protein coding changes between closely related species' orthologues. Similarly, this could lead to non-orthologous genes becoming more functionally similar to each other than they are to their orthologues, as discussed in chapter 2. This is my suggested mechanism for how ETCHbox genes can be both so divergent but retain significant overlaps in inferred function (figure 6.1).

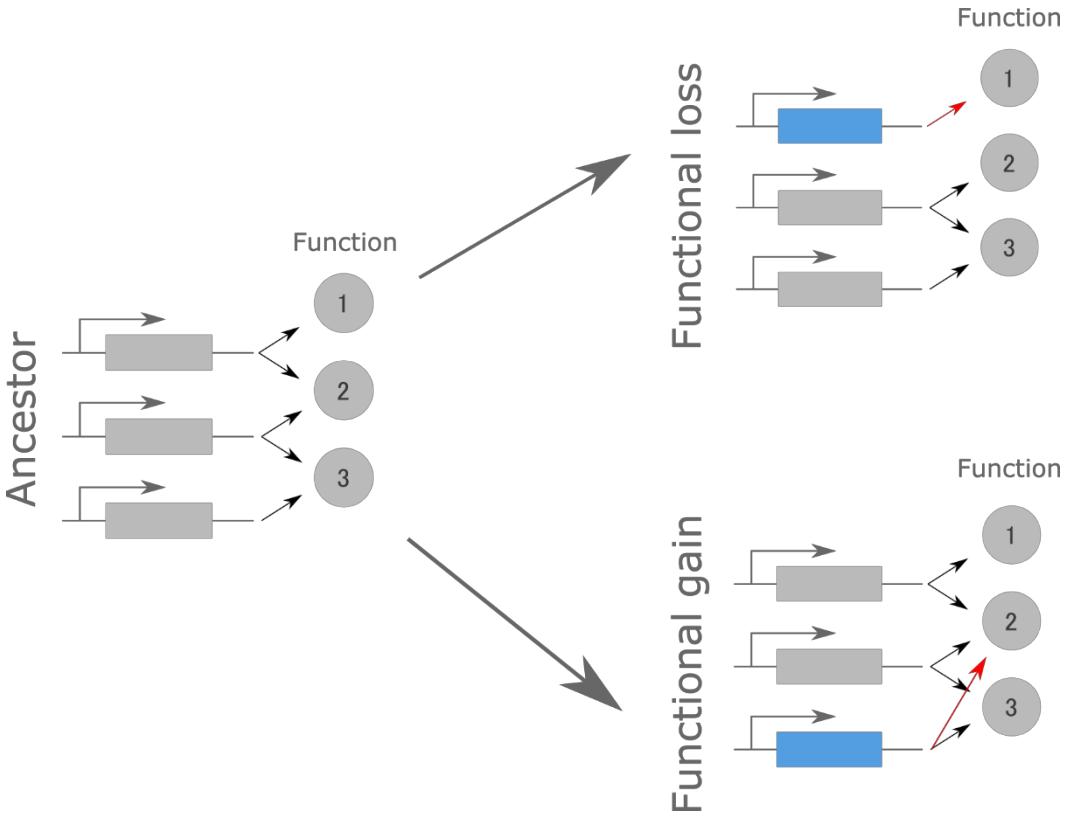


Figure 6.1. Partially overlapping functions of ETCHbox genes could allow for fast evolutionary change. In this hypothesis ETCHbox genes in an ancestor have partially overlapping roles in a number of "group functions" (grey circles). This overlap allows for gene function to be dynamic and shift, through functional gain or loss (red arrows), but these vital functions (grey circles) are all still accounted for and partial overlap can be lost and reinstated. Therefore, involvement in networks important for embryo survival, such as those in the preimplantation embryo, could allow release of evolutionary constraint and facilitate divergence in orthologues between species, whilst retaining overlap in function.

3.1. ETCHbox genes walk hand-in-hand with functional gain

In addition to overlap, I also found repeated evidence for ETCHbox genes acquiring new lineage-specific functions in their evolution. I have shown that the ETCHbox genes evolved from the ancestral *CRX* in marsupials by a mechanism we call "DDC+", combining subfunctionalisation and neofunctionalisation from the classical DDC model (Royall *et al.*, 2019; chapter 3; figure 6.2). The "+" in our mechanism refers to the gaining of additional functions by the ETCHbox genes after

their divergence from *CRX*, potentially influencing the evolution of eutherian-specific embryonic traits such as the ICM.

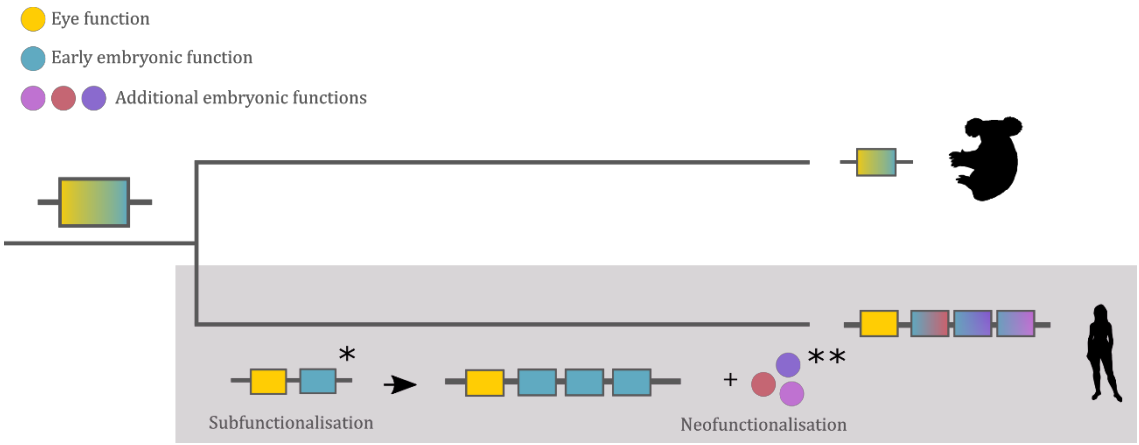


Figure 6.2 The "DDC+" hypothesis and ETCHbox genes. "DDC+" refers to a model of gene evolution whereby following a duplication of a pleiotropic gene, the daughter genes can distribute the two functions by subfunctionalisation, but that this can be combined with neofunctionalisation. In this case, the duplication of *CRX* to give ETCHbox genes in the eutherian lineage (grey box) allowed for the splitting of the eye-specific ancestral function to be retained by eutherian *CRX*, whereas the preimplantation ancestral function was retained by ETCHbox (*). However ETCHbox genes have duplicated further and developed this early embryonic function through gaining function (**), potentially influencing lineage-specific traits at these stages of development. This expansion of one inherited function is a form of neofunctionalisation and hence leads to the realisation of DDC+.

I also demonstrate ETCHbox genes gaining additional functions other than the conserved "group functions" discussed above in diverging eutherian species (figure 6.3). Whilst *ARGFX* and *Crxos* do have high levels of overlap in genes differentially regulated following their ectopic expression (evidence for my hypothesis in figure 6.1), my co-authors and I have shown that *Crxos* has an effect not recapitulated by *ARGFX* on genes differentially expressed as a morula-stage embryo becomes a blastocyst (Royall *et al.*, 2018; chapter 2), combined with a broader temporal expression pattern itself. This is not seen for DEGs following *ARGFX* (Maeso *et al.*, 2016), nor is the less specific expression pattern observed for any other ETCHbox

gene in human or cow. I also find in chapter 5 that *Crxos* exposure before reprogramming can increase the expression of *Mu-ERVL* in the resulting colonies, which is not the case for *ARGFX*. Taken together this suggests that novel functions have been gained by *Crxos* as it diverged.

Similarly, mouse *Oboxa4* is expressed maternally and following ectopic expression regulates a collection of genes which decrease in expression as the embryo heads towards EGA. This maternal expression pattern is also unique to mouse ETCHbox genes and so I suggest it is another additional function gained as Obox genes diverged. It is likely that another non-ETCHbox gene is carrying out this maternal function in other species. Additional descriptions of preimplantation development from a suite of rodents, as well as comparisons between mouse and human embryology, would allow inferences about the effect of the *Tprx2*-Obox transition (chapter 4) on rodent biology. However, it is possible that the shift in timing of EGA to the two-cell embryo in rodents requires an additional maternal factor to prepare for the huge cellular rearrangements of the EGA (Lee *et al.*, 2013; Xue *et al.*, 2013; Liu *et al.*, 2018). I have already discussed that ETCHbox genes are likely to have vital developmental roles, but how can genes with inferred functions at such vital stages of embryogenesis lack evolutionary constraint so much so that they can evolve novel function?

This gaining of function throughout evolution is common for homeobox genes. For example, where Hox genes ancestrally define position along the A-P axis, additional roles in patterning of the limb bud (Zakany and Duboule, 2007), cardiovascular development (Lescroart and Zaffran, 2018) and other functions have been gained. Also, the homeobox genes *Ultrabithorax* (U*bx*) in insects has been

shown to evolve its downstream targets and hence the appendage that develops, whilst retaining its expression domain and high hierarchical position (Weatherbee *et al.*, 1999).

Rather than rebutting it, these additional functions also contribute to my hypothesis for the evolution of the ETCHbox group by distributed robustness described above (figure 6.1). A family of genes which display high levels of variability paired with robustness through partially overlapping functions allow different species to evolve independently (figure 6.1). In this instance, rather than a gene gaining a function and restoring partial overlap, i.e. gaining one of the "group functions" which are proposed shared ETCHbox functions (see above, figure 6.1), a gene gains function which is not currently regulated by ETCHbox in one species (figure 6.3); once this gene has diverged enough to obtain new downstream targets (potentially completely novel, potentially acquired following loss of another unrelated gene but novel for ETCHbox) and/or a new expression domain in the gametes or preimplantation embryo, they would be selected for under independent evolutionary pressures (figure 6.3). This then has a direct relationship to larger evolutionary effects, such as speciation (Sutton and Wilkinson, 1997; Katayama *et al.*, 2018), if hybrid embryo survival is impacted (e.g. if the GRN controlling oocyte-embryo transition is not identical between parent species, using the *Oboxa4* example). I suggest that the partial overlapping nature of ETCHbox function allows a) genes to diverge between species whilst retaining the group function (chapter 2), b) reorganisation of how this group function is carried out by the ETCHbox genes (figure 6.1), and c) the gain of novel functions in a species-specific manner (figure 6.3).

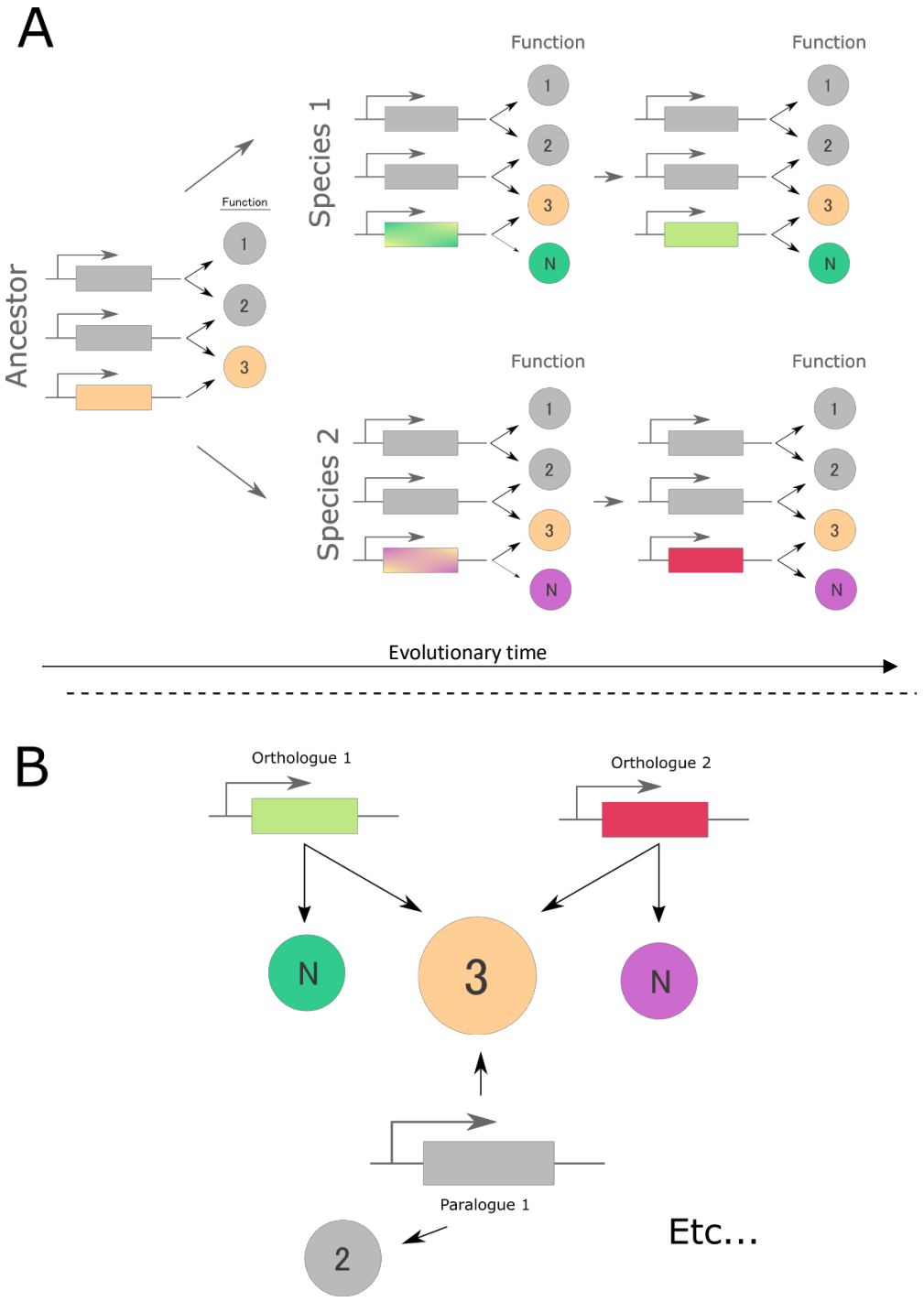


Figure 6.3. Proposed extended mechanism for the evolution of the ETCHbox gene variability and overlapping functions. (A) An ancestral repertoire of ETCHbox genes carry out partially overlapping vital functions. In each species, an ETCHbox gene can gain an additional new (function N, green/purple) function, coinciding with protein sequence change (represented by colour change in the box) as the gene adjusts. This is akin to that of figure 6.1, however here the functional gain is not a "group function". Over time, evolution fine-tunes this sequence to take on both the new function and the ancestral function, resulting in diverging protein sequences between the two species which share a common ancestral function (B). These two orthologues can then be retained under independent evolutionary pressures, but converge on their ancestral function (function 3, yellow). This model also allows for paralogous genes to maintain overlapping function as they diverge (B), as I have described for *Crxos* and *ARGFX* (Chapter 2, Royall *et al.*, 2018).

4. The Pluripotency Network

The final theme to run through my work has been the role of ETCHbox genes in the pluripotency network. I have described the expression patterns of 66 Obox genes and 3 *Crxos* transcripts, all of which are expressed in the totipotent or pluripotent cells of the early embryo. Additionally, I have shown that the ETCHbox "parental" gene *CRX* is expressed in the very early embryo in the fat-tailed dunnart, (a marsupial without ETCHbox genes) suggesting that their involvement in the pluripotency network is ancient and inherited. In addition to expression patterns, I have shown that ectopic expression of both ETCHbox genes and marsupial *CRX* in differentiated cells have specific effects on the transcriptome which I can relate back to the very early, totipotent embryo and so confidently infer very early embryonic function (Royall *et al.*, 2019).

It is becoming common to look to genes utilised in the very early embryo when working to improve the somatic cell reprogramming technology (Hirata *et al.*, 2012; Gao *et al.*, 2017; Oh, 2017; Schiebinger *et al.*, 2017; Wu *et al.*, 2017). The success in this area may be due to epigenetic modifications and physical accessibility of regulatory elements; if the protein product of an early embryonic gene can influence chromatin configuration to make pluripotency genes more accessible, it may increase the likelihood of those genes later becoming expressed. What my results and those of others suggest, is that by exposing a differentiated cell to a gene high up in the pluripotency network, the whole cell may be configured to become more receptive to reprogramming, increasing both the efficiency (Schiebinger *et al.*, 2017; Wu *et al.*, 2017) and the differentiation potential (Choi *et al.*, 2017; Oh, 2017) of resulting iPSCs.

Firstly, I have shown that the ETCHbox gene *Oboxa1* is expressed in pluripotent iPSCs, secondly, I demonstrated that pre-exposure of ETCHbox genes to adult cells prior to reprogramming using Yamanaka factors can increase the efficiency of the reprogramming protocol. Thirdly, I have shown that ETCHbox genes are able to significantly alter the expression of totipotent embryo markers, *Zscan4c* and *Mu-ERVL*, in mouse iPSCs. Taken together, my results repeatedly suggest that the ETCHbox genes of both mouse and human are implicated in the pluripotency network, both in embryonic development and *in vitro*.

There is a slight discrepancy between my results in chapter 5 which suggest that *Crxos* pre-exposure can increase the differentiation potential of stem cells, and chapter 2 which suggest that forced *Crxos* expression pushes the transcriptome out of totipotency and towards a blastocyst-like state. However, the two experiments test different aspects of *Crxos* function. While *Crxos* may not directly increase the expression of totipotency markers, it may have a role elsewhere in the network which acts to prepare the cells for pluripotency (compatible with chapter 2), which leads to more efficient restoration of differentiation potential.

5. Perspectives and future work

My work presented here has contributed to understanding of the ETCHbox genes in multiple directions including function, evolution, and application to other areas of biology. However, questions remain unanswered. Namely, further experimental work is needed to test the repeatability of the effect of exposure of differentiated cells to ETCHbox genes prior to somatic reprogramming. This line of research should also be widened to human stem cells in order to have maximum impact on the biomedical field. Additionally, further characterisation of the cell lines that I have created should test increased differentiation potential.

Secondly, I have described how ETCHbox genes changed in function following their evolution from marsupial *CRX*. An understanding of the sequence of events leading to the six ETCHbox genes in an early eutherian ancestor would contribute greatly to our knowledge of how these unusual genes evolved. For example, did the ETCHbox genes evolve in a stepwise fashion following a single *CRX* duplication, or did they emerge simultaneously from a more complex duplication event? Answering these questions could add detail to my hypothesis of the origin and contribution of overlapping ETCHbox gene functions.

My survey of rodent genomes has shown that our prior understanding of ETCHbox variability may be an underestimation (chapter 4). As more high quality genomes are produced, and more gene expression data of early embryonic stages are generated, we can increase the number and spread of species analysed across the eutherian phylogeny which will increase our understanding further.

Finally, the level of sequence divergence of the ETCHbox genes is astounding considering the importance of the correct development of the preimplantation embryo. How this variation occurs, and the consequence of this on gene function can only be illuminated by further functional work. Analysis of direct downstream targets by chromatin immunoprecipitation followed by sequencing (ChIP-Seq), and information regarding the genomic context that the ETCHbox genes exist in and influence by ATAC-seq should be priority experiments for those building on the work presented here.

My work has aimed to answer the question "How can species evolve such varied repertoires of highly and early expressed homeobox genes, but have relatively similar progression of preimplantation embryogenesis?". I have suggested that ETCHbox gene evolution is based on a mechanism whereby partially overlapping functions facilitate species to adapt to independent evolutionary pressures whilst retaining core functionality (figure 6.3). Consequentially, genes become divergent quickly which could affect hybrid viability between closely related species (figure 6.3). It is possible, however, that the divergence itself is adaptive (Katayama *et al.*, 2018) and overlapping function is a result of incomplete separation and refinement of gene functions. There may be evolutionary pressures driving the ETCHbox genes apart, rather than holding them together. Future work on ETCHbox genes by this group and others will surely unravel this.

7. References

Booth, H. A. F. and Holland, P. W. H. (2007) 'Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line', *Gene*, 387(1–2), pp. 7–14. doi: 10.1016/j.gene.2006.07.034.

Buffry, A. D., Mendes, C. C. and McGregor, A. P. (2016) 'The Functionality and Evolution of Eukaryotic Transcriptional Enhancers', in, pp. 143–206. doi: 10.1016/bs.adgen.2016.08.004.

