Fast-evolving homeobox genes in mammalian preimplantation development



Thomas David Lewin

Wolfson College

Department of Biology

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary Term 2023

Declaration and Statement of Authorship

The work presented here was conducted by myself under the supervision of Prof. Peter W. H. Holland in the Department of Biology at the University of Oxford. Any contributions of other individuals are explicitly stated and material referenced from other sources has been clearly indicated and its origins cited.

The works presented in Chapters 2 and 3 have been published as first author original research articles in the *Journal of Molecular Evolution* (2021) and *Molecular Biology and Evolution* (2022), respectively. The work presented in Chapter 4 has been published on the preprint server *bioRxiv* (2023) and is under review at *Genome Biology and Evolution* as a research article with myself as first and corresponding author. All work on these manuscripts was completed during my time as a DPhil student at the University of Oxford. Detailed statements of authorship contributions are included at the beginning of each results chapter; where published manuscripts have multiple authors the specific contributions of each author are stated.

I hereby confirm that this thesis has not been submitted for any other qualification or degree at this university or any other institution.



Thomas D. Lewin
Wolfson College
26th February 2023

Acknowledgements

I have spent more time than I would like to admit trying to write something poetic that encapsulates the last four and a half years of my life. Given how painful those drafts were to read back, I will spare everyone else the self-indulgence. But, I will say that, when I was first accepted onto this programme, I remember looking in detail at what the process for repaying my scholarship would be if I were to drop out, because I absolutely was not convinced that I could make it to the end of a PhD. It is a testament to the people below that, almost since the moment I started, I have not worried about that possibility again.

The only place to begin is to thank my supervisor **Peter**, who is genuinely everything one could hope for in a supervisor and mentor. I was recently asked what qualities made a good supervisor, and realised halfway through my answer that I was essentially just describing Peter. The most frequent topic of conversation among PhD students tends to be complaints about our supervisors; it says a lot that the only complaint I have after over four years working with Peter is his penchant for organising lab meetings at times when free food would be available elsewhere. And to be fair, even then he didn't object to me turning up 5 minutes late with a handful of cake.

The lab has been small throughout my time here, but every single member has made it an absolute joy to be a part of, and I appreciate the contributions all of you have made to getting me to this point. In particular, **Amy, Carlos, Roddy, Serena** and **Sharon** were welcoming from minute one, and coming back after the pandemic to **Mayur, Peter** and **Yi-Jyun** made what could have been a tricky time incredibly enjoyable. I am especially grateful for Peter's guidance and advice on the bioinformatic strands of my work. A special thanks also to **Sonia** for facilitating my escape from the cold Oxford winter and introducing me to the joys of both Catalan culture and nanotechnology. Outside of our lab, there are so many other people who have made the Department an enjoyable place to spend time: **Josh, Manuel**, **Michael**, **Simon** and the whole of the **Richards Lab** especially, but there have been too many to name. Every person who I have stopped and chatted to has made my days and weeks better.

I am hugely appreciative of the people who have kept the labs running throughout my time in the Department of Zoology (and Biology), particularly **Will Volpato**, without whom the entire building would grind to a halt, and **Lee Fletcher**, who was always on hand to

assess the damage to whichever piece of equipment I'd had a disagreement with. I also appreciate everyone who has taken time out of their day to give me any kind of advice. **Aziz Aboobaker**, **Tim Davies, Ali Fouladi, Nacho Maeso** and **Seb Shimeld** immediately come to mind, but I am sure there are so many more. I must also thank **Lindsay Turnbull**, my tutor throughout my undergraduate degree, without whom I am pretty certain would not be in this position today; and **Carmen Sánchez-Cañizares**, who supervised my undergraduate dissertation and whose patience and encouragement was the perfect introduction to life as a research scientist.

I have been lucky enough to have shared my time in Oxford with so many wonderful people. To **Brittany**, **Char**, **Han**, **Joe**, **Kai**, **Lucy**, **Lucy**, **Michael**, **Nat** and **Tom**, thank you for making my first three years here so much fun. To **Alex**, **Atreyi**, **Brendan**, **Christian, Emily**, **Eve** and **Pippa**, it is a testament to you that, as the above people drifted away from Oxford, it has always continued to feel like home. In particular, I have to thank my various housemates for their both their company and for tolerating my laissez-faire approach to housekeeping. **Vic**, **Fran**, **Mark** and especially **George,** who was unlucky enough to serve a two-year sentence, it has been a pleasure.

There are several people who deserve an individual mention. To **Athena**, it was an absolute joy to spend my time with you, and I learned so much from you that I take with me to this day.

Querida **Cristina**, sin duda has sido lo más destacado de mis últimos dos años en Oxford, y tu apoyo en los últimos meses ha sido muy importante para mi (y, claro, *maravilloso*). He estado aquí demasiado tiempo y necesito irme, pero tú eres lo único que me hace querer quedarme.

And to my brother **Jack**, throughout this entire process you have been the one to keep me sane. No matter what is going on I know I can always rely on you, and I really don't think I can overstate how much you have helped me.

Finally, to my parents, **Clare** and **Phil**. Your unconditional love and support has provided the foundations for everything I have achieved and will ever achieve. You encapsulate everything I hope to be as a person. I dedicate this work to you.

Table of Contents

List of Figures VIII

List of Tables X

List of Publications XI

Thesis Abstract 1

1 General Introduction 2

1.1 Homeobox genes 2

1.2 Eutherian Totipotent Cell Homeobox (ETCHbox) genes 9

1.3 ETCHbox genes and cellular potency 23

1.4 Gene duplication 28

1.5 Mammalian preimplantation development 34

1.6 Transposable elements 41

1.7 Transcriptomics and RNA-sequencing 46

1.8 Thesis aims 49

2 Dynamic molecular evolution of mammalian homeobox genes: duplication, loss, divergence and gene conversion sculpt PRD class repertoires 52

Introduction to manuscript 52

2.1 Abstract 53

2.2 Introduction 54

2.3 Materials and methods 56

2.4 Results 62

2.5 Discussion 78

2.6 Conclusions 82

2.7 Acknowledgements and declarations 83

2.8 Supplementary figures 84

3 PRD-class homeobox genes in bovine early embryos: function, evolution and overlapping roles 111

Introduction to manuscript 111

3.1 Abstract 112

3.2 Introduction 113

3.3 Results 116

3.4 Discussion 128

3.5 Conclusions 134

3.6 Materials and methods 134

3.7 Acknowledgements and declarations 139

3.8 Supplementary figures 141

4 Rapid evolution of the embryonically-expressed homeobox gene *LEUTX* within primates 147

Introduction to manuscript 147

4.1 Abstract 148

4.2 Significance 148

4.3 Introduction 149

4.4 Results 151

4.5 Discussion 162

4.6 Conclusions 165

4.7 Materials and methods 166

4.8 Acknowledgements and declarations 169

4.9 Supplementary figures 170

5 ETCHbox proteins and transposable elements 177

Statement of author contributions 177

5.1 Abstract 178

5.2 Introduction 179

5.3 Aims 181

5.5 Materials and methods 182

5.5 Results 187

5.6 Discussion 199

5.7 Conclusions 203

5.8 Supplementary figures 204

6 Spatiotemporal expression of ETCHbox genes in the preimplantation embryo 208

Statement of author contributions 208

6.1 Abstract 209

6.2 Introduction 210

6.3 Aims 213

6.4 Materials and methods 214

6.5 Results 217

6.6 Discussion 245

6.7 Conclusions 251

6.8 Supplementary figures 253

7 General Discussion 270

7.1 Synopsis 270

7.2 Functional redundancy as an explanation for rapid evolutionary change 271

7.3 The lability of mammalian early embryonic development 273

7.4 The multifunctional nature of the ETCHbox group 277

7.5 Perspectives and future work 280
7.6 Concluding remarks 282

References **284**

Appendices **333**

A Abbreviations 333
B Additional files 336
C Resulting publications 337

List of Figures

1 General Introduction

1.1 Intron-exon structure of human ETCHbox genes and *CRX* 11
1.2 Birth of ETCHbox genes 12
1.3 ETCHbox gene repertoires of nine representative eutherians 13
1.4 Expression of ETCHbox genes in human tissues 15
1.5 Two hypotheses for the roles of ETCHbox genes in preimplantation development 16
1.6 The structure of the homeodomain 19
1.7 ETCHbox expression in mouse preimplantation development 20
1.8 Potency of blastomeres during mouse preimplantation development 24
1.9 Phylogeny of Mammalia 36
1.10 Preimplantation development in eutherians 38
1.11 Blastocysts of eutherians and marsupials 41
1.12 Transposable element classification and insertion mechanisms 42
1.13 Bulk RNA-seq data analysis pipelines 48

2 Dynamic molecular evolution of mammalian homeobox genes

2.1 ETCHbox repertoires of *Homo sapiens* (humans) and *Bos taurus* (cattle) 63
2.2 Bayesian gene tree of putatively functional ETCHbox genes 65
2.3 ETCHbox repertoires of 34 eutherians 67
2.4 Bayesian phylogenies of *TPRX1*, *TPRX2* and *TPRX3* 72
2.5 Sequence similarity between *TPRX1* and *TPRX2* genes within a species 73
2.6 Bayesian phylogenies inferred using (A) partition 1 and (B) partition 2 of putatively functional *TPRX1* and *TPRX2* genes 74
2.7 Models of ETCHbox homeodomain structures including sites under positive selection 77