Bürglin, T. R. (2011) 'Homeodomain Subtypes and Functional Diversity', in Hughes, T. R. (ed.) *A Handbook of Transcription Factors*. Dordrecht: Springer Netherlands, pp. 95–122. doi: 10.1007/978-90-481-9069-0_5.

Carroll, S. B. (2008) 'Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution', *Cell*, 134(1), pp. 25–36. doi: 10.1016/j.cell.2008.06.030.

Chen, L., Tovar-Corona, J. M. and Urrutia, A. O. (2012) 'Alternative Splicing: A Potential Source of Functional Innovation in the Eukaryotic Genome', *International Journal of Evolutionary Biology*, 2012, pp. 1–10. doi: 10.1155/2012/596274.

Choi, Y. J. *et al.* (2017) 'Deficiency of microRNA miR-34a expands cell fate potential in pluripotent stem cells', *Science*, 355(6325). doi: 10.1126/science.aag1927.

Drummond, D. A. and Wilke, C. O. (2008) 'Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution', *Cell*, 134(2), pp. 341–352. doi: 10.1016/j.cell.2008.05.042.

Force, A. *et al.* (1999) 'Preservation of Duplicate Genes by Complementary , Degenerative Mutations', *Genetics*, pp. 1531–1545. Available at: https://www.ncbi.nlm.nih.gov/pubmed/10101175.

Gao, F. *et al.* (2017) 'Identifying Candidate Reprogramming Genes in Mouse Induced Pluripotent Stem Cells', *Stem Cell Reviews and Reports*. Stem Cell Reviews and Reports, 13(4), pp. 532–541. doi: 10.1007/s12015-016-9704-2.

Hirata, T. *et al.* (2012) 'Zscan4 transiently reactivates early embryonic genes during the generation of induced pluripotent stem cells', *Scientific Reports*, 2, pp. 1–11. doi: 10.1038/srep00208.

Katayama, S. *et al.* (2018) 'Phylogenetic and mutational analyses of human LEUTX, a homeobox gene implicated in embryogenesis', *Scientific Reports*, 8(1), p. 17421. doi: 10.1038/s41598-018-35547-5.

Lee, M. T. *et al.* (2013) 'Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition', *Nature*. Nature Publishing Group, 503(7476), pp. 360–364. doi: 10.1038/nature12632.

Lescroart, F. and Zaffran, S. (2018) 'Hox and Tale transcription factors in heart development and disease', *The International Journal of Developmental Biology*, 62(11–12), pp. 837–846. doi: 10.1387/ijdb.180192sz.

Liu, C. *et al.* (2018) 'Post-translational regulation of the maternal-to-zygotic transition', *Cellular and Molecular Life Sciences*. Springer International Publishing, 75(10), pp. 1707–1722. doi: 10.1007/s00018-018-2750-y.

Ludwig, M. Z. *et al.* (2005) 'Functional evolution of a cis-regulatory module', *PLoS Biology*, 3(4), pp. 0588–0598. doi: 10.1371/journal.pbio.0030093.

Maeso, I. *et al.* (2016) 'Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals', *BMC Biology*, 14(1), pp. 1–14. doi: 10.1186/s12915-016-0267-0.

Oh, S. (2017) 'Reprogramming Pluripotent Stem Cell Towards Totipotency', *Field of Northwestern University Interdepartmental Neuroscience Program*, (December, 2017).

Royall, A. H. *et al.* (2018) 'Mouse Obox and Crxos modulate preimplantation transcriptional profiles revealing similarity between paralogous mouse and human homeobox genes', *EvoDevo*. BioMed Central, 9(1), p. 2. doi: 10.1186/s13227-018-0091-4.

Royall, A. H. *et al.* (2019) 'Of eyes and embryos : subfunctionalization of the CRX homeobox gene in mammalian evolution', *Proceedings of the Royal Society B: Biological Sciences*.

Schiebinger, G. *et al.* (2017) 'Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming.', *bioRxiv (preprint)*, p. 191056. doi: 10.1101/191056.

Shaw, P. J. *et al.* (2002) 'Coevolution in bicoid-dependent promoters and the inception of regulatory incompatibilities among species of higher Diptera', *Evolution and Development*, 4(4), pp. 265–277. doi: 10.1046/j.1525-142X.2002.02016.x.

Sutton, K. A. and Wilkinson, M. F. (1997) 'Rapid evolution of a homeodomain: Evidence for positive selection', *Journal of Molecular Evolution*, 45(6), pp. 579–588. doi: 10.1007/PL00006262.

Villar, D. *et al.* (2015) 'Enhancer Evolution across 20 Mammalian Species', *Cell*, 160(3), pp. 554–566. doi: 10.1016/j.cell.2015.01.006.

Wagner, A. (2005) 'Distributed robustness versus redundancy as causes of mutational robustness', *BioEssays*, 27(2), pp. 176–188. doi: 10.1002/bies.20170.

Weatherbee, S. D. *et al.* (1999) 'Uitrabithorax function in butterfly wings and the evolution of insect wing patterns', *Current Biology*, 9(3), pp. 109–115. doi: 10.1016/S0960-9822(99)80064-5.

Wray, G. A. (2007) 'The evolutionary significance of cis-regulatory mutations', *Nature Reviews Genetics*, 8(3), pp. 206–216. doi: 10.1038/nrg2063.

Wu, L. *et al.* (2017) 'Oocyte-Specific Homeobox 1, Obox1, Facilitates Reprogramming by Promoting Mesenchymal-to-Epithelial Transition and Mitigating Cell Hyperproliferation', *Stem Cell Reports*, 9(5), pp. 1692–1705. doi: 10.1016/j.stemcr.2017.09.012.

Xue, Z. *et al.* (2013) 'Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing', *Nature*. Nature Publishing Group, 500(7464), pp. 593–597. doi: 10.1038/nature12364.

Zakany, J. and Duboule, D. (2007) 'The role of Hox genes during vertebrate limb development', *Current Opinion in Genetics and Development*, 17(4), pp. 359–366. doi: 10.1016/j.gde.2007.05.011.

Appendices

A1: List of *Mus musculus* Obox genes including old annotation, new annotation, groups, and whether I have found evidence of expression in the preimplantation embryo.

A2: Portable Document Format (PDF) of published paper that forms Chapter 2.

A3: Portable Document Format (PDF) of published paper that forms Chapter 3.

A1 *Mus musculus* Obox gene new annotation, pseudogene status, expression analysis, and groupings.

Alias	New nomenclature	Pseudogene?	Expressed in preimplantation Embryos?	Group
Gm5585/Obox8	*Oboxc*	N	Y	OboxC
Obox6	*Oboxa1*	N	Y	OboxA
Obox5	*Oboxa2*	N	N	OboxA
Obox3	*Oboxa3*	N	Y	OboxA
Obox1	*Oboxa4*	N	Y	OboxA
Gm5889/ obox3 ps6	*Oboxa5*	N	Y	OboxA
Obox2	*Oboxa6*	Y	N	OboxA
Gm8053/ obox3 ps5	*Oboxa7*	N	Y	OboxA
Gm8072/ obox3 ps4	*Oboxa8*	N	Y	OboxA
Gm8064/ obox3 ps3	*Oboxa9*	N	N	OboxA
Gm8040/obox3 ps2	*Oboxa10*	N	Y	OboxA
Gm8057/ obox3 ps1	*Oboxa11*	N	N	OboxA
Gm4745	*Oboxa12*	N	Y	OboxA
Gm18533 partial/obox3 ps7	*Oboxa13*	Y	N	OboxA
Gm4830/ obox3 ps8	--	--		Not analysed
--	*Oboxb1 partial*	Y	N	OboxB
Gm7235	*Oboxb2*	N	Y	OboxB
Gm20622/obox4 ps34	*Oboxb3*	Y	*N*	OboxB
Obox4 ps32	*Oboxb4*	Y	*N*	OboxB
--	*Oboxb5 partial*	Y	*N*	OboxB
--	*Oboxb6 partial*	Y	*N*	OboxB
--	*Oboxb7 partial*	Y	*N*	OboxB
Obox4 ps25	*Oboxb8 partial*	Y	*N*	OboxB
Gm20624 Obox4 ps24	*Oboxb9*	Y	*N*	OboxB
--	*Oboxb10*	Y	*N*	OboxB
Obox4 ps10	*Oboxb11*	Y	*N*	OboxB
--	*Oboxb12 partial*	Y	*N*	OboxB
--	*Oboxb13*	Y	*N*	OboxB
Obox4 ps7	*Oboxb14*	Y	*N*	OboxB
Gm27258/obox4 ps39	*Oboxb15 partial*	Y	*N*	OboxB
--	*Oboxb16*	Y	*N*	OboxB
Obox4 ps5	*Oboxb17*	Y	*N*	OboxB
--	*Oboxb18*	Y	*N*	OboxB
Obox4 ps4	*Oboxb19*	Y	*N*	OboxB
--	*Oboxb20 partial*	Y	*N*	OboxB
--	*Oboxb21*	Y	*N*	OboxB
Gm207180/Obox4 ps3	*Oboxb22*	Y	*N*	OboxB
--	*Oboxb23 partial*	Y	*N*	OboxB
Gm27172/Obox4 ps2	*Oboxb24 partial*	Y	*N*	OboxB
Gm27171/Obox4 ps1	*Oboxb25 partial*	Y	*N*	OboxB
--	*Oboxb26*	Y	*N*	OboxB
Obox4	*Obox4*	--	*N*	Not annotated
Gm4023/ obox4 ps33	*Oboxd1*	N	*N*	OboxD
Gm4018/obox4 ps31	*Oboxd2*	N	*N*	OboxD
Gm4016/obox4 ps30	*Oboxd3 partial*	Y	*N*	OboxD
Gm4014/ obox4 ps29	*Oboxd4*	Y	*N*	OboxD
Gm10678/obox4 ps28	*Oboxd5*	N	*N*	OboxD
Gm10679/obox4 ps27	*Oboxd6*	N	*N*	OboxD
Gm20623/ obox4 ps26	*Oboxd7 partial*	Y	*N*	OboxD
Gm4897/obox4 ps23	*Oboxd8*	N	*N*	OboxD
Gm4003/obox4 ps22	*Oboxd9*	N	*N*	OboxD
Gm4001/obox4 ps21	*Oboxd10*	N	*Y*	OboxD
Gm3998/obox4 ps20	*Oboxd11*	N	*N*	OboxD
Gm3996/obox4 ps19	*Oboxd12*	N	*N*	OboxD
Gm3994/obox4 ps18	*Oboxd13*	N	*N*	OboxD
Gm3989/obox4 ps17	*Oboxd14*	N	*N*	OboxD
Gm3987/obox4 ps16	*Oboxd15*	N	*N*	OboxD
Gm3684/obox4 ps15	*Oboxd16*	N	*N*	OboxD
Gm3981/obox4 ps14	*Oboxd17*	N	*N*	OboxD
Gm3980/obox4 ps13	*Oboxd18*	N	*N*	OboxD
Gm3977/obox4 ps12	*Oboxd19*	N	*N*	OboxD
Gm3968/obox4 ps11	*Oboxd20 partial*	N	*N*	OboxD
Gm3965/obox4 ps9	*Oboxd21*	N	*N*	OboxD
Gm3959/obox4 ps8	*Oboxd22 partial*	Y	*N*	OboxD
Gm20710/ obox4 ps6	*Oboxd23 partial*	N	*N*	OboxD
Gm20709/obox4 ps38	*Oboxd24 partial*	Y	*N*	OboxD
Gm20664/obox4 ps37	*Oboxd25 partial*	Y	*N*	OboxD
Gm20665/obox4 ps36	*Oboxd26 partial*	Y	*N*	OboxD
Gm7235/Obox4 ps35	--	--		Not annotated

Royall *et al. EvoDevo (2018) 9:2*
https://doi.org/10.1186/s13227-018-0091-4

EvoDevo

RESEARCH

Open Access

Mouse *Obox* and *Crxos* modulate preimplantation transcriptional profiles revealing similarity between paralogous mouse and human homeobox genes

Amy H. Royall[1], Ignacio Maeso[2], Thomas L. Dunwell[1*] and Peter W. H. Holland[1*]

Abstract

Background: ETCHbox genes are eutherian-specific homeobox genes expressed during preimplantation development at a time when the first cell lineage decisions are being made. The mouse has an unusual repertoire of ETCHbox genes with several gene families lost in evolution and the remaining two, *Crxos* and *Obox*, greatly divergent in sequence and number. Each has undergone duplication to give a double homeodomain *Crxos* locus and a large cluster of over 60 *Obox* loci. The gene content differences between species raise important questions about how evolution can tolerate loss of genes implicated in key developmental events.

Results: We find that *Crxos* internal duplication occurred in the mouse lineage, while *Obox* duplication was stepwise, generating subgroups with distinct sequence and expression. Ectopic expression of three *Obox* genes and a *Crxos* transcript in primary mouse embryonic cells followed by transcriptome sequencing allowed investigation into their functional roles. We find distinct transcriptomic influences for different *Obox* subgroups and *Crxos*, including modulation of genes related to zygotic genome activation and preparation for blastocyst formation. Comparison with similar experiments performed using human homeobox genes reveals striking overlap between genes downstream of mouse *Crxos* and genes downstream of human *ARGFX*.

Conclusions: Mouse *Crxos* and human *ARGFX* homeobox genes are paralogous rather than orthologous, yet they have evolved to regulate a common set of genes. This suggests there was compensation of function alongside gene loss through co-option of a different locus. Functional compensation by non-orthologous genes with dissimilar sequences is unusual but may indicate underlying distributed robustness. Compensation may be driven by the strong evolutionary pressure for successful early embryo development.

Keywords: Gene duplication, Gene loss, Homeodomain, PRD class, *ARGFX*, Transcription factor, Compensation, Blastocyst

Background

Members of the homeobox superclass are widespread across eukaryotes, and their encoded proteins function mainly as transcription factors [1]. The proteins contain a DNA-binding domain known as the homeodomain which is highly variable at the primary sequence level. In animals, homeobox genes can be assigned to eleven classes, containing ~ 100 gene families in humans [1]. The largest homeobox gene class in animals is the ANTP class, with the PRD class being second largest. Some homeobox genes are highly conserved throughout animals; others, such as the mammalian X-linked *Rhox* cluster, have been evolving rapidly through duplication and sequence divergence [2]. Some homeobox gene families

*Correspondence: thomasdunwell@gmail.com; peter.holland@zoo.ox.ac.uk
[1] Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK
Full list of author information is available at the end of the article

BioMed Central

© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

have been lost from certain evolutionary lineages secondarily [3, 4].

Ten years ago, several PRD class homeobox genes with highly divergent sequences were identified in the human genome (*TPRX1*, *TPRX2*, *DPRX*, *LEUTX* and *ARGFX*); it is now known that these are specific to eutherian mammals and arose through duplication of an Otx family gene *Crx* [4–7]. Each of these genes has expression limited to the germ line and early embryos in humans, and collectively they have been named Eutherian Totipotent Cell Homeobox (ETCHbox) genes [5, 7, 8]. Ectopic expression of several human ETCHbox genes revealed roles in transcriptional regulation of genes that have a peak of expression in the eight-cell embryo and morula stage, around the time of embryo compaction and following embryonic genome activation in humans [7, 9].

Until recently, these genes were thought to be lost from the mouse genome due to the absence of genes with highly similar homeobox sequences to human ETCHbox genes. The genomic regions syntenic to *TPRX1* and *TPRX2*, flanking the *Crx* gene on mouse chromosome 7, contain homeobox genes, but these are so distinct in sequence from the human genes that it was hypothesised they had arisen through independent tandem gene duplications in mouse ancestry [4, 10]. Each of these mouse genes, *Crx*-opposite strand (*Crxos*) [11] and oocyte-specific homeobox (*Obox*), have also undergone additional tandem duplication events not evident at the human locus. However, with increased sampling of rodents, the long phylogenetic branch lengths leading to *Crxos* and *Obox* could be broken and the mouse genes were shown to be highly divergent orthologues of *TPRX1* and *TPRX2*, respectively [7]. The mouse genome has no orthologous homeobox genes in loci syntenic to any other ETCHbox gene (Fig. 1; [7]). This unique genome organisation is quite different from human and most other mammals and raises important questions about the function of these genes in the mouse. For example, does extensive sequence divergence of the mouse genes imply they have acquired distinct functions from their human orthologues? What is the functional consequence of secondary duplications experienced by these genes in mice? With loss of *Dprx*, *Leutx*, *Pargfx* and *Argfx* in mice, have any compensatory mechanisms evolved such that *Crxos* or *Obox* genes take over the function of the lost genes?

Previous studies of mouse *Crxos* revealed that the locus gives rise to three distinct transcripts: a long transcript encompassing the two duplicate genes (and two homeoboxes) and one transcript from each duplicate. The complete *Crxos* locus has six exons with homeobox sequences spanning exons 2–3, and 5–6. The long transcript consists of exons 1, 2, 3 (partially), 5 and 6, with the shorter transcripts containing either exons 1, 2 and 3, or exons 4, 5 and 6 [12, 13]. Experiments involving ectopic expression in embryonic stem cells followed by quantification of candidate downstream targets have suggested that *Crxos* is involved in cell pluripotency with each shorter variant inhibiting differentiation markers [13–15]. Furthermore, an RNAi screen has identified *Crxos* as important for correct formation of the inner cell mass in the blastocyst and for hatching and outgrowth [16]. Another study has suggested that the long isoform is involved in specifying the primitive endoderm lineage [17].

The *Obox* genes were initially described as transcripts in unfertilised mouse eggs [18], and expression has now been reported throughout early murine development [19, 20]. Secondary duplication of *Obox* genes generated multiple loci. In a recent genome assembly (GRCm38/mm10), an array of five *Obox* loci is annotated, *Obox1*, *Obox2*, *Obox3*, *Obox5* and *Obox6*, with an additional locus *Obox7* given 'provisional' status. Previously, *Obox4* was annotated, but this is now listed as a 'partial' and unplaced annotation. Cheng et al. [19] showed that *Obox1* and *Obox2* have high sequence similarity and suggested they have the same expression pattern in the preimplantation embryo; they also found that *Obox1/2*, *Obox3* and *Obox5* are most highly expressed in the one-cell stage embryo with expression decreasing until no expression is detected in the morula. The *Obox6* locus, in contrast, has elevated expression between the two-cell and morula stages. When *Obox6* was knocked out in

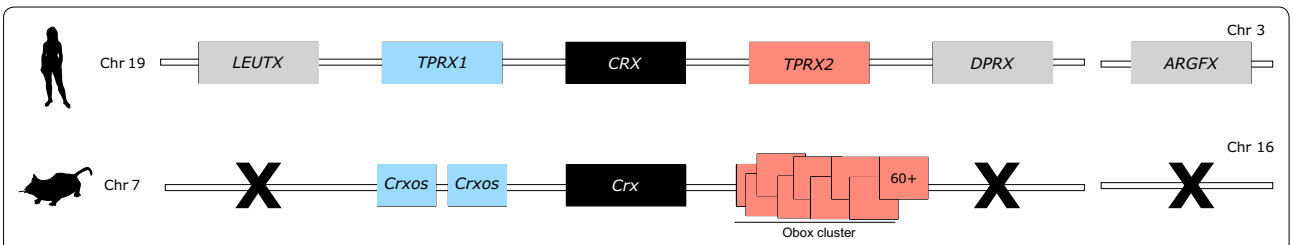


Fig. 1 ETCHbox genes in human and mouse. The ETCHbox group of homeobox genes has six members: *Leutx*, *Tprx1*, *Tprx2*, *Dprx*, *Argfx* and *Pargfx*. In human, *Pargfx* has been lost in evolution. In mouse, only *Tprx1* and *Tprx2* orthologues remain and are referred to as *Crxos* and *Obox*, respectively; each has undergone sequence divergence and gene duplication

mice by homologous recombination, offspring exhibited normal development and were fertile [19]. More recent genome exploration has revealed that the *Obox* duplications were far more extensive than previously recognised and have generated over 60 distinct loci [7]. Not all loci have complete homeobox sequences, but it is clear that the diversity of *Obox* genes had been vastly underestimated. This raises new questions regarding the function of the *Obox* cluster in the light of these extensive duplications.

In this study, we use ectopic expression and transcriptomic analysis to investigate the function of mouse ETCHbox genes (*Crxos* and *Obox*). Our goals were to investigate whether these divergent genes have similar or distinct developmental functions to their human homologues, and whether *Crxos* or *Obox* genes have taken over roles associated with ETCHbox genes secondarily lost in mouse evolution. We find that ectopic expression of *Crxos* or *Obox* genes in cultured mouse embryonic fibroblasts induces large transcriptomic changes which can be related to notable events in the preimplantation embryo. We also argue that *Crxos* functions have evolved to compensate partially for loss of the *Argfx* homeobox gene.

Results

Sequence and expression diversity within expanded mouse ETCHbox clusters

The orthologues of both *TPRX1* (*Crxos*) and *TPRX2* (*Obox*) genes are duplicated in mouse. We previously identified 67 *Obox* loci (including pseudogenes) clustered on chromosome 7 in the region syntenic to the *TPRX2* locus in humans [7]. To investigate the pathway of duplication and search for groupings of highly similar loci, we conducted phylogenetic analyses using *Obox* nucleotide sequences (Additional files 1A and 2). One locus previously identified was found to not contain a homeodomain or high sequence similarity to annotated *Obox* genes. This sequence was therefore disregarded resulting in a total of 66 loci. Phylogenetic analysis revealed three main categories, named here OboxA, OboxB and OboxD (Additional file 1A) containing 13, 26 and 26 loci, respectively, arranged in an interspersed manner (Additional file 1B). These include putatively functional loci and pseudogenes. Since there are many more loci than previously named, we propose a new Obox nomenclature system based on DNA sequence (see Additional file 1C). We identify one additional locus that does not fit within these main groups, *Oboxc*.

Of these 66 *Obox* loci, we suggest that 28 have potential to be translated into protein sequences with a full homeodomain (Additional file 2). These deduced protein sequences were then used in a second phylogenetic analysis which indicated three subgroups within the OboxA group (Table 1; Fig. 2b; Additional file 1D). Analysis of RNA sequencing data from preimplantation mouse embryos revealed that temporal expression is similar within *Obox* groups or subgroups, but can be subtly different between them (Fig. 2b). Expression of the vast majority loci within the OboxD group was not detected in the early embryo, with the exception of *Oboxd10* which shows low-level expression (FPKM 4.09) at the two-cell stage (Table 1). Overall, we identified four expression profiles: oocyte to two cell (OboxAb), two cell to four cell (OboxAa, OboxB), two cell to eight cell (OboxC) and two cell to blastocyst (OboxAc) (Fig. 2b). The observed expression patterns for *Oboxa1*, *Oboxa2*, *Oboxa3*, *Oboxa4* and *Oboxa6* are consistent with previous studies [18–20]. Based on sequence and expression similarity, we hypothesise there is likely to be functional redundancy within, but not usually between, *Obox* groups and subgroups.

The murine gene *Crxos*, orthologous to human *TPRX1*, has duplicated in tandem with the locus generating three distinct transcripts (Fig. 2a). To date the duplication, we used a conserved domain search to examine the genomic region between *Crx* and *Sepw1* loci (the location of *TPRX1*/*Crxos*) for homeobox sequences. In cow (*Bos taurus*), guinea pig (*Cavia porcellus*), naked mole rat (*Heterocephalus glaber*) and brown rat (*Rattus norvegicus*), only one homeobox was found (Fig. 2a, Additional file 1E), although

Table 1 Groupings of mouse *Obox* genes with corresponding expression patterns

Human orthologue	Group	Subgroup	Members with ORF and HD	Members expressed in embryo	Example	Expression
TPRX2	OboxA	OboxAa	5	5	*Oboxa7*	Two cell to four cell
		OboxAb	4	4	*Oboxa4*	Oocyte to two cell
		OboxAc	1	1	*Oboxa1*	Two cell to morula
	OboxB	OboxB	1	1	*Oboxb2*	Two cell to four cell
	OboxC	OboxC	1	1	*Oboxc*	Two cell to eight cell
	OboxD	OboxD	16	1*	*Oboxd10*	Two cell
TPRX1	Crxos	NA	3	3	*Crxos*	Two cell to blastocyst

* Exact number uncertain because of 100% sequence similarity

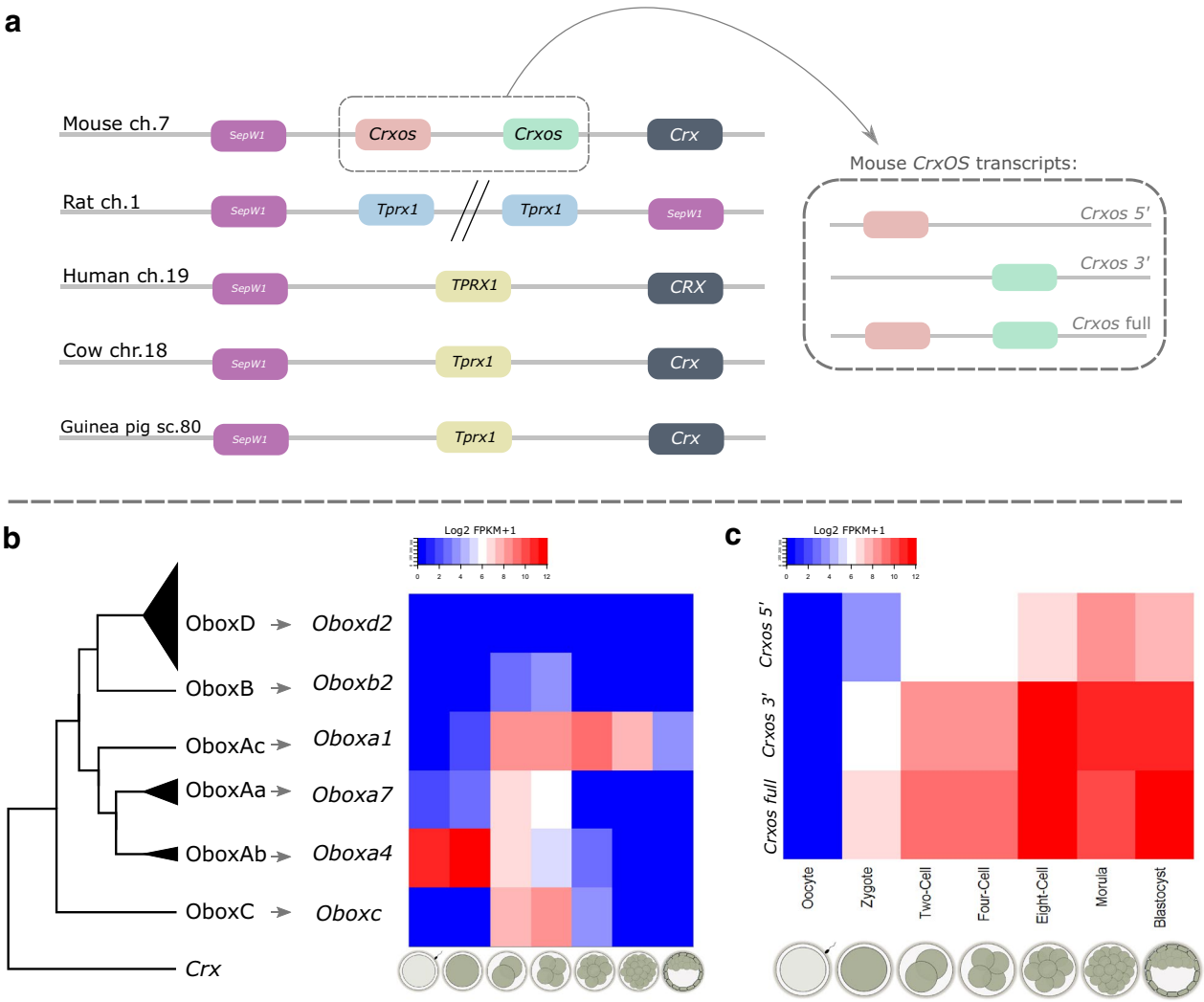


Fig. 2 ETCHbox organisation and expression in preimplantation development: **a** The *Crxos* locus has undergone a mouse-specific duplication and generates three transcripts: two with a single homeobox and a composite double-homeobox transcript. *Crxos* is the *Tprx1* orthologue although originally named as a distinct gene. **b** Analysis of protein and nucleotide sequences of 66 *Obox* genes in mouse identified four groupings; sequences and expression profiles enable OboxA to be split into three subgroups. **c** The three *Crxos* transcripts have the same preimplantation temporal expression pattern, with the 3′ single-homeobox transcript generally having higher RNA levels

R. norvegicus has an apparently independent duplication of the region containing extra copies of both *Crxos* and *Sepw1*. In contrast, using TBLASTN in the Algerian mouse (*Mus spretus*, Jackson laboratory strain #001146), we found two homeobox sequences and regions homologous to all *M. musculus* exons (Additional file 1E and F). These data reveal that the internal duplication to generate a double-homeobox *Crxos* gene occurred in the ancestry of the *Mus* genus after the split of the mouse and rat lineages.

Analysis of RNA sequencing data from mouse preimplantation embryos revealed that the 3′ single-homeobox *Crxos* transcript has higher expression than the 5′ single-homeobox transcript, although the temporal profile is the same (Fig. 2c). The 3′ *Crxos* transcript has also been previously reported to encode sequences essential for nuclear localisation of the protein [21].

Ectopic expression of *Crxos* and *Obox* genes induces transcriptomic changes

As homeobox genes encode transcription factors, we asked whether ectopic expression of mouse ETCHbox genes could induce transcriptional changes and how such changes are related to possible in vivo roles. Due to the complex genomic organisation, we used the above evolutionary analyses to inform the choice of genes. We selected three *Obox* genes with different expression patterns and from different subgroups to test the hypothesis that genes from different subgroups have distinct

activities. We also selected the 3′ *Crxos* single homeodomain transcript as the most highly expressed splice variant of *Crxos;* this will now be referred to as '*Crxos*' throughout.

We expressed ectopically *Oboxa1* (subgroup OboxAc, *Obox6* in earlier nomenclature), *Oboxa4* (subgroup OboxAb, formerly *Obox1*), *Oboxa7* (subgroup OboxAa) or *Crxos* under a constitutive promoter in primary mouse embryonic fibroblasts and assayed transcriptome-wide effects through RNA sequencing. In each case, we determined catalogues of genes that were up- or down-regulated 48 h after ectopic expression (Additional file 3). These catalogues could include direct and indirect targets. To identify embryonic processes potentially influenced by *Crxos*, *Oboxa1*, *Oboxa4*, and *Oboxa7,* we examined whether each catalogue of up- or down-regulated genes had overlap with defined 'temporal expression profiles' (these are sets of genes grouped on the basis of expression pattern in mouse preimplantation development; shown in Additional files 1G and 4). We also used gene ontology (GO terms) to provide further functional information on groups of genes. The numbers of genes affected significantly by each treatment, and the overlaps between experiments are shown in Additional file 1H.

Crxos up- and down-regulated profiles

The catalogue of genes down-regulated following *Crxos* ectopic expression is enriched in genes from a single temporal expression profile, number 5 (Fisher's test $p = 0.0007$; Fig. 3a, Additional file 5). Profile 5 contains genes with high levels of mRNA in the oocyte, which decrease as development proceeds (Fig. 3a). Genes up-regulated by *Crxos* ectopic expression are enriched for genes in profile 59 (Fisher's test $p = 0.02$, Additional file 5) composed of genes with a sharp peak of expression in the two-cell embryo, then lower expression from four cell to morula, before a second, larger expression peak in blastocyst. There is also enrichment for profile 216, but this profile is not consistent with normal *Crxos* expression so may be off-target. Given that expression of *Crxos* begins in the two-cell embryo and extends to the blastocyst, we suggest that the second, larger peak of expression in profile 59 includes in vivo downstream targets of *Crxos* (Fig. 3a). Together, these results suggest that when *Crxos* transcription is initiated in the two-cell stage, it serves a role in preparing the embryo for blastocyst formation.

Obox up- and down-regulated profiles

The catalogue of genes up-regulated by ectopic expression of *Oboxa1*, *Oboxa4* or *Oboxa7* showed extensive overlap. Each of these *Obox* genes elicited higher expression of genes found within temporal expression profile 101 (*Oboxa1* $p = 0.007$, *Oboxa4* $p = 0.005$, *Oboxa7* $p = 0.001$; Fig. 3b, Additional file 5). In mouse development, genes in profile 101 are not expressed during the earliest stages of preimplantation development but show a sharp pulse of expression at the eight-cell stage. These genes are likely to have roles before the earliest cell fate decisions. The finding that three *Obox* genes, but not *Crxos*, elicit increased expression of similar sets of genes may be a consequence of the three closely related Obox proteins activating target genes through recognition of the same enhancer motifs [20]. Thus, after ectopic expression, an *Obox* gene may up-regulate targets usually regulated by a different *Obox* gene. Considering the expression profiles of *Oboxa1*, *Oboxa4* and *Oboxa7*, we suggest that genes within profile 101 are possible in vivo downstream targets of *Oboxa1* (OboxAc subclass) or *Oboxa7* (OboxAa subclass).

Considered across all temporal profiles, the three *Obox* genes examined up-regulate a common set of 343 genes and down-regulate a common set of 268 genes (Fig. 4). GO analysis reveals extracellular matrix ($p = 4.9 \times 10^{-33}$) and focal adhesion ($p = 7.2 \times 10^{-8}$) as enriched functions among the commonly up-regulated genes, and DNA replication ($p = 1.7 \times 10^{-5}$) enriched among the down-regulated genes (Additional file 1I).

The comparison of all downstream genes also reveals that ectopic expression of *Oboxa4* has the most distinct effects of the three *Obox* genes tested, affecting more than 600 genes not significantly affected by the other *Obox* genes (Fig. 4). The distinctiveness of *Oboxa4* is also evident at the level of temporal profiles. In addition to the shared profile 101, genes up-regulated following ectopic expression of *Oboxa4* were enriched for three further temporal profiles: 216, 219 and 226 (Fig. 3C, Additional file 5). The first two of these comprise genes with in vivo expression peaks in the zygote and two-cell embryo. Since *Oboxa4* is predominantly expressed as a maternal transcript, these profiles are consistent with being downstream in vivo direct or indirect targets (Fig. 3c). Genes down-regulated following *Oboxa4* ectopic expression are significantly enriched for six expression profiles, each of which has an mRNA peak after zygotic genome activation (ZGA; [22]): profiles 202, 84, 79, 124, 149, 129; collective Fisher's test $p = 2.7 \times 10^{-13}$ (Fig. 3b, Additional file 5). These profiles, therefore, are consistent with being negatively regulated downstream in vivo targets, with maternal *Oboxa4* suppressing expression until after ZGA. A gene ontology (GO) analysis of the regulated genes in these profiles shows enrichment for ribosome biogenesis functions ($p < 0.001$). Profile 219 is also enriched in the genes up-regulated downstream of *Oboxa1*, although considering the temporal expression of *Oboxa1* in the embryo this is more likely to be an in vivo target of *Oboxa4* (Fig. 3c).

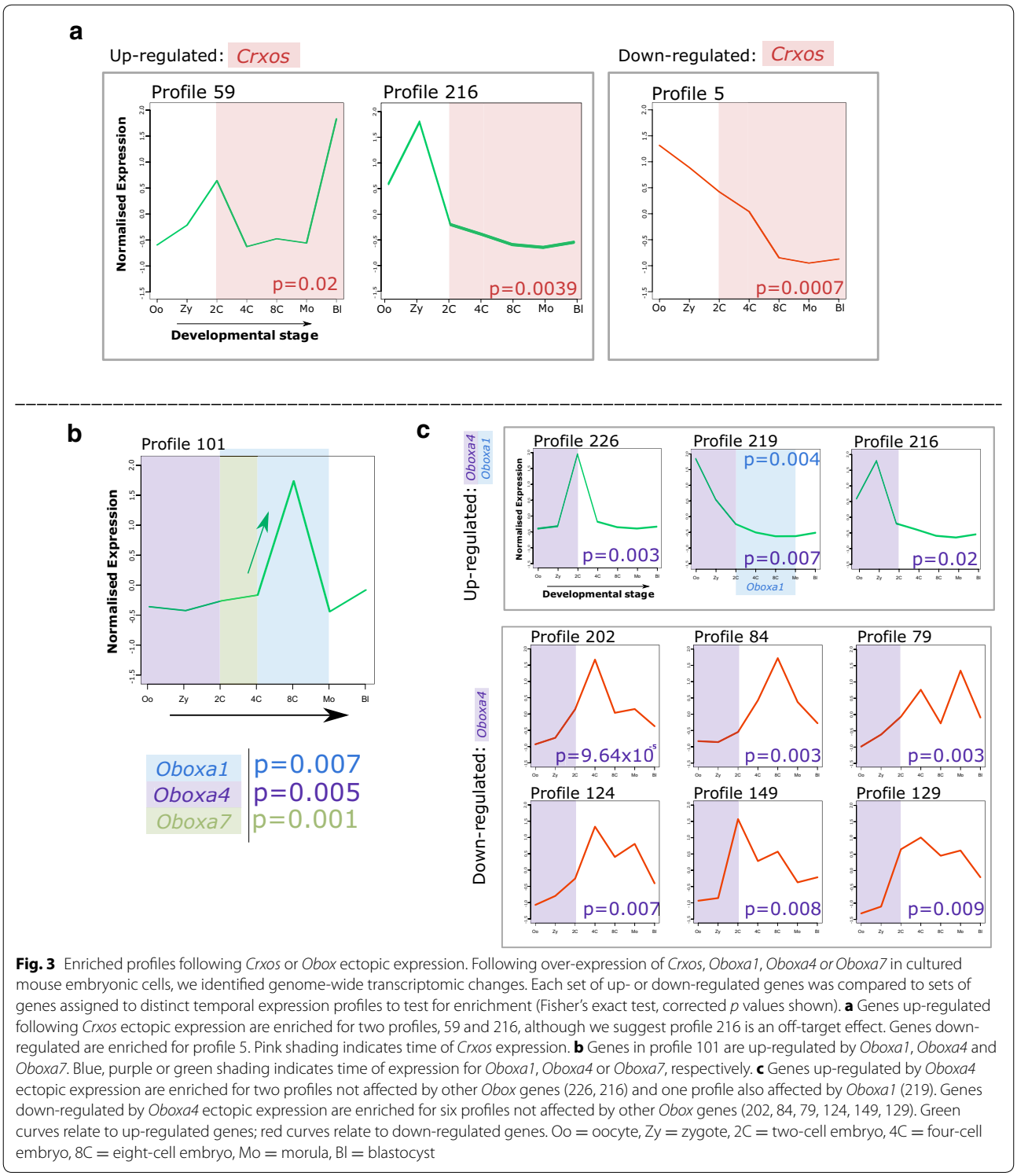


Fig. 3 Enriched profiles following *Crxos* or *Obox* ectopic expression. Following over-expression of *Crxos*, *Oboxa1*, *Oboxa4 or Oboxa7* in cultured mouse embryonic cells, we identified genome-wide transcriptomic changes. Each set of up- or down-regulated genes was compared to sets of genes assigned to distinct temporal expression profiles to test for enrichment (Fisher's exact test, corrected *p* values shown). **a** Genes up-regulated following *Crxos* ectopic expression are enriched for two profiles, 59 and 216, although we suggest profile 216 is an off-target effect. Genes down-regulated are enriched for profile 5. Pink shading indicates time of *Crxos* expression. **b** Genes in profile 101 are up-regulated by *Oboxa1*, *Oboxa4* and *Oboxa7*. Blue, purple or green shading indicates time of expression for *Oboxa1*, *Oboxa4* or *Oboxa7*, respectively. **c** Genes up-regulated by *Oboxa4* ectopic expression are enriched for two profiles not affected by other *Obox* genes (226, 216) and one profile also affected by *Oboxa1* (219). Genes down-regulated by *Oboxa4* ectopic expression are enriched for six profiles not affected by other *Obox* genes (202, 84, 79, 124, 149, 129). Green curves relate to up-regulated genes; red curves relate to down-regulated genes. Oo = oocyte, Zy = zygote, 2C = two-cell embryo, 4C = four-cell embryo, 8C = eight-cell embryo, Mo = morula, Bl = blastocyst

Comparison between human and mouse ETCHbox gene functions

The evolutionary loss of *Argfx*, *Leutx* and *Dprx* ETCHbox genes in mouse, and duplication of the remaining *Tprx1* and *Tprx2* genes, raises questions about functional similarities and differences between mouse and human. We compared the sets of genes significantly up- and down-regulated after ectopic expression of human *TPRX1*,