3 PRD-class homeobox genes in bovine early embryos

3.1 ETCHbox gene repertoires of human, mouse and cattle 114
3.2 Immunocytochemistry analysis of BFF cells transfected with bovine ETCHbox genes 117
3.3 Transcriptional response to ETCHbox gene ectopic expression 119
3.4 Overlap between genes differentially expressed in response to ectopic expression of (A) bovine and human and (B) bovine and mouse ETCHbox genes 122
3.5 Temporal expression profiles for bovine genes enriched in the downstream target gene sets of ARGFX and LEUTX 124
3.6 Characterisation of a mutant bovine *ARGFX* allele 127
3.7 ETCHbox target gene expression profiles in human and bovine 131

4 Rapid evolution of the embryonically-expressed homeobox gene *LEUTX* within primates

4.1 Phylogenetic analysis of primate LEUTX sequences 152
4.2 Evolution of LEUTX within primates 153
4.3 Positive selection in LEUTX homeodomains 155
4.4 LEUTX expression in primate preimplantation embryos 158
4.5 Ectopic expression of human and marmoset *LEUTX* genes 159
4.6 Differing transcriptional responses to human and marmoset LEUTX expression 161

5 ETCHbox proteins and transposable elements

5.1 Retrotransposition assay construct 184
5.2 Human ETCHbox genes regulate genes that repress transposons 189
5.3 Bovine ETCHbox genes regulate genes that repress transposons 192
5.4 Effect of TPRX1 on LINE-1 retrotransposition 195
5.5 Effect of ETCHbox proteins on TE expression 197
5.6 TE enrichment at ETCHbox loci 198

6 Spatiotemporal expression of ETCHbox genes in the preimplantation embryo

6.1 Preimplantation embryonic development in humans 210
6.2 Clustering of 1,525 human preimplantation embryo cells based on their transcriptome 219
6.3 ETCHbox gene expression in human single-cell RNA-seq data 221
6.4 Expression of *ARGFX* and key ICM and epiblast marker genes 223
6.5 Clustering of human morula and transitory cells 224
6.6 Lineage-specificity of genes regulated by ARGFX 226
6.7 GRN reconstruction reveals DUXA as an activator of *ARGFX* and *LEUTX* 228
6.8 Clustering of porcine preimplantation embryo cells based on their transcriptome 230
6.9 ETCHbox expression in porcine scRNA-seq data 232
6.10 Lineage specification in the porcine morula (Dataset 2) 236
6.11 Clustering of mouse preimplantation embryo cells based on their transcriptome 237
6.12 Expression of *Obox* genes in scRNA-seq data 239
6.13 Expression of *OboxD* genes in scRNA-seq data using uniquely mapped reads only 241
6.14 Expression profiles of Kallisto ECs containing multiple *OboxD* genes 243
6.15 Expression of *Crxos* and *OboxA1* genes in the mouse preimplantation embryo 244
6.16 Expression of ETCHbox genes in human and pig embryos 247
6.17 Cascade of ETCHbox activation in the preimplantation embryo 249
6.18 An update to the model for ETCHbox function in the human preimplantation embryo 250

7 General Discussion

7.1 The hourglass model of embryonic development 275

List of Tables

2 Dynamic molecular evolution of mammalian homeobox genes

2.1 Duplication and loss of ETCHbox genes 70

3 PRD-class homeobox genes in bovine early embryos

3.1 Interferon-related GO terms enriched in *ARGFX* and *LEUTX*-downregulated gene sets 132

5 ETCHbox proteins and transposable elements

5.1 List of conditions for retrotransposition assay 186

5.2 GO biological process terms related to the antiviral response which are enriched in ETCHbox DE gene sets. 193

List of Publications

Work for the following chapters has resulted in publications:

Chapter 2

Lewin TD, Royall AH, Holland PWH (2021). Dynamic molecular evolution of mammalian homeobox genes: duplication, loss, divergence and gene conversion sculpt PRD-class repertoires. *J Mol Evol*. 89(6):396-414.

Chapter 3

Lewin TD, Fouladi-Nashta AA, Holland PWH (2022). PRD-class homeobox genes in bovine early embryos: function, evolution, and overlapping roles. *Mol Biol Evol.* 39(5):msac098.

Work for the following chapter has been published on the preprint server *bioRxiv*. It has also been submitted and is under review at *Genome Biology and Evolution*:

Chapter 4

Lewin TD, Blagrove JR, Holland PWH (2023). Rapid evolution of the embryonically-expressed homeobox gene *LEUTX* within primates. *bioRxiv*. 2023.02.02.526829.

Darwin Tree of Life Genome Notes

In addition, during the course of my DPhil, I also contributed to the following 'Genome Note' publications for species sequenced as part of the Darwin Tree of Life Project (https://www.darwintreeoflife.org/):

Boyes D, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, **Lewin TD**, Darwin Tree of Life Consortium (2022). The genome sequence of the acorn piercer, *Pammene fasciana* (Linnaeus, 1761). *Wellcome Open Res*. 7:258

Boyes D, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, **Lewin TD**, Darwin Tree of Life Consortium (2022). The

genome sequence of the mottled umber, *Erannis defoliaria* (Clerk, 1759). *Wellcome Open Res*. 7:208

Boyes D, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, **Lewin TD**, Darwin Tree of Life Consortium (2022). The genome sequence of the double-striped pug, *Gymnoscelis rufifasciata* (Haworth, 1809). *Wellcome Open Res.* 7:135

Boyes D, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, **Lewin TD**, Darwin Tree of Life Consortium (2022). The genome sequence of the pebble prominent, *Notodonta ziczac* (Linnaeus, 1758). *Wellcome Open Res.* 7:111

Thesis Abstract

Transcription factor proteins containing the homeodomain motif orchestrate myriad key functions during embryonic development and are therefore often conserved to a spectacular degree over extensive evolutionary timescales. Given this paradigm, the recent discovery of several homeobox families with roles in embryogenesis but which appear to be rapidly-evolving is a puzzling development. In this thesis I focus on one such group, the mammalian-specific Eutherian Totipotent Cell Homeobox (ETCHbox) genes, with the overarching aim of advancing our understanding of the function fast-evolving genes perform in the embryo and the forces that have combined to produce their unusual evolutionary trajectories. Analysis of single-cell RNA-sequencing data finds that ETCHbox genes are activated during the major wave of embryonic genome activation across various mammalian species. Ectopic expression experiments of ETCHbox proteins in cell culture suggest them to be multifunctional regulators of several key processes during preimplantation development, including blastocyst formation; inner cell mass development; the activity of transposable elements; and cellular potency. From an evolutionary perspective, dramatic changes in ETCHbox protein-coding sequences and copy number have occurred between lineages, and this has been driven at least in part by positive selection. Rapid protein-coding sequence evolution has resulted in small alterations to gene functions between closely related species (e.g. within primates) but highly divergent transcription factor functions between humans and cattle. Overall, I conclude that the lability of preimplantation development, combined with functional redundancy of ETCHbox proteins and positive selection acting on ETCHbox sequences, have combined to produce the diverse repertoires we observe today.

Chapter 1: General Introduction

1.1 Homeobox genes

1.1.1 The homeobox

Hox genes are so deeply embedded into the fabric of evolutionary developmental biology (or evo-devo) and their properties so familiar that one could be forgiven for overlooking just how astonishing their roles in development are. The Hox cluster is the largest gene cluster to exhibit the remarkable properties of spatial and temporal collinearity, that is the order of the genes on the chromosome matches their spatial expression in the embryo and the timing of their activation[1–5]. What is more, following Lewis' seminal work of 1978 which first identified spatial collinearity in the classical model *Drosophila*, it was shown that this had been conserved over more than 550 million years[6] of evolution since the divergence of fruit flies and mice[3,7–9]. Simultaneously in 1984, two groups uncovered a conserved motif shared between these homeotic genes and, it would transpire, many further non-Hox genes, which came to be known as the homeobox[10,11].

The apparent simplicity of the homeobox belies the transformative role it has played in the evolution of life on earth. It consists of a variable but recognisable sequence of usually 180 base pairs (bp) encoding a 60 amino acid residue globular homeodomain (HD) with the capability of sequence-specific DNA-binding. Homeodomain proteins act with cofactors to bind DNA and function as transcription factors (TFs), enhancing or repressing the expression of other genes. Though often thought of in the context of Metazoa, homeobox genes are present across the Eukaryota, where they coordinate numerous fundamental developmental processes[12,13]. While fungi and unicellular eukaryotes tend to possess a small complement of fewer than 12 homeobox genes, the genome of the plant model *Arabidopsis thaliana* contains over 100, and the two rounds (2R) of whole genome duplication (WGD) at the base of the vertebrates have resulted in

vertebrate-specific diversification, with repertoires of over 250 genes in mouse and human[12,14]. Homeobox gene numbers are not static in evolution, and gene families can be completely lost from lineages while gene duplications can also occur[15].

In the Bilateria, homeodomain proteins coordinate axial patterning; although it is for this function that they are renowned, they are also involved in the regulation of numerous further key developmental processes like cellular differentiation, proliferation, migration and adhesion[16–21]. For instance, relevant to this thesis, the first cell fate decisions in mammalian embryos are coordinated by homeodomain proteins such as CDX2, OCT4 and NANOG[22–24]. Moving from the micro to macro scale, expansions of homeobox gene repertoires prior to the Cambrian explosion are thought to have facilitated the evolution of the *Bauplan* of bilaterians and therefore the commencement of a high-energy, active lifestyle with directional locomotion; in doing so, they revolutionised the entire composition of the Earth's fauna[21,25]. Subsequent duplication and co-option of homeobox genes for lineage-specific functions has permitted further elaborations to this body plan, enabling the remarkable morphological diversity observed in bilaterians today[20,26]. A minor but informative example is the deployment of homeobox genes in the specification and patterning of mammalian-specific ectodermal appendages, such as hair, teeth and mammary glands[27].