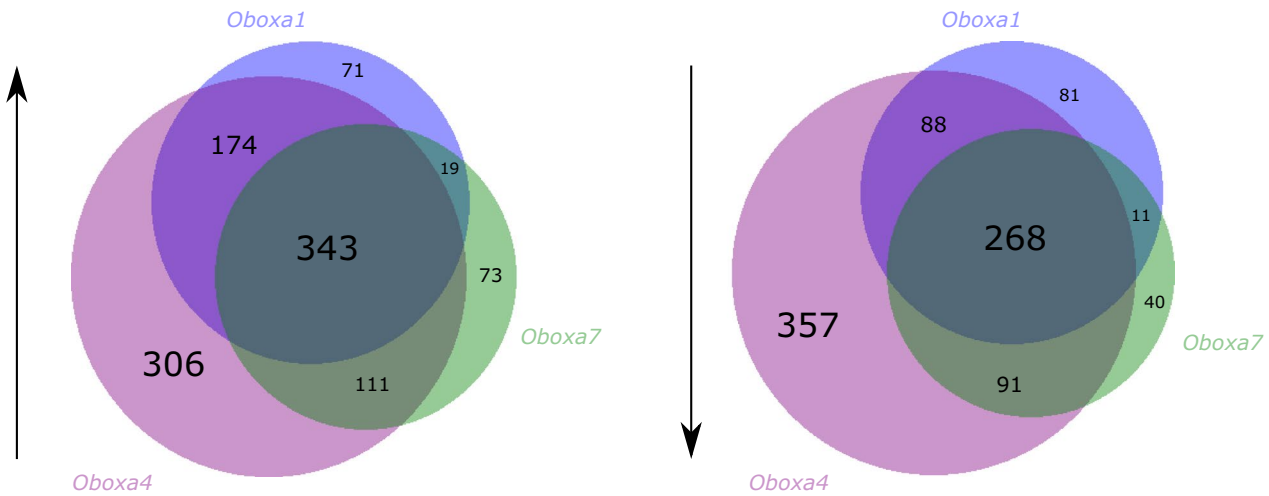


Fig. 4 Overlap between genes downstream of *Obox* genes. Comparison of genes up- or down-regulated following ectopic expression of *Obox* genes; ectopic expression of the maternally expressed *Oboxa4* gene affects expression of additional downstream genes not affected by *Oboxa1* or *Oboxa7* expression

ARGFX, *DPRX* and *LEUTX* [7] with each of the gene sets up- or down-regulated by mouse *Crxos*, *Oboxa1*, *Oboxa4* and *Oboxa7*. Two of these comparisons revealed striking overlap (Fig. 5a; Additional files 6 and 7). We found that for 96 of the genes up-regulated by *ARGFX* in human cells, their mouse orthologues were also up-regulated by *Crxos* in mouse cells (Fig. 5b; Fisher's test $p = 3 \times 10^{-52}$; Additional file 7). Similarly, the set of genes down-regulated by *ARGFX* in human cells showed significant similarity to the set of genes down-regulated by *Crxos* in mouse cells (Fig. 5b; 125 one-to-one orthologues, $p = 2.4 \times 10^{-102}$; Additional file 7). Of these, 98% of the jointly down-regulated orthologues are expressed in the mouse preimplantation embryo, as are 50% of the jointly up-regulated genes. These data suggest close similarity of function between human *ARGFX* and mouse *Crxos*, despite these being non-orthologous proteins. The direct orthologue of mouse *Crxos* is human *TPRX1*, not *ARGFX*; these orthologous genes seem to have contrasting functions following ectopic expression since many genes down-regulated by *Crxos* over-expression are up-regulated by *TPRX1* ($p = 1.58^{-36}$; Additional file 6). There are also additional significant overlaps observed between genes (Fig. 5a), suggesting further overlapping functions of human and mouse ETCHbox genes.

Targets with strong response are associated with embryonic milestones

The analysis of function using temporal gene profiles pays equal attention to genes with low and high expression and does not distinguish between mildly or strongly up- and down-regulated targets. We therefore asked which target genes were most strongly affected by ectopic expression of mouse *Crxos* and *Obox* genes (highest fold change up-regulated in Table 2; highest fold change down-regulated Additional file 1J). Three of the 10 most strongly up-regulated genes following *Crxos* ectopic expression have functions related to the extracellular matrix (ECM). Similarly, when ordered by fold change, three of the genes up-regulated most strongly following *Oboxa1* and *Oboxa4* ectopic expression are ECM components or function as mediators of ECM communication: *Col8a2 (Oboxa1* and *Oboxa4)*, *Eln (Oboxa1* and *Oboxa4)*, *Egfl6 (Oboxa1)* and *Aoc3 (Oboxa4)*.

The most strongly down-regulated genes have functions in processes distinct from the up-regulated genes (Additional file 1J). Among the top 10 down-regulated genes following *Oboxa1* ectopic expression is *Sox2*, a transcription factor involved in specification of the inner cell mass of the blastocyst and deployed in iPS cell production as one of the Yamanaka factors [23]. Similarly, a gene strongly down-regulated after *Crxos* ectopic expression is implicated in trophectoderm specification and development, *Slco2a1* ($\log_2$ fold change of -0.94) [24, 25]. The effect of ectopic ETCHbox expression on other genes with known roles in preimplantation development is given in Additional file 1K.

Discussion

Mammalian preimplantation development encompasses several processes common to a wide range of species, such as generation of a hollow blastocyst and implantation into maternal endometrial tissue. Mice have long been used judiciously as models for human

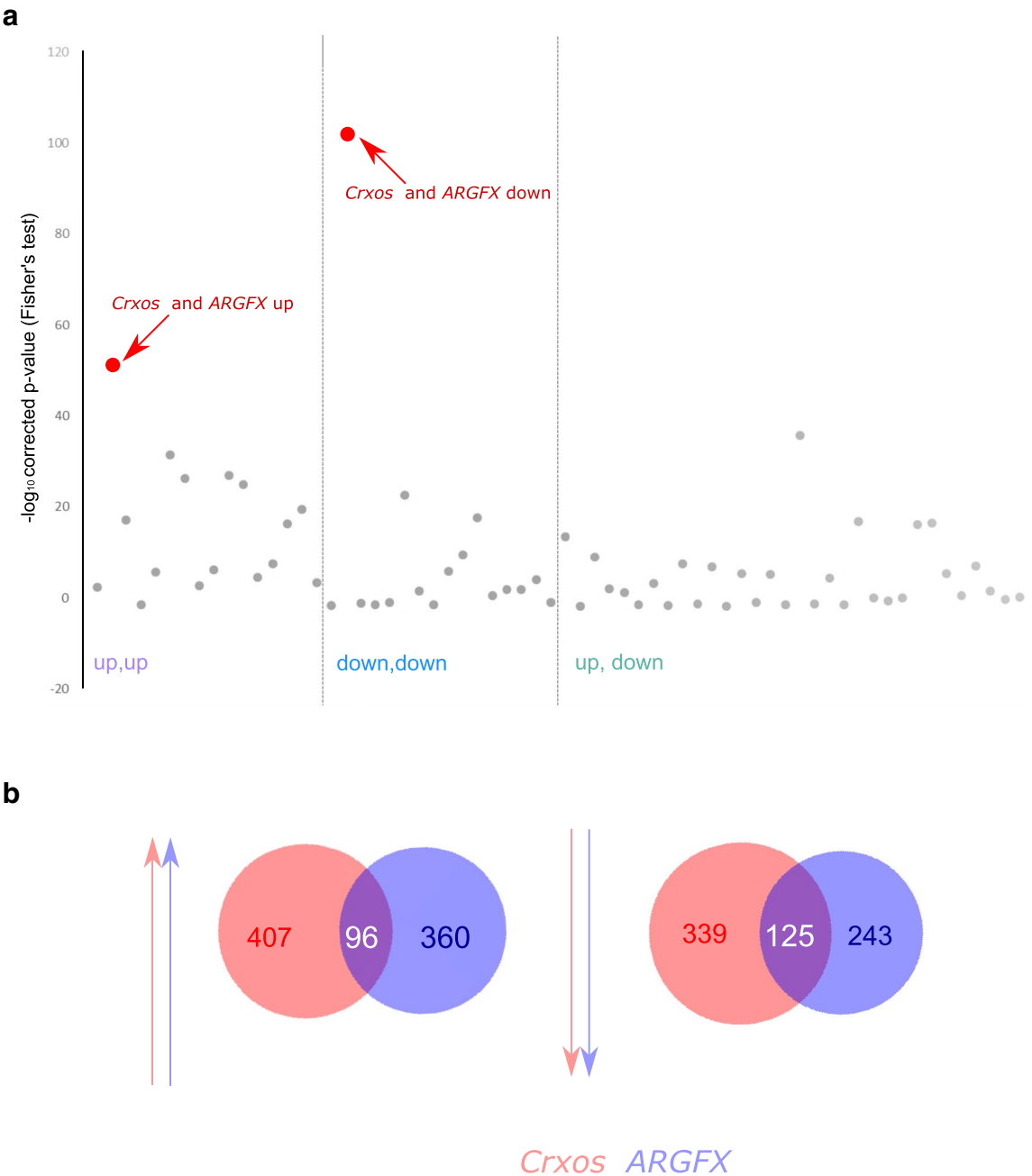


Fig. 5 Similarity between genes downstream of human *ARGFX* and mouse *Crxos*. **a** The sets of genes up- or down-regulated following ectopic expression are compared between each mouse and human ETCHbox gene after filtering for one-to-one orthologues (human data from Ref. [7]); *y*-axis shows -log(*p* values) derived from pairwise Fisher's exact test. Many comparisons are significant; the most striking similarities are between downstream targets of mouse *Crxos* and human *ARGFX* (up-regulated genes $p = 3 \times 10^{-52}$; down-regulated genes $p = 4.2 \times 10^{-102}$). Order of comparisons, from left to right, given in Additional file 6. **b** Proportional Venn diagrams showing extent of overlap between human and mouse one-to-one orthologues affected by *Crxos* ectopic expression in primary mouse embryonic fibroblasts and *ARGFX* in primary adult human fibroblasts

embryogenesis [26–29], with caution urged about extrapolation between species [30]. Furthermore, consistent with the hourglass model which describes propensity for evolutionary change early and late in development, transcriptomic analyses have highlighted variation in the earliest stages of development between different mammalian species [31, 32]. Similarly, several of the ETCHbox genes implicated in regulation of human preimplantation

Table 2 Genes with highest up-regulated expression fold change following ectopic expression of mouse ETCHbox genes

Rank	*Oboxa1*		*Oboxa4*		*Oboxa7*		*Crxos*	
	Gene	Log_2 fold change	Gene	Log_2 fold change	Gene	Log_2 fold change	Gene	Log_2 fold change
1	*Lgr5*	2.02	*Megf6*	2.24	*Fmo1*	2.45	*Apod*	1.85
2	*Megf6*	1.82	*Lgr5*	1.91	*Megf6*	1.65	*Serpina3g*	1.64
3	*Myh1*	1.66	*Myh1*	1.87	*Myh11*	1.58	*Tnn*	1.56
4	*Myh11*	1.54	*Eln*	1.78	*Egfl6*	1.57	*Megf6*	1.53
5	*Col8a2*	1.48	*Gja5*	1.78	*Cdo1*	1.55	*Cxcl14*	1.53
6	*Eln*	1.47	*Gdpd2*	1.71	*Rbp3*	1.44	*Mfap4*	1.53
7	*Gja5*	1.38	*Col8a2*	1.71	*Apoe*	1.39	*Dpt*	1.52
8	*Egfl6*	1.38	*Gucy1a3*	1.69	*Chit1*	1.35	*Egfl6*	1.49
9	*Itga11*	1.36	*Aoc3*	1.68	*Gja5*	1.32	*Ch25h*	1.48
10	*Epha3*	1.36	*Hmcn1*	1.64	*Slpi*	1.31	*Serping1*	1.43

gene expression have been secondarily lost in mice [7]. These findings raise questions about how far pathways and regulatory networks discovered in mice can be applied to human preimplantation development and vice versa. We have studied the mouse ETCHbox genes which, like their human orthologues, are expressed specifically in the preimplantation embryo but which have very different numbers and genomic composition to humans. Extensive gene loss, sequence divergence and duplication of the remaining genes in mouse allowed us to investigate how differences at the genomic level relate to species-specific differences or similarities in preimplantation development.

In mouse, four of the six ancestral ETCHbox gene families have been lost leaving orthologues of just *Tprx1* and *Trpx2*. The mouse *Crxos* gene is the orthologue of human *TPRX1*. *Crxos* has duplicated and is processed to give three transcripts with common temporal expression profiles in the preimplantation mouse embryo. The 66 *Obox* loci are orthologous to human *TPRX2* and can be divided into four groups and six subgroups, with expression profiles mirroring molecular phylogenetic classification. The concordance between sequence and expression suggests there may be functional redundancy within *Obox* subgroups during preimplantation mouse development.

To investigate downstream activities and functional similarity to human genes, we ectopically expressed three *Obox* genes from different subgroups and the most highly expressed *Crxos* transcript in mouse embryonic fibroblasts. It could be argued that these cell types are very different from cells of the preimplantation embryo, but to conduct evolutionarily meaningful comparisons with data previously obtained using human fibroblasts [7] it is important to use comparable cell types. Additionally, ectopic expression of ETCHbox genes in primary cells is comparable to reprogramming experiments, which commonly use primary fibroblasts as the initial cell population.

We found that each of the genes elicited transcriptomic changes related to gene expression profiles of preimplantation stages, suggesting we have partially recapitulated the in vivo roles of mouse ETCHbox genes. For example, we find that over-expression of *Crxos* modifies the transcriptome of embryonic fibroblasts to partially mimic the transcriptome of the blastocyst, down-regulating transcripts that fall in abundance during cleavage stages and up-regulating genes with a blastocyst peak. These results are exciting since the blastocyst is composed of the descendent cell types from the first cell lineage decision: trophectoderm and pluripotent inner cell mass. It is also the time when the embryonic secretome communicates with maternal tissue before subsequent blastocyst invasion into the endoderm [27]. A role in formation of these cell types is emphasised by the finding that orthologues of the genes most strongly up-regulated by *Crxos* are implicated in implantation in human development: *Apod* and *Serpina3* (Table 2).

Similarly, over-expression of *Oboxa4* caused transcriptomic changes mirroring those of the early embryo. *Oboxa4* is expressed maternally in the oocyte, and high levels of RNA are detected in the zygote. Ectopic expression of this gene caused down-regulation of genes that are not detectable as RNA in the earliest developmental stages, and up-regulation of those that are. In mouse ontogeny, zygotic genome activation (ZGA) is initiated earlier than in humans, with an early major wave of activation at the 2-cell stage and further waves of activation occurring until the morula stage [33]. The expression profiles of genes downstream of *Oboxa4* suggest that it may have a role in delaying or suppressing expression of embryonic genes, affecting timing of ZGA (Fig. 6).

For all *Obox* genes tested, we detect evidence of regulation of genes expressed in the eight-cell embryo and thus substantially earlier than the effect driven by *Crxos* ectopic expression. However, because this effect is seen for three different *Obox* genes, this may partially reflect cross-regulation between targets of closely related genes. On the basis of the normal expression profiles of *Obox* genes, we suggest the common effect is most likely to reflect in vivo function of *Oboxa1* or *Oboxa7* (Table 1). The eight-cell stage of mouse development, when many putative *Obox* target genes are expressed, is a particularly interesting developmental period. At this stage, the cells of the embryo increase their cell–cell contacts and elongate, and the embryo compacts in a critical morphological change which prepares the embryo for implantation (Fig. 6) [34–36]. Functions of genes affected include interactions with the ECM, adhesion to the external environment and DNA replication.

One of the most striking findings of this study was the discovery of highly significant overlap between the set of genes up- and down-regulated by ectopic mouse ETCHbox genes with the set of genes affected by ectopic expression of human ETCHbox genes. We find many comparisons show significant overlaps (Fig. 5a), which may reflect widespread overlapping functions between the majority of ETCHbox genes. Most notably, many of the inferred downstream targets of *Crxos* in mouse are orthologous to the inferred downstream targets of *ARGFX* in human (96 orthologues commonly up-regulated, 125 commonly down-regulated). Much smaller overlap was detected for other pairs of mouse and human ETCHbox genes. There are differences between the biological activities of the two genes, however. For example, expressing *ARGFX* in human fibroblasts induced transcriptional changes that mirror a pulse of expression in the eight-cell human embryo [7]. In contrast, *Crxos*-induced transcriptional changes more closely parallel changes occurring at the blastocyst stage. Furthermore, *Crxos* has broader expression in mouse preimplantation development than does *ARGFX* in human. Together, these data suggest that before or after the loss of *ARGFX* in murid evolution, the *Crxos* gene took over roles originally undertaken by *ARGFX* and has also acquired additional targets and biological functions (Fig. 7). It is particularly intriguing that this functional compensation involved deployment of a paralogous rather than an orthologous homeobox gene. Indeed, *Crxos* seems to have contrasting transcriptomic effects to its direct orthologue *TPRX1*, with significant overlap between genes down-regulated by *Crxos* and those up-regulated by *TPRX1* ($p = 1.58^{-36}$, Additional file 7).

The convergence of function between human and mouse ETCHbox genes provides an intriguing example of compensation and refinement of gene function alongside dynamic lineage-specific gene loss and expansions (Fig. 7). Such compensation would be favoured if

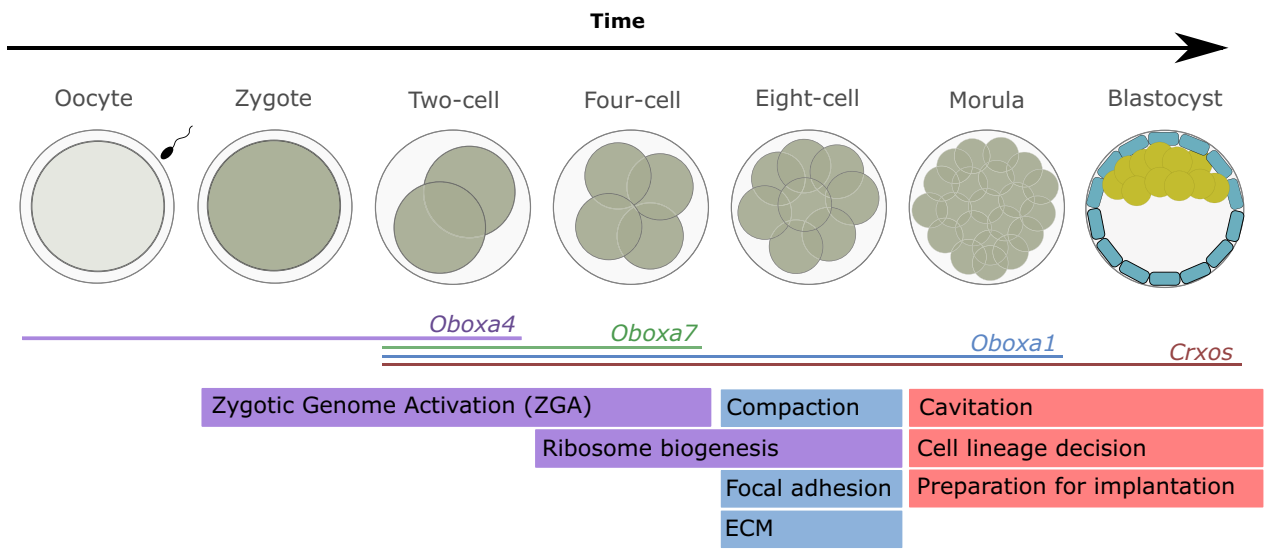


Fig. 6 Cellular and embryonic processes potentially regulated by mouse ETCHbox genes. Global transcriptomic changes elicited by ectopic expression, and the embryonic expression profiles of ETCHbox genes themselves, highlight possible developmental milestones regulated by ETCHbox genes. We suggest that *Crxos* is involved in preparing the embryo for the first cell fate decision prior to the early blastocyst stage. *Obox* genes likely regulate a range of biological processes in vivo; we suggest *Oboxa4* is involved with early milestones including induction of zygotic gene expression, whereas *Oboxa1* and *Oboxa7* are involved in later events such as embryo compaction. Expression of the ETCHbox genes tested is represented by coloured lines (*Crxos* red, *Oboxa1* blue, *Oboxa4* purple, *Oboxa7* green)

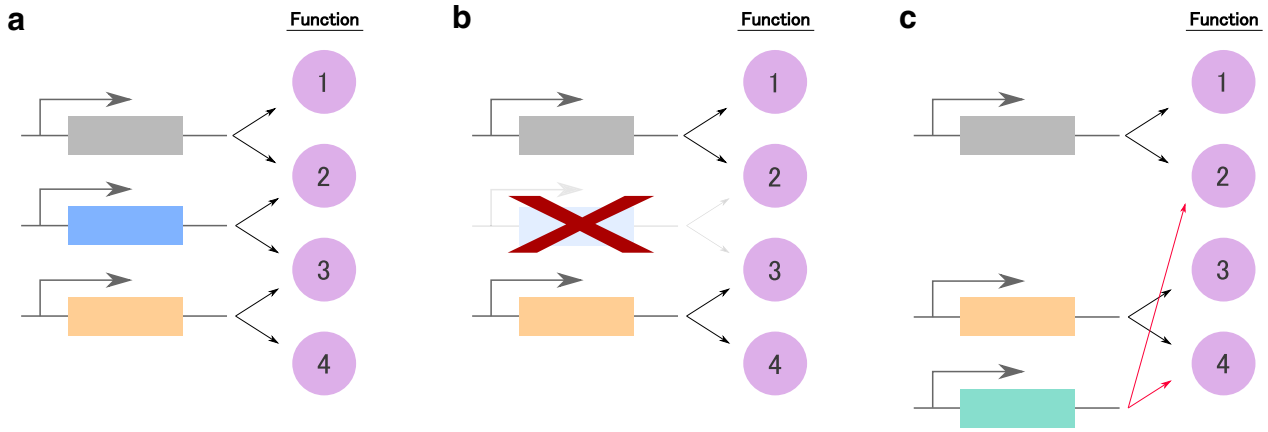


Fig. 7 Simplified scenario for the evolution of functional redundancy. **a** A group of genes contribute to several biological processes with partial overlap of functions. **b** Partial functional overlap permits gene loss to persist as a temporary state. **c** Redundancy can be restored through gene duplication and divergence. The temporal sequence of (**b**) and (**c**) can be reversed. Partial redundancy affords buffering against somatic failure of a gene or process

the system displays 'distributed robustness' as defined by Wagner [37]. Under this model, individual genes may have overlapping rather than redundant functions, such that collectively the system has multiple routes to achieve the same endpoints. Extrapolating, we suggest that a variable set of ETCHbox genes regulates a suite of functions in eutherian preimplantation embryos that are shared between species, regardless of the particular repertoire of ETCHbox genes retained in a genome.