1.1.2 The genetic toolkit paradigm

A substantial proportion of research in the field of evo-devo has been based upon the notion that bilaterians deploy a set of so-called 'toolkit' genes to produce their basic body plan, and that these genes, especially TFs at the 'top' of the network hierarchy, are highly conserved across large evolutionary timescales[28]. It has been argued that it is changes in the expression patterns of these genes caused by *cis*-regulatory modifications, rather than changes in the gene repertoires of species or alterations to the genes' sequences, that facilitates the diversity of forms and morphologies that exist across bilaterians[29,30].

Early studies identifying remarkable conservation of homeobox genes between vertebrates and invertebrates supported the idea that the molecular mechanisms underpinning development are conserved across disparate taxa, with highly conserved genes performing the same functional role in different species[31]. These were amongst the first demonstrations that at least some of the fundamental molecular mechanisms controlling animal development are conserved across species. Data since accumulated have been widely interpreted to support this idea, and it is of course true that many homeobox genes do show remarkable conservation across great evolutionary distances. Hox genes are an archetypal example, and this sequence conservation combined with their spectacular collinearity drew significant attention to the genetic toolkit paradigm. Furthermore, variation in the *expression* of Hox genes (as opposed to in their amino acid sequence) is indeed associated with modifications to anterior-posterior axis morphology[32,33], reinforcing the idea that gene regulatory change is a major driver of morphological diversity. Further well-known results associating regulatory changes with novel morphological features, for instance butterfly eyespots[34] and stickleback pelvic reduction[35,36], have provided additional support to this idea.

One legacy of this (justified) fascination with Hox genes and the idea of the molecular toolkit has been that the vast majority of work on homeobox genes (and developmental genes in general) has centred on examples that are highly conserved in sequence and copy number across extensive evolutionary timescales. Furthermore, candidate gene approaches based on discoveries in models like *Drosophila* and, as noted previously by Holland et al. (2017)[37], methods like PCR used to discover new genes, are inherently biased towards those with similar sequences, missing lineage-specific and fast-evolving examples. My point here is not to argue that this has been misguided or unfruitful in any way; quite the contrary, from this approach have emerged some of the most significant discoveries in the history of evo-devo. Instead, I emphasise the fact that, as a result of both technical biases and a sharp focus on the idea of a conserved genetic toolkit, fast-

evolving gene families had, until recently, largely been overlooked as major contributors to morphological diversity.

1.1.3 The neglected homeobox genes

Fast-evolving genes are those with an elevated rate of either DNA sequence change, copy number change (via duplications and losses) or both of these characteristics. Other properties, such as changes in the biological function of the gene, could also be included in such a definition. Rapid evolution of a gene is most clearly identifiable when contrasted to a closely related gene without these traits, as is the case when there is asymmetric divergence of gene duplicates. This occurs when, following duplication, the sequence and function of one daughter is conserved while the other diverges rapidly[38]. Such asymmetric evolution appears to be common after tandem duplication events, and is exciting because it suggests the possibility of the evolution of a novel functional role in the divergent copy[37].

As a result of the factors discussed above, homeobox genes exhibiting faster evolution and lacking conservation have, in the past, been considered less important to understanding the processes that govern the evolution of animal development and were subsequently largely overlooked. Since the turn of the century, and at an accelerating rate, fast-evolving genes are being exposed as playing critical roles in animal development, and underpinning many lineage-specific novelties. Intriguingly, despite their reputation as conserved mediators of the fundamental patterning stages of bilaterian development, several cases of spectacular asymmetric evolution have recently been identified in the homeobox superclass. The tide may be beginning to turn away from the classic examples of conserved homeobox gene function and towards these paradigm-defying groups.

Within the stem Ditrysia, a clade comprising the vast majority of Lepidoptera, four *Special Homeobox* genes (*ShxA* to *ShxD*) evolved as tandem duplicates of the Hox gene *Zerknüllt* (*zen*)[39,40]. The *Shx* genes' homeobox sequence subsequently diverged rapidly from *zen*, while the sequence of *zen* remained incredibly highly conserved. Importantly, these lineage-specific duplicates appear to have been recruited to key embryonic roles, marking the location of cells fated to become an extra-embryonic membrane called the serosa[40]. Moreover, further copy number evolution has occurred in specific Lepidoptera lineages, with *ShxD* being repeatedly lost and further tandem duplications present in 18 of the 123 Lepidoptera species sampled in one study[41]. *Apamea monoglypha* has an astonishing 165 tandem Shx loci[41] but the biological significance of the presence of such large tandem arrays remains unclear. Curiously, in contrast to other insect Hox genes, of which duplication is exceptionally rare[42], *zen* has been the subject of several other independent duplications too. One example is another case of asymmetric evolution which resulted in *Bicoid,* the famed anterior-specifying morphogen of cyclorrhaphan flies[43–45]. Overall, this suggests that homeobox genes such as *zen* can possess inherent properties making them more prone to spawning duplicates which are retained by selection.

The above case of fast-evolving homeobox genes contributing to early animal development is not a unique phenomenon. Spiralians, a group of protostomes exhibiting spiral cleavage and including phyla such as molluscs and annelids[46], possess a lineage-specific duplication of TALE class homeobox genes known as Spiralian TALE (SPILE), which display rapid sequence evolution[47,48]. The ancestral role of SPILE genes is in specifying quartets, the sets of micromeres produced by spiral cleavage[47], but the overall picture across Spiralians is more complex as repertoires of SPILE genes vary dramatically between species[48–50]. Bivalves use an unusual unequal cleavage mechanism, resulting in one abnormally large cell (the D blastomere) at the four-cell stage[51]. In addition to a conserved role in quartet patterning, genes in a bivalve-specific

expansion of the SPILE family have been recruited to specify the D quadrant[50]. This is a striking example of the recruitment of recent gene duplicates for a lineage-specific innovation. In contrast, cephalopods' loss of spiral cleavage has been accompanied by a complete loss of SPILE genes[48]. Overall, a lineage-specific group of homeobox genes underpins spiralian development, and divergence in gene copy number within the clade can be correlated with further developmental divergence.

Asymmetrically evolving homeobox genes are also involved in the preimplantation development of mammals. The ANTP-class gene *Nanog Neighbour* (*NANOGNB*) is a cryptic tandem duplicate of the classical pluripotency gene *NANOG*; while the duplication occurred in the ancestor of birds, reptiles and mammals, the gene has undergone lineage-specific rapid asymmetric evolution in mammals, diverging dramatically from the sequence of *NANOG*[52]. *NANOGNB* is expressed in a sharp peak at the 8-cell and morula stages in human and cow; while its embryonic function is unknown, ectopic expression experiments have shown it to repress a group of preimplantation embryo genes that peak at the 8-cell stage[52]. Interestingly, no *NANOGNB* duplicates have been reported in the literature, so its rapid evolution appears to be limited to sequence change.

1.1.4 Fast-evolving PRD-class homeobox genes in mammalian development

Lineage-specific genes are those which, within the context of a larger group, arose on the branch leading to a specific clade and are therefore unique to that taxon or taxa within that clade. Such genes are abundant in genomes but comparatively understudied; they are thought to be a significant driver of evolutionary innovations and lineage-specific adaptations, and in mammals they most often take on functions in reproduction or the immune response[53,54]. Curiously, several mammalian-specific gene families in the PRD-class of homeobox genes all show rapid evolution and have been recruited to functions in reproduction and embryonic development.

The eutherian-specific *Double Homeobox* (*DUX*) family is one of the most striking examples[55,56]. *Dux* is a pioneer factor in mouse zygotic genome activation (ZGA), being the critical activator of 2-cell stage genes, which it achieves largely via the binding of murine endogenous retrovirus (ERV) transposable elements (TEs) in their promoters[57–61]. Surprisingly, its human orthologue, *DUX4*, shows minimal sequence conservation, with only 33% identity of their first homeodomains[62], despite performing an apparently similar function in activating embryonic genome activation (EGA) genes[61,63,64]. Remarkably, this is caused by the divergence of ERV elements between mouse and human, and subsequent coevolutionary selection for the continued binding of these diverging elements by Dux/DUX4[65,66]. In addition to this, *DUX* genes also exhibit further dynamism, with gene losses especially frequent, and functional retrotransposed sequences[56,60,67]. The case of *DUX* demonstrates that fast-evolving PRD-class homeobox genes can possess critical functions in the developing embryo.

The *Reproductive Homeobox* (*RHOX*) genes, mammalian PRD-class genes which exist in a cluster on the X chromosome, offer another example of this phenomenon[68]. *RHOX* gene repertoires are highly divergent even between closely related species[69–71]. For instance, the rat genome houses 11 *Rhox* genes but species-specific expansions have resulted in a cluster of 33 genes in mouse[72]. In addition to copy number variation, *RHOX* paralogues have undergone selection-driven rapid divergence and show minimal sequence identity outside the HD[69], and strong natural selection has also caused orthologues to be highly divergent between species (e.g. only 63% identity between mouse and rat *Rhox5* HDs)[72–74]. Furthermore, two *RHOX* duplicates have undergone rapid asymmetric evolution away from a *RHOX*-like sequence and are now named *ARX* and *ESX1*[72]. Despite these rapid evolutionary changes, mouse and human *RHOX* genes are known to be involved in several important functions in developing and adult reproductive tissues and the early embryo[68,70,72].