Conclusions

The evolutionary history of ETCHbox genes, with extensive gene duplication, loss and sequence change during mammalian radiation, contrasts sharply to that of most other homeobox genes. We show that ectopic expression of these genes in mouse embryonic fibroblasts elicits transcriptomic changes that mirror transcriptomic changes in early development. For example, the maternally expressed *Oboxa4* gene can reduce expression of genes normally active after ZGA, other *Obox* genes elevate expression of genes that peak at the eight-cell stage, and *Crxos* expression causes transcriptomic changes mirroring pre-blastocyst development. Most strikingly, we find close similarity between the genes downstream of mouse *Crxos* and human *ARGFX*, despite these not being orthologous homeobox genes. These results point to an evolutionarily labile system where a set of regulatory genes can duplicate and be lost in evolution, but shift functionally to compensate for gene loss. Functional replacement by a paralogous gene might be favoured when evolution builds an inherently redundant system to ensure the robustness of critical developmental events.

Methods

Embryonic expression and phylogenetic analyses

To examine endogenous expression patterns, RNA sequencing files were acquired from the SRA database (SRA identifier SRP034543; Additional file 8) and aligned to the GRCm38/mm10 *M. musculus* genome using the STAR alignment tool [38]. The Cufflinks tool was used to extract FPKM expression values (Additional file 9). To investigate evolutionary relationships between genes, maximum likelihood phylogenies were generated from nucleotide or deduced amino acid sequences using RAxML with 500 bootstraps [39].

Ectopic gene expression

Mouse embryonic fibroblasts (Sciencell #M7540-57) were cultured in Dulbecco's modified Eagle medium (DMEM, ThermoFisher #41966-029) supplemented with 10% foetal calf serum (FCS) and 1% Pen–Strep in 5% CO_2 at 37 °C on 0.001% poly-L-lysine-coated plasticware. Coding sequences of *Oboxa1*, *Oboxa4*, *Oboxa7* and *Crxos* were obtained from GenScript and cloned into a vector co-expressing daGFP; genes were C-terminally tagged with V5 and under control of a CMV promoter (Oxford Genetics #OG244). For ectopic expression, 10^6 cells in 100 µl Opti-Mem (Gibco 11058-021) were electroporated (NEPA GENE; poring pulse: 175 volts, 5 ms, 4 pulses) with 10 µg of plasmid DNA and seeded in DMEM, 10% FCS, 1% Pen–Strep antibiotic (ThermoFisher #15140-122). Media was changed after 24 h and cells collected for FACS sorting after 48 h to enrich for transfected cells. RNA was extracted using the RNeasy microkit (Qiagen) and quality checked on an Experion electrophoresis station (Bio-Rad). RNA sequencing was performed on three

replicates from each experimental condition and control (empty vector transfection); libraries were prepared using Illumina TruSeq, and paired-end RNA sequencing reads generated on the Hi-Seq 4000 platform (Oxford Genomics Centre), yielding 38.2–58.2 million reads per sample. Sequence reads were aligned to the GRCm38/mm10 *Mus musculus* genome using the STAR alignment tool, and gene expression levels for protein-coding genes assessed according to NCBI annotations (ftp://ftp.ncbi.nlm.nih.gov/genomes/Mus_musculus/ accessed 13 July 2016, FPKM read-outs in Additional file 9). The ectopically expressed genes were expressed at a level (assessed by FPKM) comparable to their embryonic expression levels (Additional file 1L).

Differential gene expression analysis

DESeq2 was used to identify genes differentially expressed in response to homeobox gene transfection following FeatureCounts to retrieve raw read counts (Additional file 10) [40]. Experiments were matched by date to minimise the effects of day-to-day variation. The Cufflinks tool was used to estimate mean FPKM expression values (Additional file 11). Lists of up- and down-regulated genes for each experimental condition were produced using criteria of $p < 0.05$ (with Benjamini–Hochberg correction) and expression fold change greater than 1.25 or − 1.25 (Additional file 3). Transcripts with an FPKM > 2 were considered to be actively expressed.

To generate sets of genes with similar temporal expression profiles across preimplantation mouse development (oocyte, zygote, two cell, four cell, eight cell, morula and blastocyst), we took mouse RNAseq data processed as above, filtered to identify genes with a FPKM variance > 5 across preimplantation development, and clustered these into profiles using Mfuzz [41, 42]. This analysis generated 150 initial temporal profiles of gene expression. Sets of profiles with a correlation coefficient of over 0.95 were merged into composite clusters which are identified by IDs > 200 (Additional files 1G and 4). When high correlation coefficients were not all reciprocal within a set of profiles, an expression dendrogram was used to guide the process of merging.

Profile enrichment

Pearson's Chi-square test was used to test the null hypothesis that the number of differentially expressed genes is proportional across all temporal profiles. If the proportions are not equal between profiles ($p < 0.05$), Pearson's statistic was used to identify the contribution of each profile to the overall difference. After removal of profiles inferred to contribute to the difference, Fisher's exact test was used to verify that the differentially expressed genes were proportionally distributed between the remaining profiles and to find statistical difference between occupation of enriched and non-enriched profiles.

Additional files

Additional file 1. (A) *Obox* phylogeny using nucleotide sequences. (B) Genomic organisation of *Mus musculus* ETCHbox genes. (C) Proposed new *Obox* nomenclature. The term 'partial' indicates an Obox sequence lacking a complete homeobox, and the term is not part of proposed nomenclature. (D) *Obox* phylogeny using protein sequences that contain a complete homeodomain. (E) Summary of number of homeoboxes in *Tprx1* syntenic region in different species. * In rat, the two homeobox sequences reflect an independent regional duplication. (F) Alignment of *M. musculus* and *M. spretus Tprx1/Crxos* homeodomains. (G) Normalised expression graphs of each temporal profile generated by Mfuzz. (H) Venn diagrams shown numbers of genes affected significantly by each treatment, and overlaps between experiments. (I) Enriched GO terms for genes affected by *Obox* ectopic expression. (J) Genes most strongly down-regulated following ectopic expression. (K) Effect of ectopic ETCHbox expression on selected genes with notable preimplantation or stem cell functions. (L) In vivo expression levels of ETCHbox genes at mouse preimplantation stages compared to ectopic expression levels of transfected genes.

Additional file 2. Nucleotide and protein sequences used for *Obox* phylogenetic analyses.

Additional file 3. Lists of genes significantly up- or down-regulated following ectopic expression.

Additional file 4. Mfuzz clusters and genes within each cluster.

Additional file 5. Observed and expected numbers of significantly regulated genes populating temporal expression profiles used for enrichment analysis.

Additional file 6. List of one-to-one orthologues similarly regulated following ectopic expression of human *ARGFX* and mouse *Crxos*.

Additional file 7. Data used for statistical tests of overlap between sets of one-to-one orthologues regulated by mouse and human ETCHbox genes.

Additional file 8. SRA data information used to generate embryonic gene expression data.

Additional file 9. Gene expression (FPKM) across mouse embryonic stages.

Additional file 10. Gene expression (raw reads) from triplicate RNA sequencing of mouse embryonic fibroblasts following ectopic expression of mouse ETCHbox genes.

Additional file 11. Average FPKM values for experimental datasets.

Abbreviations

CMV: cytomegalovirus; DMEM: Dulbecco's modified Eagle medium; ECM: extracellular matrix; FACS: fluorescence-activated cell sorting; FCS: foetal calf serum; FPKM: fragments per kilobase of transcript per million mapped reads; NCBI: National Center for Biotechnology Information; SRA: sequence read archive; ZGA: zygotic genome activation.

Authors' contributions

AHR performed molecular and cell biology experiments, transcriptomics and statistical analyses of data. IM undertook genomic annotation of *Obox* genes. AHR, TLD and PWHH designed the ectopic expression experiments and analysis strategy. AHR and PWHH wrote the manuscript with input from TLD and IM. All authors read and approved the final manuscript.

Author details

[1] Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. [2] Centro Andaluz de Biología del Desarrollo, Consejo Superior de Investigaciones Científicas/Universidad Pablo de Olavide, 41013 Seville, Spain.

Acknowledgements
We acknowledge the Oxford Genomics Centre, Wellcome Trust Centre for Human Genetics, for transcriptome sequencing, and the Flow Cytometry Facility, Experimental Medicine Division, University of Oxford, for assistance with cell sorting. We thank Adam Hargreaves, Daniel Lunn and Monica McAndrews for discussions and advice.

Competing interests
PWHH, TLD and IM have patent applications in progress related to the topic of this research.

Availability of data and materials
The RNA sequencing raw datasets generated in this study are available in the NCBI Gene Expression Omnibus repository under accession GSE108060. The mouse developmental stage RNA sequencing data analysed in this study were obtained from NCBI SRA Project SRP034543 [42]. All transformed data are included in this published article and its additional information files.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
Funding was provided through the Oxford Interdisciplinary Bioscience BBSRC Doctoral Training Partnership (BB/M011224/1), the EPA Cephalosporin Fund and a Royal Society International Exchanges Grant.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 19 October 2017 Accepted: 7 January 2018
Published online: 27 January 2018

References

1. Holland PWH. Evolution of homeobox genes. Wiley Interdiscip Rev Dev Biol. 2013. https://doi.org/10.1002/wdev.78.
2. MacLean JA II, Chen MA, Wayne CM, Bruce SR, Rao M, Meistrich ML, Macleod C, Wilkinson MF. *Rhox*: a new homeobox gene cluster. Cell Press. 2005. https://doi.org/10.1016/j.cell.2004.12.022.
3. Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sanchez-Flores A, Brooks KL, Tracey A, Bobes RJ, Fragoso G, Sciutto E, Aslett M, Beasley H, Bennett HM, Cai J, Camicia F, Clark R, Cucher M, De Silva N, Day TA, Deplazes P, Estrada K, Fernandez C, Holland PWH. The genomes of four tapeworm species reveal adaptations to parasitism. Nature. 2013. https://doi.org/10.1038/nature12031.
4. Zhong Y, Holland PWH. The dynamics of vertebrate homeobox gene evolution: gain and loss of genes in mouse and human lineages. BMC Evol Biol. 2011. https://doi.org/10.1186/1471-2148-11-169.
5. Booth HAF, Holland PWH. Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. Gene. 2007. https://doi.org/10.1016/j.gene.2006.07.034.
6. Holland PWH, Booth HAF, Bruford EA. Classification and nomenclature of all human homeobox genes. BMC Biol. 2007. https://doi.org/10.1186/1741-7007-5-47.
7. Maeso I, Dunwell TL, Wyatt CDR, Marlétaz F, Vetö B, Bernal JA, Quah S, Irimia M, Holland PWH. Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals. BMC Biol. 2016. https://doi.org/10.1186/s12915-016-0267-0.
8. Töhönen V, Katayama S, Vesterlund L, Jouhilahti EM, Sheikhi M, Madissoon E, Filippini-Cattaneo G, Jaconi M, Johnsson A, Bürglin TR, Linnarsson S, Hovatta O, Kere J. Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. Nat Commun. 2015. https://doi.org/10.1038/ncomms9207.
9. Madissoon E, Jouhilahti EM, Vesterlund L, Töhönen V, Krjutškov K, Petropoulos S, Einarsdottir E, Linnarsson S, Lanner F, Månsson R, Hovatta O, Bürglin TR, Katayama S, Kere J. Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. Sci Rep. 2016. https://doi.org/10.1038/srep28995.
10. Wilming LG, Boychenko V, Harrow JL. Comprehensive comparative homeobox gene annotation in human and mouse. Database (Oxford). 2015. https://doi.org/10.1093/database/bav091.
11. Alfano G, Vitiello C, Caccioppoli C, Caramico T, Carola A, Szego MJ, McInnes RR, Auricchio A, Banfi S. Natural antisense transcripts associated with genes involved in eye development. Hum Mol Genet. 2005. https://doi.org/10.1093/hmg/ddi084.
12. Saito R, Yamasaki T, Nagai Y, Wu J, Kajiho H, Yokoi T, Noda E, Nishina S, Niwa H, Azuma N, Katada T, Nishina H. *Crxos* maintains the self-renewal capacity of murine embryonic stem cells. BBRC. 2009. https://doi.org/10.1016/j.bbrc.2009.09.118.
13. Saito K, Abe H, Nakazawa M, Irokawa E, Watanabe M, Hosoi Y, Soma M, Kasuga K, Kojima I, Kobayashi M. Cloning of complementary DNAs encoding structurally related homeoproteins from preimplantation mouse embryos: their involvement in the differentiation of embryonic stem cells. Biol Reprod. 2010. https://doi.org/10.1095/biolreprod.108.075697.
14. Iha M, Watanabe M, Kihara Y, Sugawara S, Saito K, Soma M, Sato S, Mori Y, Kasuga K, Kojima I, Sasamura R, Murata J, Kobayashi M. Effect of ectopic expression of homeoprotein EGAM1C on the cell morphology, growth, and differentiation in a mouse embryonic stem cell line, MG1.19 cells. Reproduction. 2012. https://doi.org/10.1530/REP-11-0379.
15. Cui W, Dai X, Marcho C, Han Z, Zhang K, Tremblay KD, Mager J. Towards functional annotation of the preimplantation transcriptome: an RNAi screen in mammalian embryos. Sci Rep. 2016. https://doi.org/10.1038/srep37396.
16. Soma M, Iha M, Kihara Y, Sato S, Sato Y, Sato S, Mori Y, Sugawara S, Kasuga K, Kojima I, Kobayashi M. Preferential emergence of cell types expressing markers for primitive endoderm lineages in mouse embryonic stem cells expressing exogenous EGAM1 homeoprotein. J Biosci Bioeng. 2012;114:342–6. https://doi.org/10.1016/j.jbiosc.2012.04.012.
17. Sato S, Nakazawa M, Kihara Y, Kubo Y, Sato Y, Kikuchi T, Nonaka A, Sasaki A, Iwashita J, Murata J, Hosaka M, Kobayashi M. Partial inhibition of differentiation associated with elevated protein levels of pluripotency factors in mouse embryonic stem cells expressing exogenous EGAM1N homeoprotein. J Biosci Bioeng. 2015. https://doi.org/10.1016/j.jbiosc.2015.03.001.
18. Rajkovic A, Yan C, Yan W, Klysik M, Matzuk MM. *Obox*, a family of homeobox genes preferentially expressed in germ cells. Genomics. 2002. https://doi.org/10.1006/geno.2002.6759.
19. Cheng W-C, Hsieh-Li HM, Yeh Y-J, Li H. Mice lacking the *Obox6* homeobox gene undergo normal early embryonic development and are fertile. Dev Dyn. 2007. https://doi.org/10.1002/dvdy.21261.
20. Ge SX. Exploratory bioinformatics investigation reveals importance of "Junk" DNA in early embryo development. BMC Genom. 2017. https://doi.org/10.1186/s12864-017-3566-0.
21. Sato S, Morita S, Iha M, Mori Y, Sugawara S, Kasuga K, Kojima I, Ozaki N, Muraguchi H, Okano K, Iwashita J, Murata J, Hosaka Kobayashi M. Intact structure of EGAM1 homeoproteins and basic amino acid residues in the common homeodomain of EGAM1 and EGAM1C contribute to their nuclear localization in mouse embryonic stem cells. J Biosci Bioeng. 2013. https://doi.org/10.1016/j.jbiosc.2013.02.007.
22. Pálfy M, Joseph SR, Vastenhouw NI. The timing of zygotic genome activation. Curr Opin Genet Dev. 2017. https://doi.org/10.1016/j.gde.2016.12.001.
23. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell. 2007. https://doi.org/10.1016/j.cell.2007.11.019.
24. Assou S, Boumela I, Haouzi D, Monzo C, Dechaud H, Kadock I-J, Hamamah S. Transcriptome analysis during human trophectoderm specification suggests new roles of metabolic and epigenetic genes. PLoS ONE. 2012. https://doi.org/10.1371/journal.pone.0039306.
25. Reinholt BM, Bradley JS, Jacobs RD, Ealy AD, Johnson SE. Tissue organization alters gene expression in equine induced trophectoderm cells. Gen Comp Endocrinol. 2017. https://doi.org/10.1016/j.ygcen.2017.01.030.