A poorly characterised but promising candidate for a fast-evolving PRD-class gene with a role in mammalian preimplantation development is *Cytoplasmic Polyadenylated Homeobox* (*CPHX*). Little is known of its function or evolutionary history apart from its presence in multiple copies in human, mouse and rat; its expression in the preimplantation embryo; and the results of one ectopic expression experiment which found human CPHX1 and CPHX2 to be gene expression activators with overlapping functions[75–77]. Overall, there appears to be a repeated tendency for rapidly-evolving PRD-class homeobox genes to be recruited to roles in the very early development of mammals. Many of these genes are lineage-specific, and may therefore contribute to mammalian developmental novelties. In this thesis, I turn the focus on one such group of genes as a model to study this process, and I now explore the current knowledge about this group in greater detail.

1.2 Eutherian Totipotent Cell Homeobox genes

1.2.1 Discovery of ETCHbox genes

Cone-rod Homeobox (*CRX*) is a PRD-like homeobox gene[78] which in placental mammals is expressed in photoreceptor cells[79] and the pineal gland[80,81]. *CRX* is highly conserved across eutherians and mutations are associated with numerous clinical pathologies, including cone-rod dystrophy[82–84]. The non-mammalian orthologue of *CRX*, known as *OTX5,* has conserved functions in the pineal gland and eye, but is frequently co-opted to other lineage-specific functions such as the formation of the *Xenopus* cement gland[85,86]. *CRX*, it transpires, is a cryptic paralogue of *Orthodenticle Homeobox 1* (*OTX1*) and *OTX2*[79,85], two genes with critical functions in vertebrate head development[87] and a recently-discovered role in the preimplantation embryo[88], but has undergone dramatic divergence away from OTX-like sequences[85]. The ancestor of jawed vertebrates possessed *CRX* (*OTX5*), *OTX1* and *OTX2*[85], three paralogous genes generated by the 2R WGD at the base of vertebrates. These three genes are part of the PRD-like class,

encoding PRD-class homeodomains but without the paired box that is typical of PRD-class genes[14,89].

The term Eutherian Totipotent Cell Homeobox (ETCHbox) was coined in 2016 to describe a group of six PRD-like gene families that are duplicates of *CRX*[90]*: Arginine-Fifty Homeobox* (*ARGFX*), *Divergent Paired-Related Homeobox* (*DPRX*), *Leucine-Twenty Homeobox* (*LEUTX*), *Parent of ARGFX* (*PARGFX*), *Tetrapeptide Repeat Homeobox 1* (*TPRX1*) and *Tetrapeptide Repeat Homeobox 2* (*TPRX2*). *TPRX1* and *TPRX2* lay either side of *CRX* on human chromosome 19, forming a tight cluster; *DPRX* and *LEUTX* are a greater distance away on the same chromosome, and *PARGFX* is adjacent to *DPRX*. *ARGFX* has been translocated, and now sits on human chromosome 3. All ETCHbox genes encode a single 60 amino acid HD split across their final two exons, with the canonical PRD-like intron between HD residues 46 and 47 (Figure 1.1). ETCHbox genes are expressed as multiple isoforms, but each has a dominant (most highly expressed) isoform encoding a full HD[75].

ARGFX, DPRX, LEUTX and the two *TPRX* gene families were first described from the human genome sequence in 2007[78,91], and *PARGFX* was discovered almost ten years later once genomes of more diverse mammalian species were available[90]. Their names are drawn from the unique properties of their proteins: human *TPRX1* has a 4 amino acid (tetrapeptide) motif repeated imperfectly 58 times[78] and *LEUTX* has a leucine residue at the functionally-important homeodomain position 20, which is unique for PRD-class homeodomains in humans[78]. *ARGFX* was named after its possession of an unusual arginine residue at position 50 of the homeodomain[91]. This is notable as residue 50 is a key part of the homeodomain's specificity-determining recognition helix, and arginine is only found at this location in two other human proteins, neither of which are PRD-class[91].

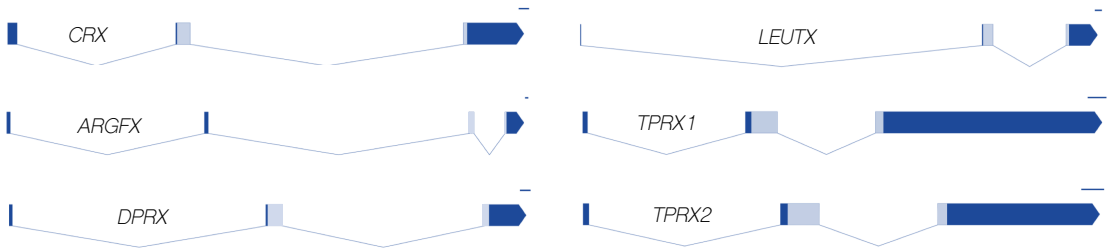


Figure 1.1 Intron-exon structure of human ETCHbox genes and *CRX*. For each gene, the most highly expressed transcript is shown. Solid blocks show exons and connecting lines represent introns. Homeoboxes are in pale blue, other coding sequence in dark blue. Untranslated regions are not shown. Scale bars at 3' end of genes = 100 bp. Figure based on data from Maeso et al. (2016)[90].

1.2.2 Evolution of ETCHbox genes

The evolution of ETCHbox genes can be considered to have occurred in two main phases. Firstly, all six ETCHbox genes arose as tandem duplicates of *CRX* in the stem lineage leading to eutherians after the divergence from metatherians, making them a synapomorphy of placental mammals; the placental ancestor is likely to have had one copy of each gene (Figure 1.2)[90]. These tandem duplication event(s) were then followed by extensive asymmetric evolution: while *CRX* remained highly constrained and its sequence invariable, ETCHbox sequences diverged rapidly from each other and that of their sister and exhibit remarkably low sequence identity with *CRX*[37,90]. In this first phase of evolution, it appears that ETCHbox genes were retained by selection due to subfunctionalisation via the duplication-degeneration-complementation (DDC) mechanism (see section 1.4.2). In metatherians, their 'ancestor' *CRX* is expressed in both the retina and in the preimplantation embryo[92]. In eutherians, subfunctionalisation has resulted in *CRX* possessing only a retinal function and ETCHbox proteins performing the ancestral *CRX* embryonic function[92]. Notably, ETCHbox proteins also appear to have gained further additional functions not possessed by the ancestral *CRX*, and so their mode of evolution has been labelled DDC+[92].

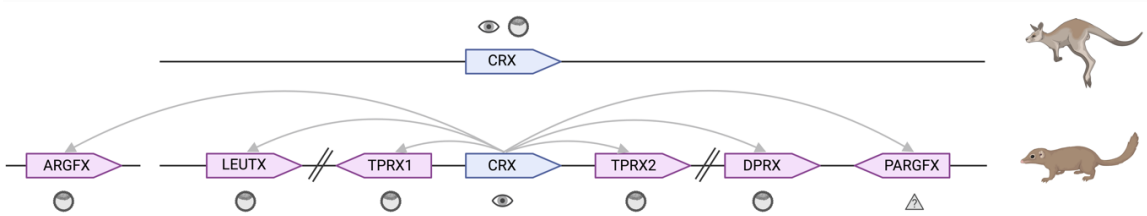


Figure 1.2 Birth of ETCHbox genes. Metatherian mammals (e.g. kangaroos) possess only *CRX* and no ETCHbox genes. In the stem lineage of eutherians, duplication of the *CRX* gene produced six ETCHbox genes. *TPRX1* and *TPRX2* form a tight cluster around *CRX*; *LEUTX*, *DPRX* and *PARGFX* form a loose cluster on the same chromosome. Double lines represent a large chromosomal distance. *ARGFX* is found on a separate chromosome. Gene 'arrows' show the direction of transcription. Metatherian *CRX* is expressed in the eye and the embryo; eutherian *CRX* is expressed only in the eye and ETCHbox genes are expressed in the embryo. The expression of location *PARGFX* is unconfirmed, but recent retrogenes suggest that it has germline expression. Figure based on data from Maeso et al. (2016)[90] and Royall et al. (2019)[92], and produced with BioRender.

The second phase of evolution has consisted of lineage-specific evolutionary events which have resulted in variation in the ETCHbox repertoires possessed by different eutherian species. Li and Holland (2010)[93] noted in characterising *ARGFX* across species that it has been completely lost or pseudogenised in evolution on numerous independent occasions. While some of the orthologues annotated as pseudogenes appear with updated genomic data likely to be functional, this work was key as it was the first to allude to the idea of ETCHbox genes as evolving rapidly. Further work by Maeso et al. (2016)[90] has suggested that indeed all six ETCHbox genes display unexpected dynamism between species, being duplicated and lost at a rate that is particularly unusual for homeobox genes (Figure 1.3). These authors identified no species possessing a representative of all six ancestral gene families, and several extraordinary ETCHbox locus architectures were unearthed. Humans have a repertoire relatively similar to the predicted ancestral state, with one copy of each gene except *PARGFX*. In addition, retrotransposition of one of the *TPRX* genes spawned another potentially functional gene, *TPRXL*[91]. *TPRXL* has undergone dramatic changes to gene structure and divergence away from the sequence and structure of typical ETCHbox loci and until the last few months was considered to be non-functional[91,94]; as such, in concordance

with previous work[90], it is not studied as part of this thesis. In contrast to the relatively simple evolutionary history of the human ETCHbox cluster, mice have lost completely *ARGFX*, *DPRX* and *PARGFX* while *LEUTX* is a highly degraded pseudogene[90,95]. *TPRX1* and *TPRX2*, meanwhile, are present and have been tandemly duplicated, with two *TPRX1* copies and 66 *TPRX2* loci now annotated in the mouse genome[96]. In a study of nine eutherian species, duplications were also found of, for instance, guinea pig (*Cavia porcellus*) *LEUTX* (14 copies), bat (*Myotis lucifugus*) *TPRX2* (10 copies) and cow (*Bos taurus*) *TPRX2* (2 copies). Overall, the diversity of ETCHbox repertoires captured in a relatively shallow study suggests that drastic variation is likely to be present across the eutherian phylogeny, implying a lack of constraint on this aspect of ETCHbox evolution.

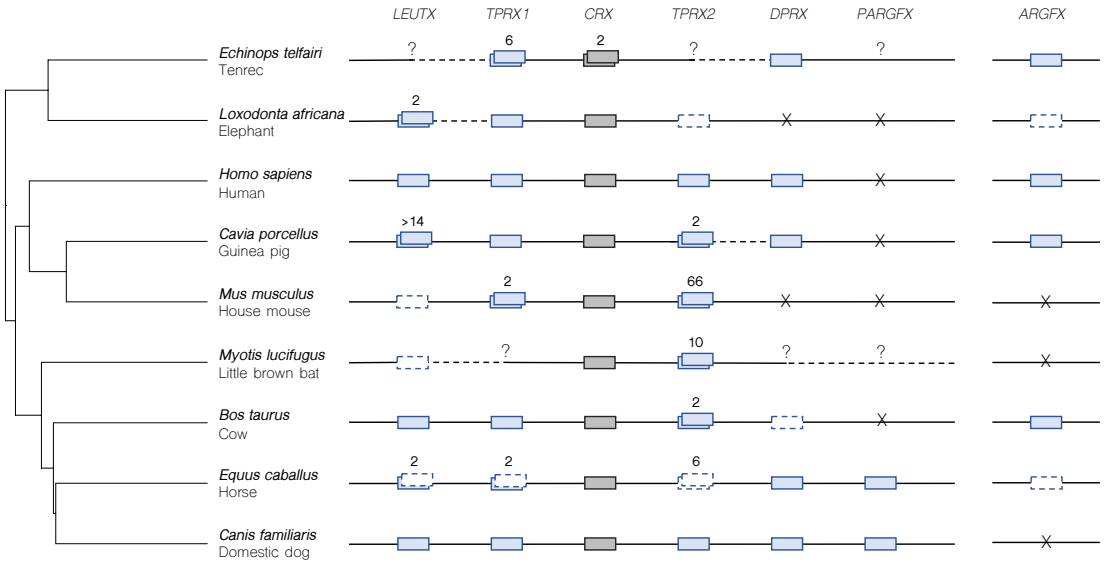


Figure 1.3 ETCHbox gene repertoires of nine representative eutherians. Blue rectangles represent putatively functional ETCHbox genes, with *CRX* in grey. White rectangles with dotted borders represent pseudogenes and an X shows complete gene loss. Double boxes indicate duplicated genes, with the number of duplicates shown above. Question marks indicate incomplete assembly in the genome region, and dotted lines indicate different scaffolds. Figure based on data from Maeso et al. (2016)[90] and Royall et al. (2018)[96].

This second phase of evolution, the divergence between lineages, consisted of not only changes in gene number but also further, rapid sequence changes. Evolution has been asymmetric, as *CRX* is highly conserved in sequence and copy number across eutherians, while gene trees show long branches leading to ETCHbox genes, and ETCHbox sequences show extensive variation between species[90,95]. While over 600 bp in *CRX* belongs to conserved elements, a maximum of 279 bp in any ETCHbox gene does so, and *TPRX1* and *TPRX2* have just 13 and 19 bp in conserved elements respectively[90]. This suggests that only small sections of ETCHbox sequences have been constrained by evolution. Indeed, mouse *TPRX1* and *TPRX2* have evolved so rapidly that their orthology was not initially recognised, and they were annotated as new gene families called *Crxos* and *Obox* respectively[90].

1.2.3 Are ETCHbox genes non-functional?

Though the evolutionarily labile nature of ETCHbox genes might lead one to suggest that they are non-functional and evolving by genetic drift alone, there is evidence that this is not the case. Firstly, the above analysis found that the recognition helix of the homeodomain was part of a conserved element at all human ETCHbox loci; this suggests that there has been selection to maintain sequence-specific DNA-binding capabilities[90], which would only be the case if there was a selectable function. More convincing evidence can be drawn from the genes' expression patterns. The first clue to their expression came from the presence of retrocopies (or 'processed pseudogenes'), intronless copies of genes produced by retrotransposition; these have been identified from all six ETCHbox genes, suggesting that all are expressed in the germline[90,91,97,98].

Subsequently, quantification of expression in RNA-seq data[90,99,100] identified a striking pattern: human ETCHbox genes are expressed only in the preimplantation embryo, peaking at the 8-cell stage and morula, and then never expressed again in the life of an individual (Figure 1.4). The transcription of all five genes is upregulated sharply at the 4-

cell to 8-cell stage and then quickly downregulated after just one or two cell divisions, though *DPRX* also shows low-level expression starting in the oocyte; this rapid downregulation is at least in part a product of epigenetic silencing by DNA methylation[75]. Moreover, this timing of expression was found to be conserved in bovine[90]; this conservation of expression pattern offers a further clue that there is constraint on ETCHbox evolution due to the possession of a selectable function.

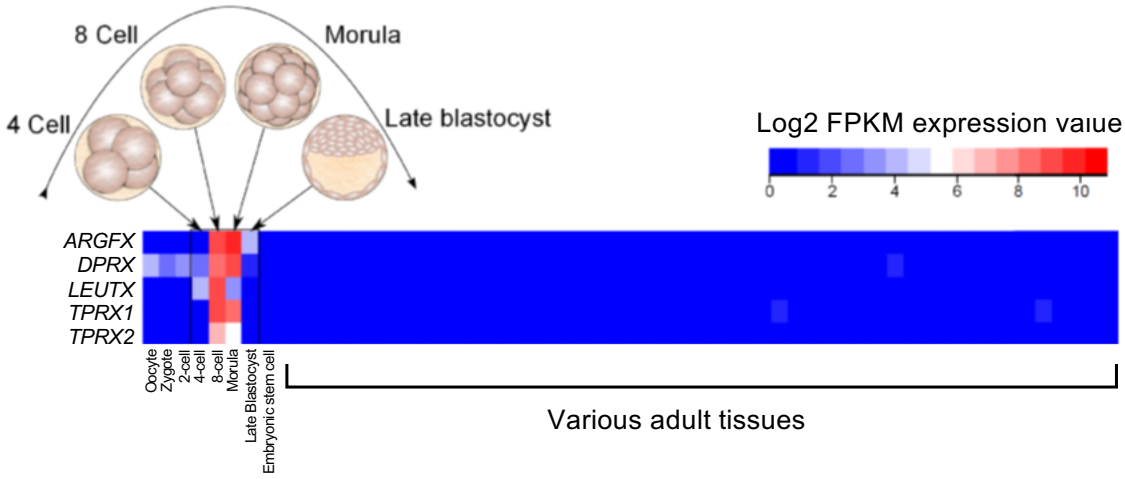


Figure 1.4 Expression of ETCHbox genes in human tissues. ETCHbox genes are only expressed in the preimplantation embryo and are most highly expressed at the 8-cell stage and in the morula. No expression is observed post-blastocyst. Figure adapted from Maeso et al. (2016)[90].

1.2.4 Function of ETCHbox proteins

ETCHbox genes are an example of the recruitment of lineage-specific homeobox genes to early development, but their functional role in the preimplantation embryo remains to be completely elucidated. Due to the developmental stage at which they are expressed, experimental manipulation in their native setting is highly challenging and, particularly in humans, limited by ethical considerations. In addition, until 2022[101–103] no known cell line expressed human ETCHbox genes to a level sufficient for meaningful knockdown or knockout experiments. As such, the approach largely used for functional characterisation has been ectopic expression in a cell type that does not naturally express the gene. In studies like this and in similar experiments throughout this thesis, the 'function' of a

transcription factor is taken to be the set of gene targets that it regulates, as determined by differential expression analysis. Maeso et al. (2016)[90] found, expressing human ETCHbox genes in human fibroblasts, that these transcription factors are effective at changing the expression of embryonically expressed genes in somatic cell culture. This result, including particularly the activation of embryonically expressed genes, validates this ectopic expression approach. Their most significant finding was that ARGFX acts antagonistically to LEUTX and TPRX1, whose roles overlap, to coordinate the expression of a group of genes expressed in a sharp pulse from the 8-cell to morula stages of human development (Figure 1.5). This is consistent with the ETCHbox temporal expression profiles reported above and, given the timing of this pulse of expression, it was suggested that it might be involved in the first lineage specification events of development, which occur around this time.

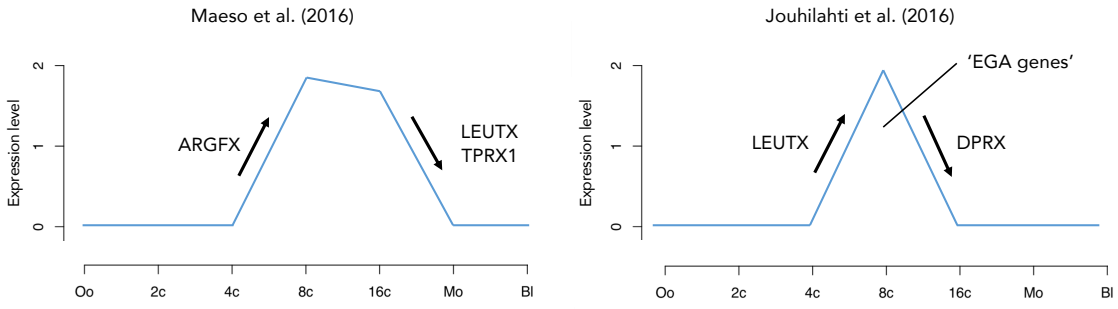


Figure 1.5 Two hypotheses for the roles of ETCHbox genes in preimplantation development. Left: temporal expression profile for genes upregulated in response to ARGFX expression and downregulated in response to LEUTX and TPRX1 expression when human genes were expressed in fibroblasts (data from Maeso et al. 2016[90]). Right: LEUTX upregulates and DPRX downregulates a group of genes activated at EGA (data from expression in embryonic stem cells, Jouhilahti et al. (2016)[104]). Abbreviations: Oo = oocyte; 2c = 2-cell; 4c = 4-cell; 8c = 8-cell; 16c = 16-cell; Mo = morula; Bl = early blastocyst.