26. Rossant J, Cross JC. Placental development: lessons from mouse mutants. Nat Rev Genet. 2001. https://doi.org/10.1038/35080570.
27. Wang H, Dey SK. Roadmap to embryo implantation: clues from mouse models. Nat Rev Genet. 2006. https://doi.org/10.1038/nrg1808.
28. Juriloff DM, Harris MJ. Mouse models for neural tube closure defects. Hum Mol Genet. 2000. https://doi.org/10.1093/hmg/9.6.993.
29. Sulik KK, Johnston MC, Webb MA. Fetal alcohol syndrome: embryogenesis in a mouse model. Science. 1981. https://doi.org/10.1126/science.6795717.
30. Ménézo YJR, Hérubel F. Mouse and bovine models for human IVF. Reprod BioMed Online. 2002. https://doi.org/10.1016/S1472-6483(10)61936-0.
31. Kuijk E, Geijsen N, Cuppen E. Pluripotency in the light of the developmental hourglass. Biol Rev. 2014. https://doi.org/10.1111/brv.12117.
32. Breschi A, Gingeras TR, Guigó R. Comparative transcriptomics in human and mouse. Nat Rev Genet. 2017. https://doi.org/10.1038/nrg.2017.19.
33. Xue Z, Huang K, Cai C, Cai L, Jiang C, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE, Liu J, Horvath S, Fan G. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature. 2013. https://doi.org/10.1038/nature12364.
34. Fierro-González JC, White MD, Silva JC, Plachta N. Cadherin-dependent filopodia control preimplantation embryo compaction. Nat Cell Biol. 2013. https://doi.org/10.1038/ncb2875.
35. Johnson MH, Maro B, Takeichi M. The role of cell adhesion in the synchronization and orientation of polarization in 8-cell mouse blastomeres. J Embryol Exp Morphol. 1986;93:239–55.
36. Landry DW, Zucker HA, Sauer MV, Reznik M, Wiebe L. Hypocellularity and absence of compaction as criteria for embryonic death. Regener Med. 2006. https://doi.org/10.2217/17460751.1.3.367.
37. Wagner A. Distributed robustness versus redundancy as causes of mutational robustness. BioEssays. 2005. https://doi.org/10.1002/bies.20170.
38. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TRSTAR. Ultrafast universal RNA-seq aligner. Bioinformatics. 2013. https://doi.org/10.1093/bioinformatics/bts635.
39. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014. https://doi.org/10.1093/bioinformatics/btu033.
40. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014. https://doi.org/10.1186/s13059-014-0550-8.
41. Kumar L, Futschik M. Mfuzz: a software package for soft clustering of microarray data. Bioinformation. 2007. https://doi.org/10.6026/97320630002005.
42. Fan X, Zhang X. Single-cell RNA-Seq transcriptomic analysis of circular RNAs in mouse embryos. 2015; NCBI SRA SRP034543 https://www.ncbi.nlm.nih.gov//geo/query/acc.cgi?acc=GSE53386.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit



PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Royall AH, Frankenberg S, Pask AJ, Holland PWH. 2019 Of eyes and embryos: subfunctionalization of the *CRX* homeobox gene in mammalian evolution. *Proc. R. Soc. B* **286**: 20190830.
http://dx.doi.org/10.1098/rspb.2019.0830

Received: 8 April 2019
Accepted: 28 June 2019

Subject Category:
Evolution

Subject Areas:
developmental biology, evolution, genomics

Keywords:
marsupial, blastocyst, preimplantation, gene duplication, transcription factor, DDC

Author for correspondence:
Peter W. H. Holland
e-mail: peter.holland@zoo.ox.ac.uk

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4566431.

THE ROYAL SOCIETY PUBLISHING

Of eyes and embryos: subfunctionalization of the *CRX* homeobox gene in mammalian evolution

Amy H. Royall[1], Stephen Frankenberg[2], Andrew J. Pask[2] and Peter W. H. Holland[1]

[1]Department of Zoology, University of Oxford, 11a Mansfield Road, Oxford OX1 3SZ, UK
[2]School of BioSciences, University of Melbourne, Melbourne, 3010 Victoria, Australia

AJP, 0000-0002-1900-2263; PWHH, 0000-0003-1533-9376

ETCHbox genes are fast-evolving homeobox genes present only in eutherian (placental) mammals which originated by duplication and divergence from a conserved homeobox gene, *Cone-rod homeobox* (*CRX*). While expression and function of *CRX* are restricted to the retina in eutherian mammals, ETCHbox gene expression is specific to preimplantation embryos. This dramatic difference could reflect the acquisition of new functions by duplicated genes or subfunctionalization of pleiotropic roles between *CRX* and ETCHbox genes. To resolve between these hypotheses, we compared expression, sequence and inferred function between *CRX* of metatherian (marsupial) mammals and ETCHbox genes of eutherians. We find the metatherian *CRX* homeobox gene is expressed in early embryos and in eyes, unlike eutherian *CRX*, and distinct amino acid substitutions were fixed in the metatherian and eutherian evolutionary lineages consistent with altered transcription factor specificity. We find that metatherian *CRX* is capable of regulating embryonically expressed genes in cultured cells in a comparable way to eutherian ETCHbox. The data are consistent with *CRX* having a dual role in eyes and embryos of metatherians, providing an early embryonic function comparable to that of eutherian ETCHbox genes; we propose that subfunctionalization of pleiotropic functions occurred after gene duplication along the placental lineage, followed by functional elaboration.

1. Introduction

Gene duplication is postulated to facilitate evolutionary change and innovation through providing new genetic material which can gain or lose functions over evolutionary time. Although most homeobox gene families are highly conserved across Metazoa, there are examples of homeobox genes that have undergone duplication followed by extreme sequence divergence in some lineages.

The Otx homeobox gene family provides a striking example. Following characterization of the *Drosophila orthodentical* (*otd*) gene, two homologues were rapidly identified in mammals: *OTX1* and *OTX2* [1]. Subsequently, a third gene family member was identified, named *OTX5* or *CRX* [2,3]. These three Otx family genes are descendants of the whole-genome duplications that occurred in vertebrate ancestry; they are found across jawed vertebrates and share high sequence similarity. More recently, it was shown that a series of highly divergent homeobox genes, specific to eutherian mammals, originated later in evolution by tandem duplication from the *CRX* gene followed by extreme sequence divergence away from a canonical Otx-type sequence. Their pattern of sequence divergence has been described as asymmetric, meaning that the 'parental' gene (*CRX*) has undergone far less sequence change than the 'daughter' genes [3–6]. The descendent genes are so distinct in sequence

© 2019 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

from Otx family genes that they were given their own gene family names: *Tprx*, *Dprx*, *Leutx* and *Argfx* [7] and collectively ETCHbox [5]. This innovation and sequence change is of note as lineage-specific genes could have important roles in evolution [8]. In eutherian mammals, the 'parental' *CRX* gene locus is flanked by *Tprx1* (called *Crxos* in mouse) and *Tprx2* (Obox genes in mouse); the other duplicates are more distantly located or on another chromosome [5,9]. There is also variation between eutherian species as to which ETCHbox are retained or lost, and which have been duplicated further [5].

Divergent sequence evolution is mirrored by distinct expression patterns and probable functions. The ETCHbox genes of human and cow are expressed predominantly in early development, in a sharp pulse around the 8-cell to 16-cell stages [5,10,11]. Mouse ETCHbox genes are more variable in expression, but are also specific to the very early embryo [9,12,13]. Ectopic expression in cell culture, followed by RNA sequencing (RNA-seq), has uncovered putative roles in controlling a suite of downstream genes, including genes peaking in expression before blastocsyst and implicated in cell fate and fetal–maternal interactions [5,9].

By contrast, the *CRX* gene encodes a transcription factor implicated in the specification of photoreceptor fate in the developing eye [3,14,15]. There is also some evidence for trace expression of *CRX* in the early embryo in human and mouse, but the levels are too low to be identified by whole embryo RNA-seq [5,9,16]. Mouse models with perturbed *CRX* function or *CRX* gene deletion are characterized by the near or complete absence of vision from birth, and mutations in humans can cause cone–rod dystrophy [14,17,18]. Expression in the eye has also been reported for *Crx*/*Otx5* in frog, chicken, zebrafish (two genes) and dogfish [2,19–21]. However, in some non-mammalian vertebrates, *CRX* expression has also been found in embryonic tissues. For example, in the frog *Xenopus laevis*, *crx (Xotx5)* RNA is detected in the Spemann organizer of the gastrula and in anterior neuroectoderm [2,22]; in dogfish *Scyliorhinus canicula*, *Crx (Otx5)* expression can be detected in the embryonic brain by *in situ* hybridization in addition to the developing eye [21,23].

It is unclear whether the vertebrate *CRX* gene had ancestral roles in the early embryo, which have been secondarily lost in the eutherian mammal lineage [3], or alternatively whether some sites of embryonic expression were independently gained in *Xenopus* and dogfish. This question is closely tied to understanding the evolution of ETCHbox genes, because these embryo-specific genes were derived from *CRX* in eutherian mammal evolution. If *CRX* was pleiotropic with roles in embryos and eyes prior to the emergence of eutherian mammals, the distinct functions in eutherians could reflect subfunctionalization after gene duplication, followed by gradual and progressive optimization of multiple ETCHbox roles to subtly different embryonic roles [9]. By contrast, if *CRX* was ancestrally eye-specific, ETCHbox genes and their functions would be very radical innovations acquired by eutherian mammals only, thus marking a clear distinction to other vertebrates. It is not clear which scenario is more likely, since eutherian mammal embryogenesis has a mixture of plesiomorphic and derived features. The eutherian mammal lineage encompasses the emergence of the extant Placentalia, characterized by extended embryogenesis and development of a highly invasive placenta that contributes to embryo nutrient exchange throughout embryogenesis. A comparison to non-eutherian mammals is key to resolving these questions. Marsupials (metatherians) are the immediate outgroup to eutherian mammals, and their embryos do establish maternal interactions albeit with a less invasive placenta than in eutherian mammals [24]. Furthermore, metatherians do not possess ETCHbox genes, as these arose from *CRX* specifically on the eutherian stem lineage.

In this study, we ask whether a metatherian *CRX* gene is eye-specific or whether it also has expression and possible function in the early embryo, comparable to ETCHbox genes. Using sequence comparisons, we first show that specific amino acid changes occurred in the *CRX* proteins of metatherians and eutherians, compatible with alterations to transcription factor function during mammalian evolution. We examined the expression of the *CRX* gene in metatherian development, detecting expression in early embryos and in the eye. To enable comparisons to previously characterized functions of eutherian ETCHbox genes, we expressed metatherian and eutherian *CRX* genes ectopically in cell culture and examined transcriptomic responses using RNA-seq and qPCR. These experiments uncovered similarities in activity between metatherian *CRX* and eutherian ETCHbox genes, consistent with subfunctionalization and progressive specialization after gene duplication in eutherian mammal evolution.

2. Material and methods

(a) Protein sequence analysis

Deduced protein sequences were gathered from NCBI and aligned using MAFFT to identify amino acid substitutions in the homeodomain. The inferred vertebrate ancestral sequence was estimated using FastML [25] (electronic supplementary material, S1A). We tested for positive selection using CodeML implemented in PAML [26], using a branch-model to estimate the dN/dS ratio by assigning two independent ratios, specifying the branch leading to eutherian mammals (model = 2, NSsites = 0).

(b) Gene expression analysis

To examine mouse gene expression, publicly available RNA-seq datasets (electronic supplementary material, S2A) from across preimplantation development were analysed using kallisto v0.42.4 using GRCm38 coding sequences and default settings to obtain a number of transcripts-per-million mapped reads (TPM) [27]. Coding sequences for *Oboxa7*, *Oboxb2* and *Oboxd2* were manually added to the GRCm38 coding sequence dataset as they are currently unannotated. Other mouse transcripts examined were *Crx* (ENSMUST00000044434.12), *Crxos* (ENSMUST00000171280.2), *Oboxa1* (ENSMUST00000108513.4), *Oboxa4* (ENSMUST00000067288.14), *Otx1* (ENSMUST00000006071.13) and *Otx2* (ENSMUST00000226501.1).

(c) Dunnart conceptus expression

Dunnart single-cell RNA-seq data were derived from a larger study to be published in a separate manuscript. Dissociated single cells from a range of conceptus stages were processed using Clontech SMART-Seq v4 3′ DE kit, with sample volumes scaled down five-fold. Libraries were prepared using the Illumina Nextera XT DNA Sample Preparation Kit and sequenced at 1 million reads per cell (12 cells per multiplexed library) on an Illumina Nextseq 500 platform. Using the Galaxy Web-based analysis platform (https://usegalaxy.org/), de-multiplexed reads were mapped to dunnart genomic scaffolds (S.F. & A.J.P. 2019, unpublished) using Bowtie for Illumina (default parameters) and

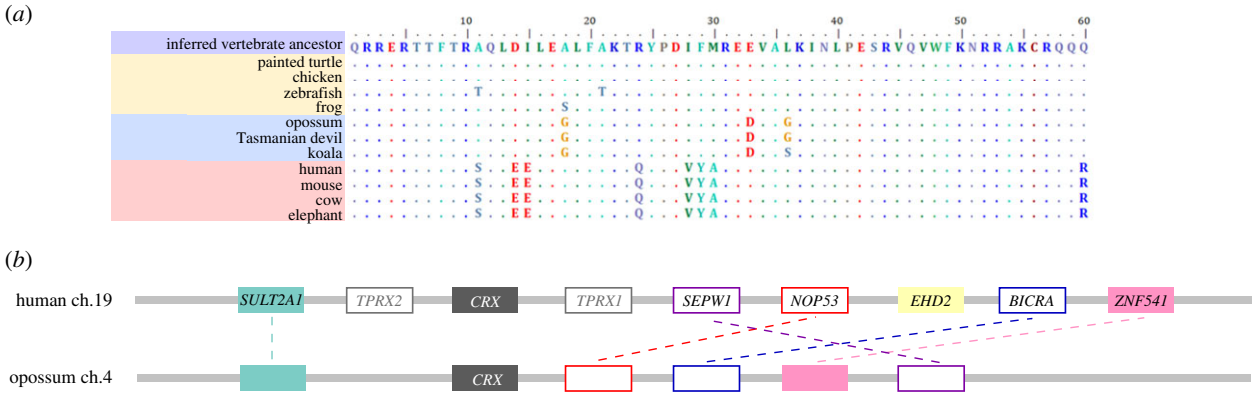


Figure 1. (*a*) Eight amino acid changes are present in all eutherian mammal CRX homeodomains and absent in non-eutherian vertebrates. Two conserved amino acid changes are metatherian-specific and not found outside of metatherians. (*b*) Synteny of opossum *CRX* relative to orthologous human genomic region confirms the identity of opossum *CRX* locus. Assembly of dunnart genome is not complete enough to describe the syntenic region. (Online version in colour.)

assembled into transcripts using Stringtie and Stringtie Merge. Stringtie was then reapplied to the mapped reads using the Stringtie Merge output as a transcripts reference file to produce the TPM (transcripts per million) dataset reported here. *CRX* was identified among the Stringtie Merge assembled transcripts using BLASTn.

(d) Reverse transcriptase polymerase chain reaction

Fat-tailed dunnart tissue collections were approved by the University of Melbourne Animal Experimentation Ethics Committees and were in accordance with the National Health and Medical Research Council of Australia (2004) guidelines. Total RNA was extracted from frozen adult female dunnart tissues using the RNeasy Mini Kit (Qiagen), followed by treatment with DNA-*free* (Ambion) to eliminate genomic DNA. Oligo(dt)-primed first-strand cDNA was reverse transcribed using Superscript III (Invitrogen). PCR was performed using GoTaq Green (Promega) in a 20-µl reaction volume for 40 cycles of 95°C for 15 s, 61°C for 20 s and 72°C for 20 s. Dunnart-specific primer sequences were: *CRX*- forward AAGATCAATCTCCCAGAATCCCGAG; *CRX*-reverse AGTC AGTGCATAGGAAGAGGAGG; *TBP*- forward TCCCCAAT-GACCCCTATGACTC; *TBP*-reverse TTTCTGGCTGCTAA CCTGGAC. Products were analysed by electrophoresis.

(e) Ectopic expression and transcriptomic responses

Plasmids containing V5-tagged human *CRX*, V5-tagged opossum *CRX* (RefSeq: XM_001372308.2, Genscript CloneID: Omc23124) or no homeobox gene, along with a puromycin resistance gene (Oxford Genetics; OG3422), were electroporated using a NEPA21 system (NEPAGENE) into Mouse Embryonic Fibroblasts (MEFs, ScienCell: M7540-57) and allowed to recover for 48 h in DMEM (Gibco: 11965092; 1% PenStrep (Gibco: 15140122), 10% FBS). Cells for immunocytochemistry were fixed in 4% formaldehyde in PBS for twenty minutes before permeabilization using 0.1% Triton-X in 1× PBS, 0.1% Tween-20. Antibodies and visualization are described previously [9].

Cells for RNA harvesting had media replaced for DMEM containing puromycin (0.5 µg ml^{-1}). Puromycin was kept in the media for 5 days with a media change on day 3. Puromycin was washed off using multiple short (5 min) incubations in DMEM and cultured overnight. RNA was harvested in triplicate per sample (control, metatherian *CRX*, eutherian *CRX*) using RNeasy Micro Plus kit (Qiagen) following manufacturer's instructions. RNA was prepared for sequencing using the Illumina TruSeq library preparation kit and size of the fragments checked (Agilent D1000 ScreenTape). RNA was sequenced using the Illumina NextSeq500 platform using 2 × 75 bp paired end reads giving between 68 and 121 million reads per sample. Transcriptome reads were aligned used STAR aligner v2.4.0 (genome assembly: mm10 GRCm38) and Cufflinks v2.2.1 software used to deduce FPKM values; FeatureCounts v1.4.6 produced raw read counts [9]. Clustering using Mfuzz software [28] and gene list overlap analysis was performed as described in Royall *et al.* [9]. Quantitative PCR (qPCR) experiments used New England BioScience Luna Universal One-Step qPCR kit according to manufacturer's instructions and Prime Pro 48 Real-Time qPCR system (Techne). qPCR analysis was performed relative to the *Gapdh* housekeeping gene and relative expression was calculated by comparison to control MEFs transfected with the 'no homeobox gene' plasmid.

Mouse-specific qPCR primers are *Gapdh*-forward CGGGTT CCTATAAATACGGACTG, *Gapdh*-reverse AATACGGCCAAAT CCGTTCA, *Hbegf*-forward TCCCTCTTGCAAATGCCTC, *Hbegf*-reverse GGACGACAGTACTACAGCCA, *Ptgs2*-forward TGGG TGTGAAGGGAAATAAGG, *Ptgs2*-reverse GAAGTGCTGGGC AAAGAATG, *Smarca1*-forward TATGCCCTTGAAAGCAGACC, *Smarca1*-reverse TGTTGGAGACTTCTGTGCTG.

To enable comparisons between transcriptomic responses following *CRX* and ETCHbox ectopic expression, differential expression analysis was carried out on FPKM counts using R package DESeq2 [29]. The resulting list of differentially expressed genes was compared with those differentially expressed following ETCHbox ectopic expression (previously obtained using the same methodology as in [9]) and overlapping responses calculated as a percentage of total differentially regulated genes in this study. The extent of overlap was assessed using Fisher's exact test. The GO term enrichment analysis used the functional annotation chart from DAVID Bioinformatics Resources v6.8 [30,31].

3. Results

(a) Eutherian-specific substitutions in CRX homeodomain

We hypothesized that if a pleiotropic CRX protein underwent subfunctionalization during the evolution of eutherian mammals, with some functions partitioned to ETCHbox genes, this may have resulted in novel amino acid substitutions associated with CRX specialization in the eutherian lineage. We searched for *CRX* sequences across vertebrates and confirmed the identity of opossum *CRX* using synteny to mouse and human loci (figure 1). As *CRX* encodes a transcription factor, we focused on the homeodomain sequence responsible for sequence-specific DNA binding and target gene specificity.

Comparison of the CRX homeodomain sequence from a range of vertebrates revealed high conservation, but with substitutions in some lineages (figure 1*a*).

We find substitutions at eight amino acid sites in the CRX homeodomain of eutherian mammals, all completely conserved between eutherians examined (figure 1*a*). By contrast, metatherian CRX has undergone substitutions at three sites (two conserved); other lineages have undergone limited change. A phylogenetic tree of CRX protein sequences clearly separates metatherian and eutherian *CRX* genes (electronic supplementary material, S2B).

We tested if the eutherian-specific substitutions were driven by positive selection, as indicated by an elevated dN/dS (ω) ratio, but found the long divergence times were associated with saturation of dS and gave unreliable ω estimation. Nonetheless, the amino acid changes in eutherian mammals are consistent with an altered function compared to other vertebrates, compatible with the subfunctionalization hypothesis. We find no evidence for reciprocal compensatory changes in ETCHbox homeodomains (electronic supplementary material, S2C).

(b) *CRX* expression in metatherian embryos and eyes

If subfunctionalization occurred after gene duplication, with retinal functions retained by eutherian *CRX* and embryonic roles retained by diverging ETCHbox genes, then the ancestral mammalian *CRX* gene would have had expression in eyes and in early embryos. If neofunctionalization occurred, with ETCHbox genes acquiring novel roles, the ancestral gene would be eye specific. To resolve between these alternatives we examined *CRX* gene expression in a metatherian.