Ectopic expression of ETCHbox genes in human embryonic stem cells found LEUTX to be an activator and DPRX to be a repressor of genes expressed during EGA, and it has been suggested that this may correspond to an in vivo upregulation of these genes at EGA and then subsequent downregulation[75,104]. Furthermore, ETCHbox genes have

been proposed to act on a 36 bp motif which is enriched in EGA genes' promoters[75,100]; LEUTX activates and DPRX represses a luciferase reporter with the motif, or even just an 11 bp core region of it, in its promoter[95,104]. It is interesting that the Maeso et al. (2016)[90] experiment found DPRX to make few transcriptional changes, but it caused the most differentially expressed genes of all PRD-like genes tested by Madissoon et al. (2016)[75]; this may be due to the different transcriptional environments (fibroblasts versus human embryonic stem cells). ARGFX was consistently found to be an activator in both studies, but Madissoon et al. did not find it to disproportionately affect genes expressed at any specific point of development.

Very recently, the first in vivo knockdown experiments of human ETCHbox genes were conducted. Zou et al. (2022)[94] used morpholinos to individually knock down *LEUTX*, *TPRX1*, *TPRX2* and *TPRXL*, and conduct double and triple knockdowns in human 3 pronuclei embryos. A *TPRX1/2/L* triple knockdown caused dramatic developmental delays at embryonic days 3 (8-cell stage) and 4 (morula) and the failure of activation of large swathes of EGA genes (31% of 8-cell-specific genes and 36% of morula-specific genes were expressed at significantly lower levels at days 3 and 4, respectively), while *TPRX1/2* double knockdown and single knockdowns showed similar but progressively less severe effects. Interestingly, the effects of *LEUTX* knockdown on both the progression of development and EGA were minimal. Overall, data suggest that ETCHbox genes function during the process of EGA, though the exact nature of their role remains to be fully elucidated.

1.2.5 ETCHbox homeodomains

The HD forms a helix-loop-helix-turn-helix structure with the capability of sequence-specific DNA-binding (Figure 1.6)[89,105]. The third alpha helix, also known as the recognition helix, binds to the major groove of the DNA double helix and is the main determinant of sequence specificity, while the N-terminal arm contacts the minor groove

and stabilises the protein-DNA interaction[105]. Residues at these sites, in addition to residues at further sites which contact the sugar-phosphate backbone and those constituting a structurally-important hydrophobic core, are critical to the proper functioning of any homeodomain, and are in many cases highly conserved across the homeobox superclass[89,105]. In reality, the specificity conferred by recognition helix-mediated HD-DNA binding is insufficient to perform the finely-tuned functions asked of homeobox genes, and interactions with cofactors are employed to further restrict specificity[106]; these interactions themselves may also be mediated by the HD[107], and can involve cooperative binding as dimers[108].

The DNA-binding properties of PRD-class HDs are well studied[107] but the relevance of these works to the divergent HDs of ETCHbox genes is uncertain. However, Katayama et al. (2018)[95] performed structural modelling of the human LEUTX protein and identified HD residues I47, K50, N51, A54 and R58 as specificity-determining and R2, R3 and R5 as contacting the minor groove. Position 50 is known to be critical to sequence-specific binding in PRD-class HDs, hence the switch to an arginine in *ARGFX* being particularly notable, and the N residue at position 51 is significant for being almost completely invariant across all HDs[89,109,110]. Interestingly, mutation assays showed that, while an A54V mutation, which is found in the Laurasiatheria, largely eradicates the activation function of LEUTX, this can be restored by a compensatory I47T mutation also present in Laurasiatheria[95]. This suggests a certain lability in the evolution of ETCHbox genes, even within the HD.

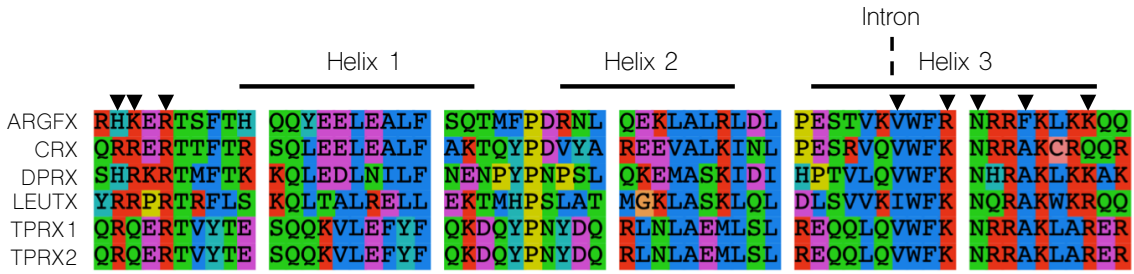


Figure 1.6 The structure of the homeodomain. Figure shows homeodomains of human ETCHbox proteins plus their sister CRX. Variation is present at many sites, while others remain highly conserved. Small black arrows mark key DNA-contacting residues (data from Katayama et al. 2018[95]) for LEUTX HD.

1.2.6 ETCHbox genes in mouse

More functional work has been completed on the ETCHbox genes of mice, though caution is urged in extrapolating conclusions to other species due to their abnormally divergent and highly duplicated ETCHbox complement and the species' more general developmental peculiarities[111]. Despite this divergence, the mouse ETCHbox genes *Crxos* (*TPRX1*, 2 copies) and *Obox* (*TPRX2*, 66 copies falling into four clades, A-D) are expressed in preimplantation development in a similar fashion to those of human and bovine, though with more variation (Figure 1.7)[96,112–114]. *Crxos* is expressed as three isoforms, one including both homeodomains (also known as Egam1), one including the N-terminal HD only (Egam1n), and one including the C-terminal HD only (Egam1c). Ectopic expression experiments in mouse embryonic fibroblasts showed that the C-terminal Crxos downregulated genes which peak in expression in the oocyte and upregulated genes peaking at the blastocyst[96]. The three Obox proteins tested (Oboxa1, Oboxa4, Oboxa7) yielded extensively overlapping differentially expressed gene sets, and all upregulated genes with a sharp peak at the 8-cell stage; Oboxa4 also appeared to repress EGA genes prior to EGA, consistent with its maternal expression pre-EGA[96]. Intriguingly, mouse Crxos and human ARGFX showed highly overlapping functions, suggesting that Crxos may have modified its function to take on the role of ARGFX prior to or following its loss[96].

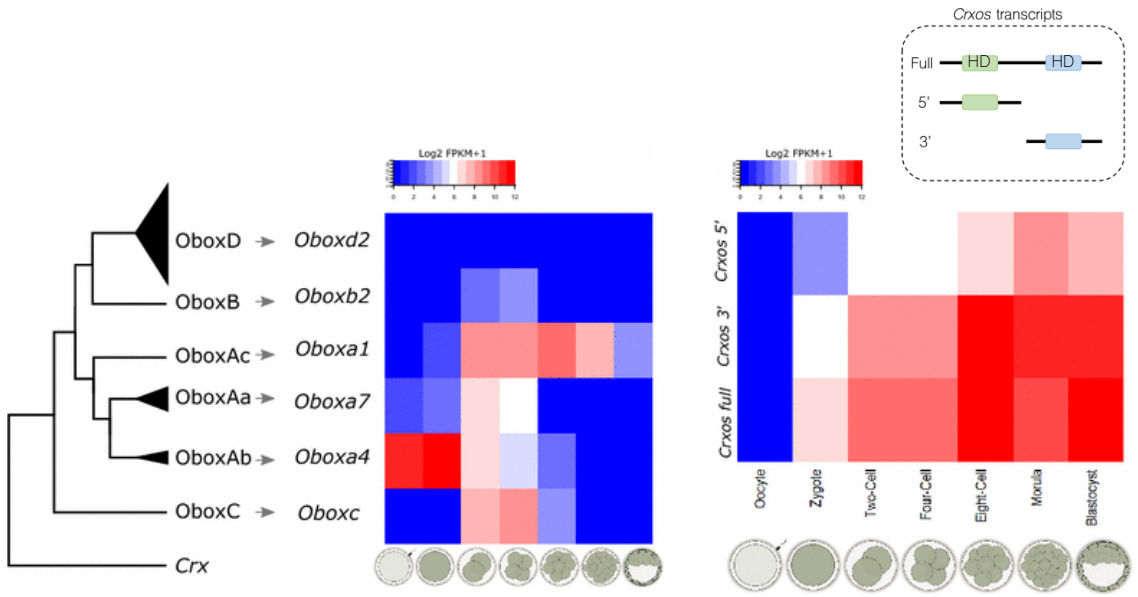


Figure 1.7 ETCHbox expression in mouse preimplantation development. Gene expression patterns from RNA-seq data of six *Obox* genes and three transcripts from the two *Crxos* loci: one with only the 5' homeobox, one with the 3' homeobox, and one with both homeoboxes. Figure adapted from Royall et al. (2018)[96].

Several further studies point to mouse ETCHbox genes as key factors in the preimplantation embryo. *Crxos* has been shown in knockdown experiments to be indispensable for normal blastocyst development and hatching[115] and *Obox* genes in particular have started to be identified as contributors to EGA. *Dux* is a key instigator of EGA in mouse[60,61], but knockdown does not arrest development[116–118]. A recent preprint[119] suggests that this is because *Obox4* is a potent activator of EGA genes and the EGA-active TEs MERVL and MERVK, and compensates for *Dux* loss due to partial redundancy; a double *Dux/Obox4* knockdown does arrest development at the 2-cell stage, with EGA genes remaining inactive. Restoring *Obox4* expression can rescue *Obox4*/*Dux* double knockdowns. Obox4 binds the promoters of over 50% of the genes bound by Dux, acting by guiding the deposition of active histone marks H3K4me3 and H3K27ac. It is of course worth remembering that, though named differently, *Obox* genes are homologous to *TPRX2* genes of other mammals, which were also recently shown to be involved in EGA in human[94].

Overall, mouse ETCHbox genes function in the preimplantation embryo, though these functions differ in their specifics from human ETCHbox genes. The evolutionary causes and functional consequences of the possession of such large arrays of ETCHbox genes is yet to be understood.

1.2.7 Hypotheses explaining rapid evolution of ETCHbox genes

The combination of properties possessed by ETCHbox genes is highly unusual. The existence of a set of homeobox genes with a conserved expression pattern at a critical point of development and apparent roles in EGA is at odds with the observed repeated gene losses and rapid sequence evolution. Several hypotheses have been put forward to explain this but no firm conclusion has currently been reached. Leading hypotheses are briefly summarised below.

Non-functionality: This hypothesis posits that ETCHbox genes evolve quickly because their function is not critical so mutation is not selected against. This is unlikely for the reasons outlined above.

Chromosomal location: This is a neutral hypothesis. With the exception of *ARGFX*, ETCHbox genes are located on human chromosome 19. Chromosome 19 is well-known for its dynamism and is home to the highest gene density and highest percentage of open chromatin of any chromosome; a high density of TEs; and an unusually elevated GC content[120–122]. Importantly, chromosome 19 also has a dramatically increased rate of gene duplications, possibly caused by its low density of recombination hotspots, with over 25% of the genes on the chromosome part of tandemly duplicated clusters[120,123]. Maeso et al. (2016)[90] suggested that the high levels of ETCHbox duplications and losses might be a product solely of the genes' position on chromosome 19; it is likely that this neutral process has contributed to the increased rate of ETCHbox duplicate production,

but the events following this must be mediated by selection. In addition, this explanation does not explain the rapid sequence change.

Mutational robustness: Mutational robustness occurs when a genetic change does not impact the functioning of a biological system[124]. Two mechanisms can underlie this: firstly, genetic functional redundancy arises when genes' functions are similar and overlapping[125,126]. This can result in relaxed selection and accelerated evolution if other genes compensate for a mutation. In support of a role for redundancy in ETCHbox evolution, multiple studies have reported overlaps in ETCHbox function, and a knockout study found that *Oboxa1*-null mice show no developmental phenotype but upregulation of other *Obox* paralogues, suggesting compensation[75,90,96]. The second mechanism is distributed robustness. This is a property of complex networks where, despite a lack of redundancy, perturbations are compensated by other parts of the network and thus have little effect on the network's output[124]. This means that mutations could occur in a gene regulatory network (GRN) such as those of which the ETCHbox genes are a part, but the network still performs as before and therefore mutations are not selected against, leading to fast evolution[96].

Placentation: ETCHbox proteins could be involved in placenta development. Placentas are very diverse between mammals[127,128], so ETCHbox genes could have evolved fast to accommodate this diversity[92]. Given the timing of ETCHbox expression, this theory would predict a role in the trophectoderm, the blastocyst lineage fated to become the placenta.

TE control: Despite being highly restricted throughout almost the entirety of an individual's lifespan, in mammals many TEs are briefly released from repression during preimplantation development at the time of ETCHbox expression[129–131]. If ETCHbox genes have been recruited to control TEs, an arms race with these elements may explain

their rapid sequence change. There are precedents for this: fast-evolving homeodomain proteins of the *RHOX* group repress LINE-1 retrotransposition during preimplantation development[132], and *DUX* sequences have diverged between mouse and human due to species-specific changes in their endogenous retrovirus targets[65].

Speciation: The production and care of sterile hybrid offspring reduces fitness. Post-zygotic selection for reproductive isolation may therefore select for incompatibilities that terminate hybrid embryos early to minimise resource wastage. Katayama et al. (2018)[95] proposed that selection could be acting on ETCHbox genes to promote divergence and result in early embryo termination of hybrid embryos due to ETCHbox gene incompatibilities. Water buffalo and domestic cow hybrid embryos fail at the 8-cell stage[133], the time of ETCHbox gene expression, lending some support to this hypothesis. MacLean and Wilkinson (2010)[72] postulated this same idea with respect to *RHOX* genes. This hypothesis is highly speculative, and there is no convincing evidence in support of it.

1.3 ETCHbox genes and cellular potency

1.3.1 Cellular potency

A cell's potency is its ability to differentiate into other cell types[134]. Maximum potency is a property of the zygote, which must be able to develop into a complete organism, and the potency of cells then becomes progressively restricted as development proceeds (Figure 1.8)[135]. In its broadest sense, totipotency is considered to be the ability of a cell to differentiate into all cell types of an organism, including embryonic and extraembryonic cells (e.g. as used by Suwińska et al. [2008][136]). However, the more stringent definition used in this thesis requires that a cell has the self-organisational capacity to initiate development and develop from a single cell into an entire fertile individual in order to be considered truly totipotent[134]. The term plenipotency, from the latin *plenus*, meaning 'full',

has been introduced to describe those cells which conform to the broad definition of totipotency, but cannot autonomously develop into a full organism[134]. The next step down the potency ladder is pluripotency. This is the ability of a cell to differentiate into cells of all three germ layers (ectoderm, mesoderm and endoderm), including all lineages required to form an adult individual, but not into extra-embryonic cells of the yolk sac and placenta[137]. After the first lineage specification event separates the inner cell mass (ICM) from the trophectoderm (TED), ICM cells are considered to be pluripotent[134,137].

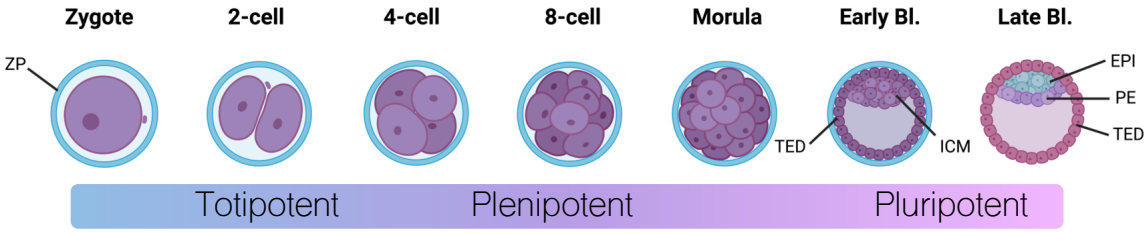


Figure 1.8 Potency of blastomeres during mouse preimplantation development. The zygote, 2-cell blastomeres and debatably 4-cell blastomeres are totipotent, 8-cell and morula blastomeres are plenipotent. The inner cell mass and epiblast of the blastocyst are pluripotent. Abbreviations: Bl = blastocyst, EPI = epiblast, PE = primitive endoderm, TED = trophectoderm, ZP = zona pellucida. Figure created with BioRender.

1.3.2 Pluripotent stem cells and the molecular basis of pluripotency

Stem cells (SCs) are cells that can self-renew but also have the potential to differentiate into at least one mature cell type[138]. Embryonic stem cells (ESCs) are pluripotent cell lines isolated from the ICM of blastocysts that are capable of differentiation but also possess indefinite proliferative potential in an undifferentiated state[139–144]. With reliable approaches for derivation and culture conditions optimised to support the long-term culture of purely undifferentiated colonies in what is known as the ground state[139,145–147], ESCs have become a workhorse for studies on potency and mammalian preimplantation development.

The revolutionary work of Takahashi and Yamanaka (2006)[148] and Takahashi et al. (2007)[149] demonstrated molecular reprogramming of mouse and human somatic cells,

respectively, into induced pluripotent stem cells (iPSCs) which highly resemble ESCs at a transcriptional, epigenetic and functional level[139,150,151]. The induction of these cells could be achieved by the addition of just four transcription factors: KLF4, MYC, OCT4 (encoded by *POU5F1*), and SOX2 (collectively referred to as OKSM or Yamanaka factors), demonstrating with clarity that pluripotency is a state induced by transcription factors. The efficiency of this process can be further enhanced, for example by adding pluripotency-associated factors such as TBX3[152,153].

In no small part due to the availability of pluripotent stem cells, significant progress has been made in understanding the molecular basis of pluripotency. OCT4 was the first pluripotency regulator to be identified[154,155]; OCT4-deletion embryos fail to form pluripotent blastomeres[23] and OCT4-deletion ESCs lose pluripotency and self-renewal capacity[156]. OCT4, alongside SOX2, with which it acts cooperatively[157–159], and the homeoprotein NANOG, together constitute a core trio of transcription factors that regulate the pluripotent state[160–163]. These three pluripotency factors cross-activate each other, and cooperate to activate genes that promote pluripotency and repress genes that are lineage-specific[160,161]. In addition, the roles of “outer” pluripotency factors including KLF2, TBX3, TCL1, STAT3 and ESRRB[164–168] are also at least partially understood, but the complexity of the network means that many details remain to be elucidated.