We analysed single-cell RNA-seq data from early developmental stages of the fat-tailed dunnart (*Sminthopsis crassicaudata*), a marsupial in the family Dasyuridae. We found expression of *CRX* in the early embryo (figure 2), with mean TPM of 79.6 at 16-cell, 14.1 at 32-cell and 40.8 at 64-cell (electronic supplementary material, S1B). By contrast, comparable analysis of mouse and human whole embryo RNA-seq data revealed no expression in the blastocyst or earlier stages, when ETCHbox genes are specifically expressed [5] (figure 2). Using reverse transcriptase polymerase chain reaction (RT-PCR) we confirmed the expected, widely conserved, expression of *CRX* in the adult metatherian eye (electronic supplementary material, S2D). These data imply that the ancestral mammalian *CRX* gene was expressed in embryos and eyes, and in the eutherian lineage these properties were separated after gene duplication.

(c) Radical restructuring of transcriptomes by the metatherian CRX protein

We hypothesized that a homeodomain protein with diverse roles (in early embryo and eye) would be capable of regulating more target genes and influencing more regulatory networks than a homeodomain protein of highly specialized function. We expressed ectopically a *CRX* gene from a metatherian (*Monodelphis domestica*) and an eutherian mammal (*Homo sapiens*) in primary mouse embryonic fibroblast cells and used RNA-seq to determine effects on global transcriptome, relative to cells transfected with a control vector. Although this experiment does not mimic closely

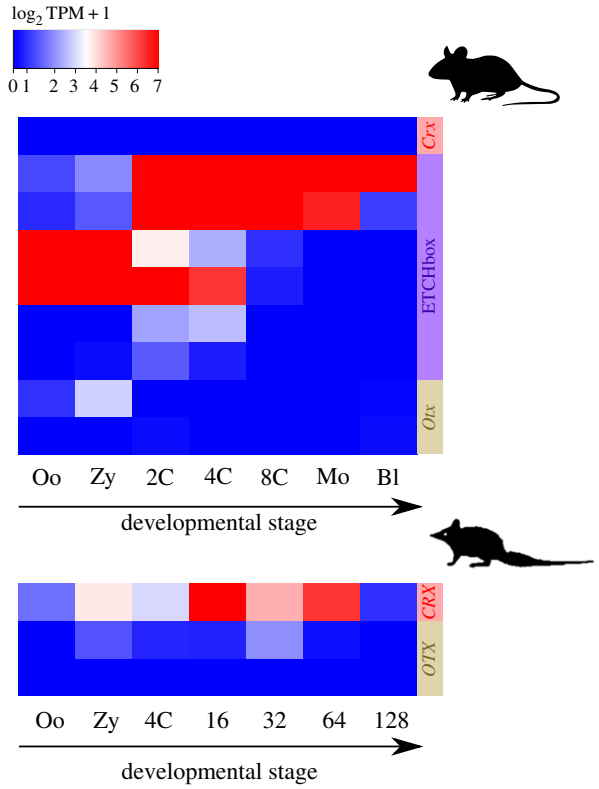


Figure 2. Comparison of eutherian (mouse, top) and metatherian (fat-tailed dunnart, bottom) *CRX* gene expression. Mouse ETCHbox genes, but not *Crx*, are expressed in early embryos. Dunnart *CRX* is expressed in early embryos. Expression levels represented as Log2 TPM + 1 (TPM in electronic supplementary material, S1B) Embryonic stages: Oo = oocyte, Zy = zygote, 2C = two cell, 4C = four cell, 8C = eight cell, Mo = morula, Bl = blastocyst, 16 = 16 cell, 32 = 32 cell, 64 = 64 cell, 128 = 128 cell. Embryonic stages not necessarily equivalent between species. ETCHbox genes from top to bottom: *Crxos*, *Oboxa1*, *Oboxa4*, *Oboxa7*, *Oboxb2*, *Oboxd2*. Otx genes from top to bottom: *Otx1*, *Otx2*.

the *in vivo* situation, previous studies have found that ectopic expression of related homeobox genes in cultured cells can drive biologically relevant transcriptomic changes [5,9]. Immunocytochemistry confirmed that both transfected genes are translated into protein (electronic supplementary material, S2E). Following overexpression, the two *CRX* orthologues had different global effects (figure 3). Ectopic expression of metatherian *CRX* resulted in differential expression of 1496 genes (661 up, 835 down; FPKM > 2, $p < 0.05$; electronic supplementary material, S1C). Ectopic expression of eutherian *CRX* resulted in differential expression of only 580 genes (118 up, 462 down; electronic supplementary material, S1C). There was partial response overlap with 61 common genes upregulated in the two experimental conditions and 278 downregulated. The discrepancy in the number of genes differentially regulated is not due to different levels of ectopic gene expression (electronic supplementary material, S2F) or variations in methodology, as samples were prepared in pairs and associated in the analysis. Gene Ontology (GO) term enrichment analysis showed that genes downregulated following eutherian CRX expression are enriched for cell cycle functions (electronic supplementary material, S1D), whereas those downregulated following metatherian CRX expression are enriched for terms relating to the extracellular matrix (electronic supplementary material, S1E). We conclude that a metatherian CRX

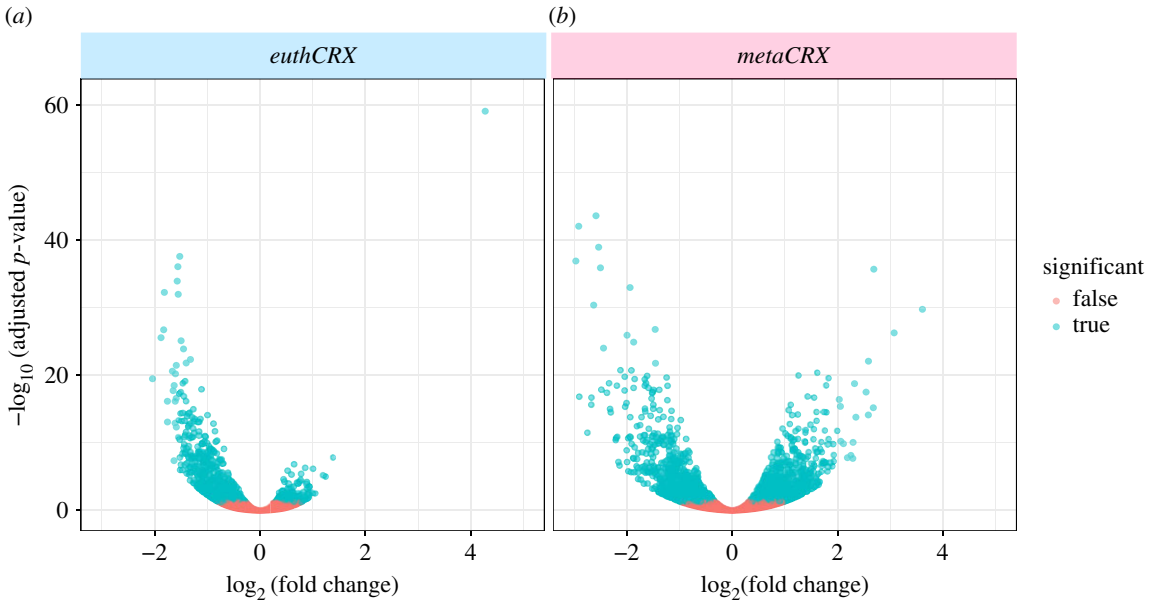


Figure 3. Effect on transcriptome of *CRX* ectopic gene expression. Metatherian *CRX* (*metaCRX*; (*b*)) induces more changes in gene expression than eutherian *CRX* (*euthCRX*; (*a*)) following ectopic expression. Both genes have more downregulatory effects than upregulatory effects. Volcano plots of −log10(adjusted *p*-value) and log2 fold-change following differential gene expression analysis. (Online version in colour.)

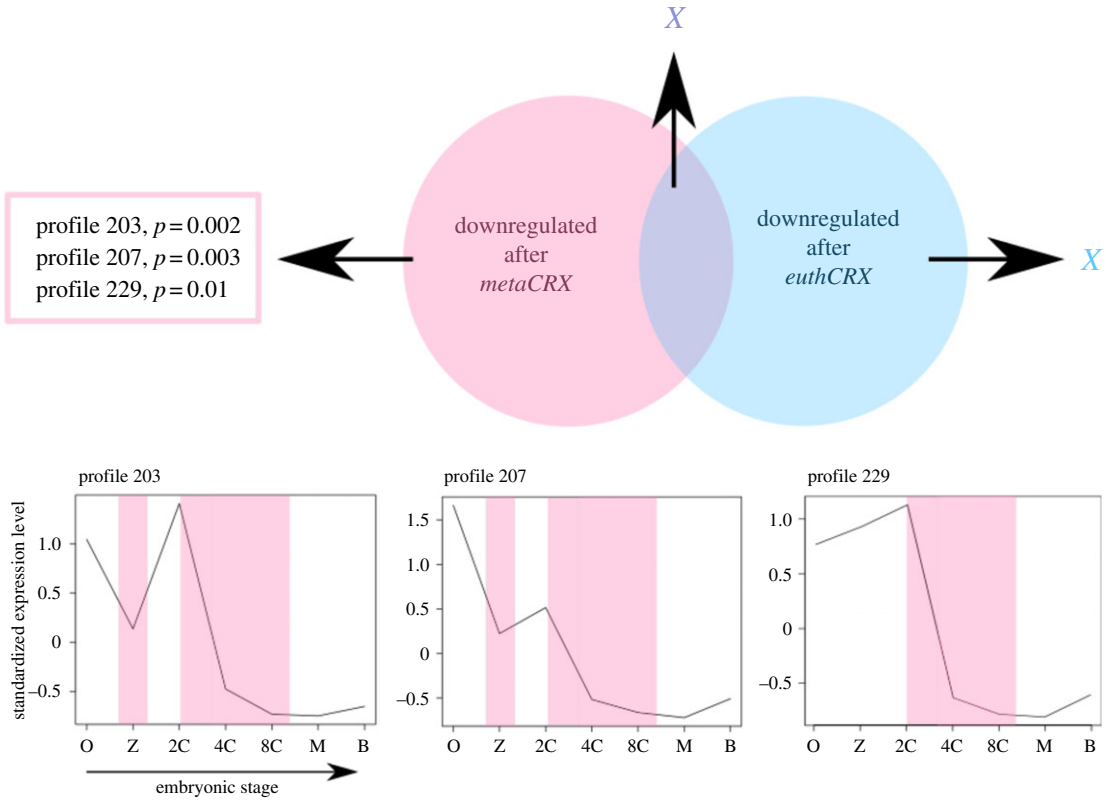


Figure 4. Mouse preimplantation clusters significantly regulated by metatherian *CRX* ectopic expression. Coloured strips identify common features of the clusters which we propose are related to the endogenous action of metatherian *CRX*. (Online version in colour.)

homeodomain protein elicits more downstream changes after ectopic expression than a eutherian CRX protein in a mouse fibroblast context.

(d) Metatherian CRX protein regulates genes expressed in preimplantation development

To assess if transcriptomic changes elicited by CRX protein expression are associated with developmental functions, we first asked how many of the differentially regulated genes are normally expressed in mouse embryos. We compiled a database of genes expressed in preimplantation mouse embryos [9] and found the transcriptomic response to metatherian *CRX* includes 732 genes that are dynamically expressed in normal preimplantation mouse development (349 up, 383 down, electronic supplementary material, S1F). Of the genes affected by ectopic eutherian *CRX*, only 277 are dynamically expressed in mouse preimplantation development (52 up, 225 down, electronic supplementary material, S1F).

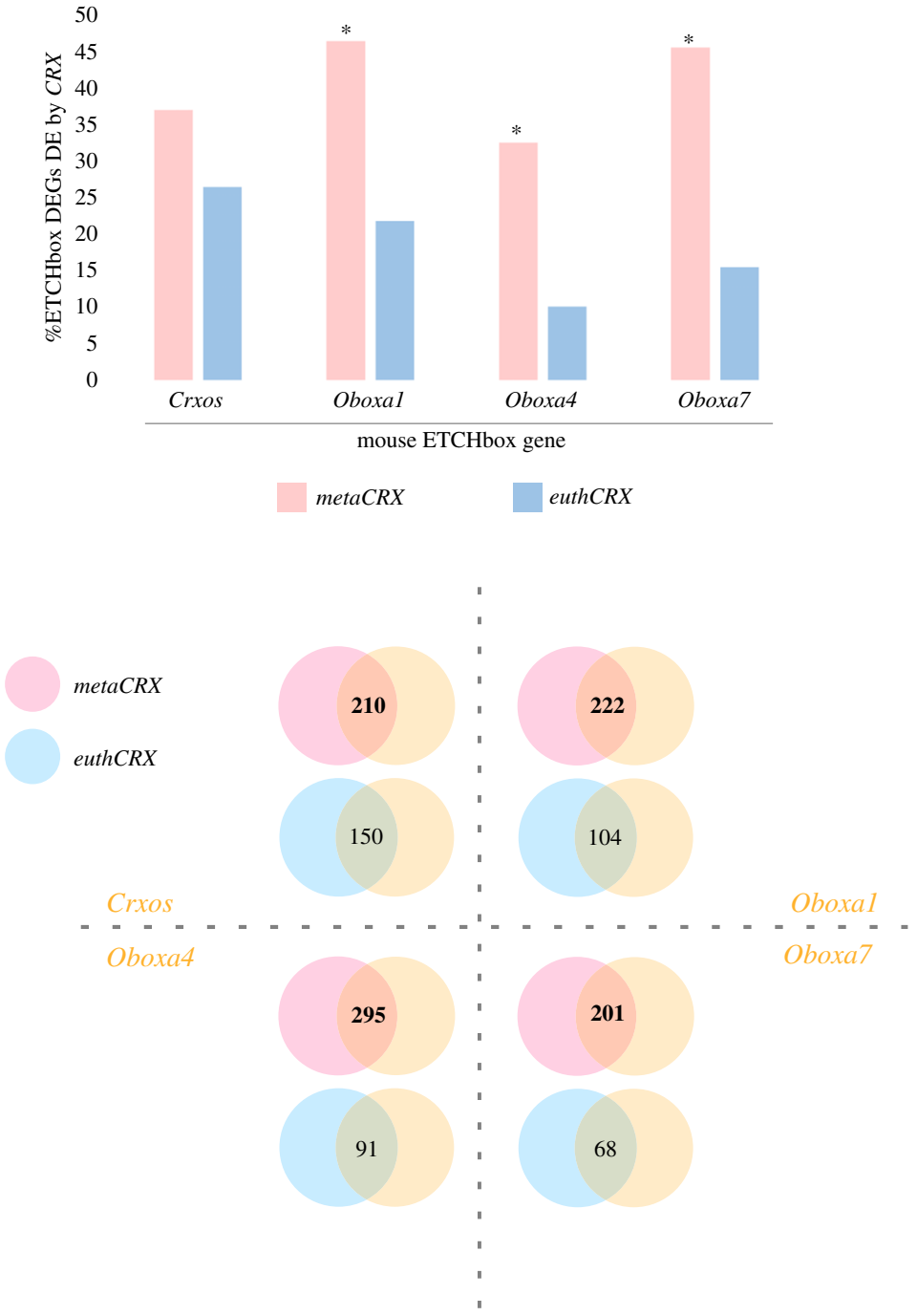


Figure 5. Gene sets downstream of mouse ETCHbox genes are more similar to gene sets downstream of metatherian *CRX* than eutherian *CRX*. Comparing lists of differentially regulated genes shows more genes are commonly regulated by metatherian *CRX* and each ETCHbox gene than by eutherian *CRX* and ETCHbox genes. (*a*) Percentage of DEGs following ETCHbox ectopic expression also regulated following *CRX* ectopic expression. Analyses show that the overlaps between Obox DEGs and *metaCRX* DEGs are stronger than the overlaps between Obox DEGs and *euthCRX* DEGs when the total sample size is taken into account (indicated by *). (*b*) Numbers of genes in the overlap between *metaCRX* and *euthCRX* with four mouse ETCHbox genes are shown by Venn diagrams.

Grouping developmentally expressed genes into sets based on their temporal expression profiles reduces the complexity of analysis and can help distinguish signal from noise [5,9,10]. We asked if the differentially expressed genes were enriched for genes that follow particular temporal profiles in development. If profile enrichment is found, this can increase confidence in the biological relevance of the transcriptomic response and provide clues to putative *in vivo* roles [5,9,10]. We used a clustering approach to assign all mouse genes to 85 different temporal expression profiles spannning early embryonic development (elctronic supplementary material, S2G), and tested if the lists of genes differentially regulated after *CRX* ectopic expression were enriched for particular temporal profiles.

We did not find profile enrichment in the lists of genes upregulated either by metatherian CRX or by eutherian CRX proteins. We also found no profile enrichment in the set of genes downregulated by eutherian *CRX* ectopic expression, or the common set of genes downregulated by both eutherian and metatherian CRX proteins (figure 4). However, we found significant enrichment for three temporal profiles in the list of genes downregulated by only metatherian *CRX* ectopic expression (profiles 203, 207 and 229; electronic supplementary material, S2G). The temporal gene expression

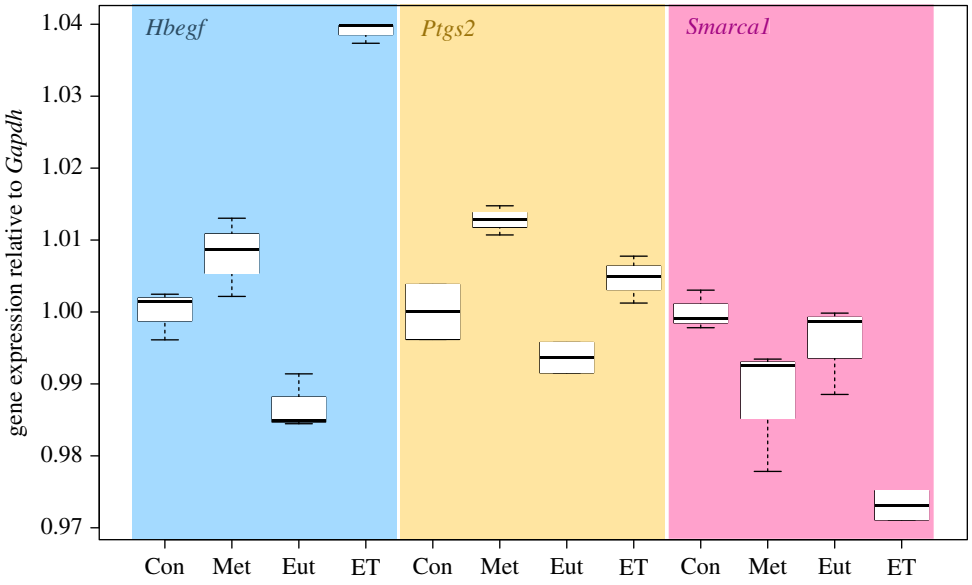


Figure 6. Genes with known roles in early eutherian development are similarly regulated following metatherian *CRX* and ETCHbox ectopic expression. A qPCR assay confirms that following metatherian *CRX* or *Oboxa1* ectopic expression, RNA levels of both *Hbegf* and *Ptgs2* are increased, but following eutherian *Crx* ectopic expression neither are upregulated. Similarly, following ETCHbox and metatherian *Crx*, but not eutherian *Crx*, ectopic expression Smarca1 expression decreases. Con = control, Met = metatherian *CRX*, Eut = eutherian *CRX*, ET = ETCHbox gene (*Oboxa1*). (Online version in colour.)

patterns encapsulated by these profiles differ in detail, but all three are composed of mouse genes showing a reduction in gene expression level between the two-cell and eight-cell stages of embryonic development (figure 4). We conclude that the metatherian CRX protein, but not the eutherian CRX protein, is capable of downregulating genes with common expression patterns in the early mouse embryo and is more likely to have *in vivo* functions at this period of development. Specifically, metatherian CRX can regulate a suite of genes that show declining expression in early mouse development. We propose that this activity reflects an *in vivo* role of the metatherian *CRX* gene: downregulating expression of specific genes before blastocyst formation.

(e) Common downstream effects of metatherian *CRX* and eutherian Obox genes

If the embryonic roles of eutherian ETCHbox genes arose initially by subfunctionalization, a prediction is that metatherian *CRX* should have comparable embryonic roles to ETCHbox. Specifically, we hypothesized that ectopic expression of metatherian CRX protein would elicit similar downstream effects to ectopic expression of ETCHbox genes. As ETCHbox genes subsequently underwent further duplication and diversification they may have gained or lost roles, but some similarities should remain. By contrast, if the embryonic functions of ETCHbox genes are a novelty that arose by neofunctionalization, expression of either metatherian or eutherian *CRX* would have equally dissimilar downstream effects to ETCHbox genes. To test these predictions, we took the lists of embryonic genes differentially regulated by ectopic *CRX* expression and compared these to similarly derived lists for mouse ETCHbox genes [9]. Figure 5*a* displays these comparisons as the percentage of ETCHbox responsive genes (differentially expressed genes, DEGs) that are also *CRX* responsive. These lists will include direct and indirect targets. Figure 5*b* shows the data displayed as absolute numbers of shared responsive genes. These comparisons show that, for all ETCHbox genes examined, the overlap in responsive genes is larger for metatherian *CRX* than for eutherian *CRX*. To test whether the latter conclusion is driven primarily by transcriptomic profile or is an artefact of sample size (eutherian *CRX* expression affects fewer genes globally than does metatherian *CRX*), we used Fisher's exact test. This showed that overlap between metatherian *CRX* and murine Obox genes is stronger than that for eutherian *CRX* even when the sample size is accounted for (figure 5*a*). We conclude that metatherian *CRX* affects the expression of a similar, but not identical, set of targets as do *Crxos* and Obox genes (mouse ETCHbox genes).

As metatherian *CRX* and mouse ETCHbox genes are capable of eliciting similar transcriptomic effects, we asked what embryonic functions might be the basis of the similarity. We focused on genes with well-characterized roles in early mammalian embryogenesis. Metatherian *CRX* ectopic expression leads to differential expression of genes with reported roles in early eutherian development including maternal communication (*Hbegf*), downregulation of the maternal inflammatory response (*Ptgs2*) and chromatin remodelling (*Smarca1*) among others (electronic supplementary material, S1G). Each of these genes was also regulated by one of more mouse ETCHbox genes, but none are regulated following eutherian *CRX* gene overexpression. Furthermore, we repeated transfections of all eutherian *CRX*, metatherian *CRX* and *Oboxa1* and treated samples in parallel. Transcriptome analysis by qPCR shows that *Hbegf*, *Ptgs2* and *Smarca1* are regulated by metatherian *CRX* and *Oboxa1* in the same direction, emphasizing the similarity of likely embryonic roles. Again, similar regulation following eutherian *CRX* expression is not observed (figure 6).

4. Discussion

The duplication, degeneration and complementation (DDC) model of gene evolution postulates that duplication of a pleiotropic gene leads to functional redundancy, followed

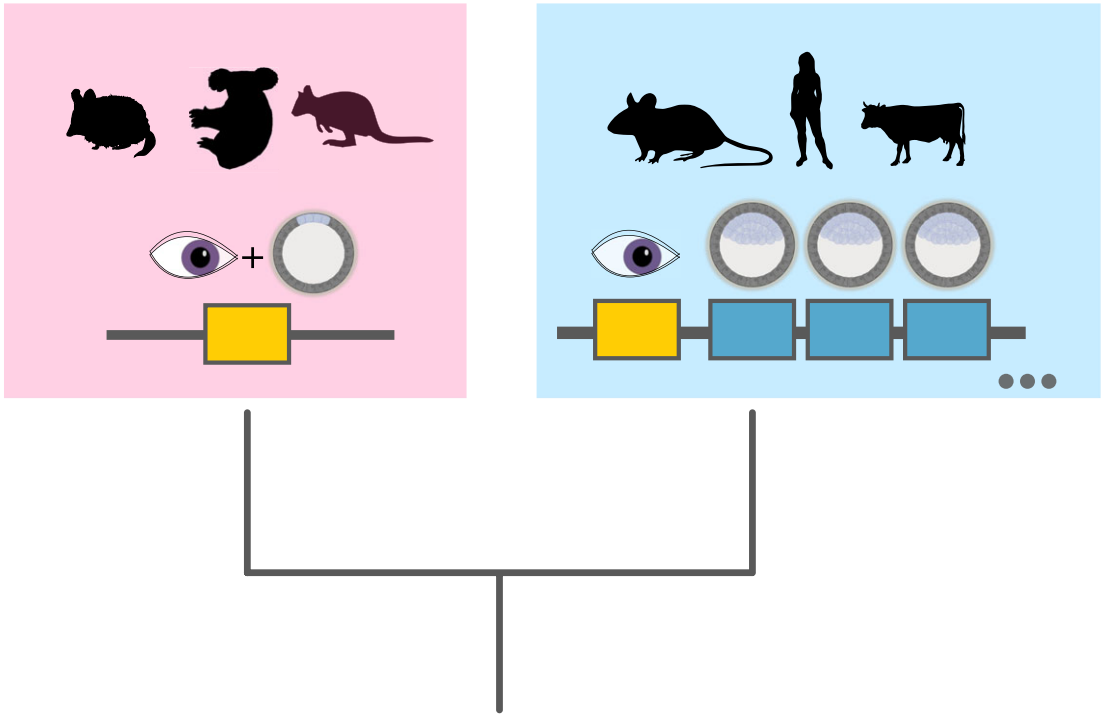


Figure 7. Duplication of *CRX* in eutherian lineage leads to specialization of duplicates into embryonic or retina-specific functions. In metatherians, *CRX* (yellow) continues to carry out both functions; in eutherians, embryonic roles are retained and elaborated on by ETCHbox genes (blue) while eutherian *CRX* retains functions in the eye. (Online version in colour.)

by stochastic mutation of different functional elements in each duplicate gene and distribution of ancestral gene functions between the duplicates [32]. This can be an intermediate step in the evolution of new or optimized gene functions, because if a pleiotropic gene is deployed for two or more roles in the same organism, it might not be possible for natural selection to optimize the gene for each role simultaneously. This resolution from compromise has been termed 'escape from adaptive conflict' [33]. Alternatively, when a new gene originates by a duplication event, it could acquire entirely new functions (or additional functions) through 'neofunctionalization'. Distinguishing between these alternatives is crucially important if we wish to understand the origin and evolution of biological traits. However, they are not always straightforward to distinguish as this needs comparable information from outgroups, ideally close outgroups.

There is currently much interest in the processes that underpin early development of humans, from fertilization and embryonic genome activation through to the determination of cell fates, as these events have practical relevance to stem cell biology and assisted reproductive technologies. The ETCHbox genes of humans and other eutherian mammals are expressed after embryonic genome activation and before cell fate determination and may be central to these processes [5,9,11,34]. We wished to understand how these gene functions arose in evolution as, paradoxically, their most closely related and parental gene (*CRX*) is expressed almost exclusively in the eye. Using marsupials (metatherians) as an outgroup, we found evidence that after the duplication of *CRX* in eutherians, subfunctionalization occurred. The *CRX* gene retained an ancestral role in the eye, while the highly divergent ETCHbox genes retained ancestral functions in the early embryo; our data argue that the single *CRX* gene of marsupials has dual properties, aligning with the DDC model (figure 7).

Subfunctionalization can occur through regulatory sequence evolution and/or coding sequence evolution. We find evidence for both. Changes in gene expression are indicative of regulatory evolution and it is striking that metatherian *CRX* is expressed in the very early stages of embryogenesis of the fat-tailed dunnart, as well as in adult eye, whereas in eutherian mammals *CRX* expression is limited to the eye. Coding sequence change is evident in the homeodomain of eutherian CRX, suggesting that sequence-specific DNA binding or transcription factor activity was altered in association with expression restriction to the eye. Changes to the encoded protein sequence clearly have functional relevance because when each is expressed in an ectopic situation, the eutherian and metatherian CRX proteins elicit different downstream effects. These differences could only be caused by protein sequence differences. When expressed ectopically in cultured cells, we found that a metatherian CRX protein caused a significant reduction in expression of sets of genes that would also normally fall in expression in the preimplantation embryo. Expressing eutherian CRX protein did not elicit the same transcriptomic response. These experiments also showed that the transcriptional effect of metatherian *CRX* ectopic expression has similarities to the effect of ETCHbox ectopic expression. For example, the ETCHbox gene *Oboxa1* and the metatherian *CRX* cause the same effect on the expression of candidate genes with known importance in both metatherian and eutherian embryo-maternal communication, *Hbegf* and *Ptgs2* [35], and on components of the SWI/SNF chromatin remodelling factor, *Smarca1*. We propose that metatherian *CRX*, and by extrapolation *CRX* of the common therian ancestor, functioned in the early embryo and the eye; following tandem gene duplication in eutherians, eye-associated functions were retained by one gene and embryonic functions were retained by a larger array of duplicates following

DDC. Paradoxically, these latter duplicates, the ETCHbox genes, diverged greatly in sequence despite apparently retaining some ancestral roles.

Under the classical DDC model, one gene with two functions duplicates and each duplicate takes on one of the functions and becomes specialized. However, in the case studied here, the fate of the duplicated genes has been more complex. The *CRX* gene duplication events in eutherians gave rise to many gene copies. One of the copies, referred to by the same name as the ancestral gene *CRX*, underwent a rather limited degree of sequence change (although this did alter protein function, as discussed above). The other gene copies underwent extreme sequence change, and are barely recognizable as duplicates of *CRX* [5]. These are the genes that inherited the embryonic functions from *CRX* in an ancestral mammal.

A key question is why eutherians deploy multiple *CRX*-derived genes in early development when metatherian development occurs with the expression of a single *CRX* orthologue? And why have these *CRX*-derived genes diverged so radically in protein sequence from the ancestral sequence and from each other? We do not have clear answers to these questions, but they may lie in differences in early embryonic development. For example, eutherians have an extended gestation compared with metatherians, in which the embryo implants into the maternal tissue through an invasive (though quite variable) placenta. This may require regulation of additional gene activities, not present in metatherians. In addition, there are morphological differences between the embryos themselves; metatherian embryos do not have an inner cell mass at the blastocyst stage, instead having embryo-fated cells as a pluriblast on the surface, nor do they undergo cellular compaction [24,36–38]. These distinctions are candidate explanations for asymmetrical gene divergence of *CRX* and ETCHbox genes in eutherians, and for selective retention of multiple duplicates. The variation in copy number between eutherian species might also relate to variation in placental structure, or it may be an evolutionary consequence of partial redundancy [9]. This reasoning implies that the evolution of these genes in eutherians probably involved more than the DDC mechanisms that we uncover here. In addition to ancestral pleiotropic roles being divided between the descendent genes, we propose that ETCHbox genes acquired additional functions to 'fine-tune' embryo development, and for additional embryo–maternal interactions that were not present in the ancestral *CRX* gene. Therefore, rather than straightforward subfunctionalization under a DDC model, we suggest that ETCHbox genes have secondarily become entwined in regulatory networks specific to eutherian embryos through specialization (figure 7). We term this subfunctionalization-specialization mechanism DDC+.

We conclude that following *CRX* duplication in the evolution of eutherian mammals, aspects of classical DDC-mediated subfunctionalization occurred. However the influence of additional gene duplication events and extreme sequence divergence allowed incorporation of additional functions, thus adding a component of neofunctionalization on top of subfunctionalization and the realization of DDC+.

Data accessibility. Raw sequencing datasets are accessible through the NCBI Gene Expression Omnibus (GSE129456). The mouse developmental stage RNA-seq datasets analysed in this study were obtained from NCBI SRA project SRP034543 [39].

Authors' contributions. A.H.R. and P.W.H.H. conceived and designed the study. A.H.R. performed sequence analyses, molecular and cell biology experiments, transcriptomics and statistical analyses of data. S.F. and A.J.P. carried out dunnart embryo sequencing and analysis. A.H.R. and P.W.H.H. wrote the manuscript and all authors contributed to and approved the final manuscript.

Competing interests. We declare we have no competing interests.

Funding. We acknowledge BBSRC funding to A.H.R. through the Oxford Interdisciplinary Bioscience Doctoral Training Partnership (BB/M011224/1) and Discovery Project funding (DP160103683) from the Australian Research Council to A.J.P. and SF.

Acknowledgements. We thank Rodrigo Pracana, Yichen Dai, Aris Katzourakis, Sounak Sahu, Aziz Aboobaker, Anish Dattani, Sonia Trigueros, Tim Davies and Paul Fairchild for discussions and advice.

References

1. Simeone A, Acampora D, Gulisano M, Stornaiuolo A, Boncinelli E. 1992 Nested expression domains of four homeobox genes in developing rostral brain. *Nature* **358**, 687–690. (doi:10.1038/358687a0)
2. Vignali R, Colombetti S, Lupo G, Zhang W, Stachel S, Harland RM, Barsacchi G. 2000 *Xotx5b*, a new member of the *Otx* gene family, may be involved in anterior and eye development in *Xenopus laevis*. *Mech. Dev.* **96**, 3–13. (doi:10.1016/S0925-4773(00)00367-1)
3. Plouhinec JL *et al.* 2003 The mammalian *Crx* genes are highly divergent representatives of the *Otx5* gene family, a gnathostome orthology class of *orthodenticle*-related homeogenes involved in the differentiation of retinal photoreceptors and circadian entrainment. *Mol. Biol. Evol.* **20**, 513–521. (doi:10.1093/molbev/msg085)
4. Germot A, Lecointre G, Plouhinec J-L, Le Mentec C, Girardot F, Mazan S. 2001 Structural evolution of *Otx* genes in craniates. *Mol. Biol. Evol.* **18**, 1668–1678. (doi:10.1093/oxfordjournals.molbev.a003955)
5. Maeso I *et al.* 2016 Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals. *BMC Biol.* **14**, 1–14. (doi:10.1186/s12915-016-0267-0)
6. Holland PWH, Marlétaz F, Maeso I, Dunwell TL, Paps J. 2017 New genes from old: asymmetric divergence of gene duplicates and the evolution of development. *Phil. Trans. R. Soc. B* **372**, 20150480. (doi:10.1098/rstb.2015.0480)
7. Booth HAF, Holland PWH. 2004 Eleven daughters of NANOG. *Genomics* **84**, 229–238. (doi:10.1016/j.ygeno.2004.02.014)
8. Villanueva-Cañas JL, Ruiz-Orera J, Agea MI, Gallo M, Andreu D. 2017 New genes and functional innovation in mammals. *Genome Biol. Evol.* **9**, 1886–1900. (doi:10.1093/gbe/evx136)
9. Royall AH, Maeso I, Dunwell TL, Holland PWH. 2018 Mouse *Obox* and *Crxos* modulate preimplantation transcriptional profiles revealing similarity between paralogous mouse and human homeobox genes. *Evodevo* **9**, 2. (doi:10.1186/s13227-018-0091-4)
10. Madissoon E, Marchetti L, Parlanti P, Landi S, Tonazzini I, Cecchini M, Piazza V, Gemmi M. 2016 Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. *Sci. Rep.* **6**, 1–14. (doi:10.1038/s41598-016-0001-8)
11. Töhönen V *et al.* 2015 Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. *Nat. Commun.* **6**, 1–9. (doi:10.1038/ncomms9207)
12. Saito K, Ogawa A, Toyofuku K, Hosoi Y, Soma M, Iha M, Kasuga K, Kojima I, Kobayashi M. 2011 Relationships between homeoprotein EGAM1C and the expression of the placental prolactin gene family in mouse placentae and trophoblast stem cells. *Reproduction* **141**, 259–268. (doi:10.1530/REP-10-0355)

13. Rajkovic A, Yan C, Yan W, Klysik M, Matzuk MM. 2002 *Obox*, a family of homeobox genes preferentially expressed in germ cells. *Genomics* **79**, 711–717. (doi:10.1006/geno.2002.6759)
14. Goldshmit Y, Galley S, Foo D, Sernagor E, Bourne JA. 2010 Anatomical changes in the primary visual cortex of the congenitally blind *Crx*-/- mouse. *Neuroscience* **166**, 886–898. (doi:10.1016/j.neuroscience.2009.12.039)
15. Andzelm MM *et al.* 2015 MEF2D drives photoreceptor development through a genome-wide competition for tissue-specific enhancers. *Neuron* **86**, 247–263. (doi:10.1016/j.neuron.2015.02.038)
16. Blakeley P *et al.* 2015 Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142**, 3613. (doi:10.1242/dev.131235)
17. Freund CL *et al.* 1997 Cone–rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gen (CRX) essential for maintenance of the photoreceptor. *Cell* **91**, 543–553. (doi:10.1016/S0092-8674(00)80440-7)
18. Wang L, Qi A, Pan H, Liu B, Feng J, Chen W, Wang B. 2018 A novel CRX frameshift mutation causing cone–rod dystrophy in a Chinese family: a case report. *Medicine* **97**, e11499. (doi:10.1097/MD.0000000000011499)
19. Viczian AS. 2003 *XOtx5b* and *XOtx2* regulate photoreceptor and bipolar fates in the *Xenopus retina*. *Development* **130**, 1281–1294. (doi:10.1242/dev.00343)
20. Shen YC, Raymond PA. 2004 Zebrafish cone–rod (*crx*) homeobox gene promotes retinogenesis. *Dev. Biol.* **269**, 237–251. (doi:10.1016/j.ydbio.2004.01.037)
21. Plouhinec JL, Leconte L, Sauka-Spengler T, Bovolenta P, Mazan S, Saule S. 2005 Comparative analysis of gnathostome Otx gene expression patterns in the developing eye: implications for the functional evolution of the multigene family. *Dev. Biol.* **278**, 560–575. (doi:10.1016/j.ydbio.2004.11.019)
22. Kuroda H, Hayata T, Eisaki A, Asashima M. 2000 Cloning a novel developmental regulating gene, *Xotx5*: its potential role in anterior formation in *Xenopus laevis*. *Dev. Growth Differ.* **42**, 87–93. (doi:10.1046/j.1440-169x.2000.00491.x)
23. Sauka-Spengler T, Baratte B, Shi D, Mazan S. 2001 Structure and expression of an Otx5-related gene in the dogfish *Scyliorhinus canicula*: evidence for a conserved role of *Otx5* and *Crx* genes in the specification of photoreceptors. *Dev. Genes Evol.* **211**, 533–544. (doi:10.1007/s00427-001-0191-2)
24. Renfree MB. 2010 Marsupials: placental mammals with a difference. *Placenta* **31**, S21–S26. (doi:10.1016/j.placenta.2009.12.023)
25. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T. 2012 FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* **40**, 580–584. (doi:10.1093/nar/gks498)
26. Yang Z. 2007 PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.
27. Bray NL, Pimentel H, Melsted P, Pachter L. 2016 Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527. (doi:10.1038/nbt.3519)
28. Kumar L, Futschik M. 2007 Bioinformation Mfuzz: a software package for soft clustering of microarray data. *Bioinformation* **2**, 5. (doi:10.6026/97320630002005)
29. Love MI, Huber W, Anders S. 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. (doi:10.1186/s13059-014-0550-8)
30. Huang DW, Sherman BT, Lempicki RA. 2009 Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57. (doi:10.1038/nprot.2008.211)
31. Huang DW, Sherman BT, Lempicki RA. 2009 Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13. (doi:10.1093/nar/gkn923)
32. Force A *et al.* 1999 Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545.
33. Hittinger CT, Carroll SB. 2007 Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449**, 677–681. (doi:10.1038/nature06151)
34. Boroviak T *et al.* 2018 Global transcriptome analysis reveals conserved and divergent features of primate preimplantation development. *Development* **145**, dev167833. (doi:10.1242/dev.167833)
35. Griffith OW, Chavan AR, Protopapas S, Maziarz J, Romero R, Wagner GP. 2017 Embryo implantation evolved from an ancestral inflammatory attachment reaction. *Proc. Natl Acad. Sci. USA* **114**, E6566–E6575. (doi:10.1073/pnas.1701129114)
36. Renfree MB, Shaw G. 2014 Embryo-endometrial interactions during early development after embryonic diapause in the marsupial tammar wallaby. **181**, 175–181. (doi:10.1387/ijdb.140059mr)
37. Frankenberg SR, de Barros FRO, Rossant J, Renfree MB. 2016 The mammalian blastocyst. *Wiley Interdiscip. Rev. Dev. Biol.* **5**, 210–232. (doi:10.1002/wdev.220)
38. Frankenberg S, Shaw G, Freyer C, Pask AJ, Renfree MB. 2013 Early cell lineage specification in a marsupial: a case for diverse mechanisms among mammals. *Development* **140**, 965–975. (doi:10.1242/dev.091629)
39. Fan X, Zhang X, Wu X, Guo H, Hu Y, Tang F, Huang Y. 2015 Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol.* **16**, 148. (doi:10.1186/s13059-015-0706-1)