1.3.3 ETCHbox genes and pluripotency

Recent single-cell RNA-seq data showed that the mouse ETCHbox gene *Oboxa1* (alternatively *Obox6*) is activated as part of iPSC reprogramming[169]. Subsequent experiments found that ectopic expression of *Oboxa1* or *Oboxa4* (alternatively *Obox1*) during molecular reprogramming was able to improve the efficiency of iPSC generation and, remarkably, that the OKSM factor Sox2 could be completely replaced by Oboxa4[169,170]. Further work suggested that Crxos may also improve reprogramming efficiency and found that the reprogramming factor Klf4 binds to the *Crxos* locus[171].

Some results suggest that human ETCHbox genes are not expressed in ESCs[90], although trace expression of some genes, particularly *ARGFX*, has been reported[75,91,93]. However, clear expression of *Oboxa1*, *Obox4* and *Crxos* was detected in mouse ESCs[172,173]. Indeed, the N-terminal HD-only isoform of *Crxos* stimulates and is necessary for ESC renewal and prevents differentiation[114,173,174]. Similarly, the C-terminal HD-only *Crxos* isoform also increases expression of pluripotency genes such as *Nanog*, *Pou5f1*, *Sox2* and *Tbx3* in ESCs and stabilises the undifferentiated state, partially inhibiting differentiation[175]. Overall, it appears that murine ETCHbox genes are deeply integrated into the pluripotency network. One key question emerging from this is whether work using mice as a model, with their highly divergent repertoire of ETCHbox genes, is in any way informative as to the ancestral role of ETCHbox genes or their wider role across eutherians.

1.3.4 ETCHbox genes and totipotency

ESCs do not exist as homogenous populations, and cells in a single culture can have heterogeneous expression of key marker genes and varying potentials for differentiation[176–180]. In mouse naïve ESC populations, a small proportion of cells (around 0.5% at any given time) exhibit reactivation of EGA genes and show strong similarity in gene and transposon expression to totipotent 2-cell stage blastomeres, regaining the ability to form both embryonic and extraembryonic tissues[57]. These two-cell-like cells (2CLCs), a powerful resource for studying mouse EGA, completely lack the presence of classical pluripotency network proteins Nanog, Oct4 and Sox2[181] which are usually highly expressed in ESCs but show large upregulation of 2-cell-specific MERVL transposons and 2-cell genes such as the *Zscan4* cluster[57]. Several mouse *Obox* ETCHbox genes that are expressed at the two-cell stage, including *Oboxa3* (alternatively *Obox3*) and *Oboxa1* are upregulated in 2CLCs[57].

In the past year, three groups independently reported on the identification of what can be considered to be a human equivalent of these mouse 2CLCs, cells in human naïve ESC cultures that show strong similarity to 8-cell blastomeres (the time of EGA), so-called 8-cell-like cells (8CLCs)[101–103]. As in mouse, such cells are present as a small proportion of naïve ESC cultures[102], but their formation can be induced by growth in a specific cell culture medium[101,103]. These cells, which exhibit marked upregulation of EGA genes, 8-cell marker genes, totipotency genes and specific TEs that are expressed in the 8-cell embryo, show high transcriptional similarity to 8-cell blastomeres, and cluster with them in single-cell RNA-seq datasets[101–103].

Unsurprisingly given their expression peaks at the 8-cell stage in human embryos, the ETCHbox genes *ARGFX* and *LEUTX* are strongly upregulated in 8CLCs[101–103]. What is more interesting, however, is that *TPRX1* appeared as one of three key regulators of the 8CLC GRN, along with *DUXA* and *ZSCAN4*, and is in the top 5 8CLC-upregulated genes; indeed, *TPRX1* knockout prevents the transition to the 8CLC state[101]. This implicates *TPRX1* particularly as a key regulatory factor of the 8-cell totipotency network. Interestingly, ectopic expression of *LEUTX* in human ESCs was found to upregulate *DPPA3*, an important factor in 8CLC conversion[101,104,182]. Overall, it is clear that there is a significant entanglement of ETCHbox proteins in the networks regulating cellular potency in the preimplantation embryo.

The broad aim of this thesis is to shed light on both the functional roles and evolution of ETCHbox genes. In light of this, following the above introduction to ETCHbox genes themselves, it is beneficial to introduce several key concepts and ideas that appear throughout the work. The following sections provide brief explorations of gene duplication, mammalian preimplantation development, TEs, and transcriptomic analyses, with the aim of facilitating easier understanding of the following data chapters.

1.4 Gene duplication

1.4.1 The importance of gene duplication

An understanding of the mechanisms and consequences of gene duplication is essential to explaining ETCHbox evolution. While by no means the first to explore the role of gene duplication[183–186], Susumu Ohno used his seminal book, *Evolution by gene duplication*[187], to argue that the duplication of genes provides the raw material required for functional innovation, and was a key factor facilitating the evolution of complex organisms. The crux of Ohno's argument was that, if a single-copy gene was duplicated into two copies, one copy would be free of the strong "policeman" (Ohno, 1970, p. 59)[187] of purifying selection, and the random mutations that would subsequently be permitted may result in a novel function which is then selectively favoured. Ohno proposed that these new loci with novel functions are necessary for the evolution of increased complexity. In the following 50 years, the idea of gene duplication as providing the raw materials on which selection can act with the ability to produce novel traits has been widely accepted, in no small part because genome sequencing has confirmed the ubiquity of gene duplicates across the tree of life[188–191]. While other mechanisms for innovation such as *de novo* gene birth[192,193], gene remodelling (fusion, fission, exon shuffling)[194] and even horizontal gene transfer (HGT)[195,196] do have roles to play, there is strong evidence supporting gene duplication as a major driver of complexity[197].

On a fundamental level, gene duplication is the production of two or more genes from one pre-existing gene copy, and there are a limited number of mechanisms that can produce this outcome[198,199]. These are 1) unequal crossing over during recombination, which causes tandem duplicates[200–202]; 2) duplicative DNA transposition, which can occur by non-homologous end-joining or non-allelic homologous recombination[203]; 3) retrotransposition, the retrotranscription of an mRNA and reinsertion into the genome, caused by the action of genes encoded by transposons[204–206]; 4) segmental or

chromosomal duplications of larger genomic regions[207–210]; or 5) whole genome duplication (polyploidisation)[187,211]. The mechanism by which a gene duplication occurs is critical to the outcome of the duplication. For instance, retrotransposed gene copies are without their original *cis*-regulatory elements, while tandem duplications copy coding and (some or all) non-coding regions alike. As such, retrocopies are more likely to be inactive and subsequently pseudogenised. The relative importance of different mechanisms varies between lineages: the high copy number of LINE-1 elements in mammalian genomes means that retrotransposition of protein-coding genes has occurred at a very high rate[206,212,213].

Gene duplication and resultant genes with novel functions are major contributors to the evolution of lineage-specific traits and adaptive innovations[197]. A well-documented example is the co-option of a duplicate of an *RNase1* gene in leaf-eating colobine monkeys to break down nutrients from symbiotic bacteria in their foregut[214]. WGD in particular has been important to the evolution of complexity and the emergence of evolutionary novelties[215]. One explanation for this is that, in contrast to the other mechanisms, whole GRNs are duplicated and can therefore be co-opted to new functional roles[189,216]. Though the fixation of a WGD is a rare event, it appears to have been critical to the emergence of several evolutionarily successful lineages, including angiosperms and vertebrates[215]. Two rounds of WGD occurred during the evolution of vertebrates[187,217–220], and it has been suggested that this facilitated their increase in morphological complexity by the release from developmental constraint and elaboration of developmental networks, such as those regulated by Hox genes[221–223]. The GRN underlying the vertebrate-specific neural crest, for example, is underpinned by products of the 2R WGDs[224–227].

1.4.2 The fate of duplicate genes

While the mechanisms that create gene duplicates are well understood, the events that occur immediately following a duplication are yet to be fully unravelled[198]. Innan and Kondrashov (2010)[228] have offered, in my opinion, the best classification of the hypotheses explaining these events, and it is that which is followed here. Category I models assume that the initial duplication event is neutral with respect to fitness and is fixed by drift, and therefore that, if no selective advantage is subsequently acquired, mutation will not be selected against and in most cases pseudogenisation will quickly occur[190,191,229]. Ohno proposed a model for the retention of duplicates in such a scenario which is now known as neofunctionalisation[187]. Under this model, the presence of two copies frees them from the strong constraint of purifying selection usually placed on a functional gene, and random mutations may cause one to acquire a novel function, while the other copy retains the ancestral function. This evolution of a new function could be initially neutral, and then selectively favoured once the organism experiences a change of environment (Dykhuizen–Hartl model)[230], or it may be caused by positive selection for adaptive mutations (adaptation model)[198]. The outcome of both models is two duplicates performing different functions which are retained by selection. At this point it is worth noting that, except in the case of retrotransposition, two gene copies produced by the duplication of one gene are of equal age, and there is no daughter and parent. While acknowledging this, I use the 'parent, daughter' terminology in parts of this thesis as it is informative in cases where there has been asymmetric evolution and one copy has retained the ancestral function while the other has diverged.

The other main category I model is duplication-degeneration-complementation (DDC)[231]. Consider for instance the duplication of a gene that has multiple enhancers, each driving expression at a different location. Should mutation disrupt one of these enhancers in one duplicate and another enhancer in another duplicate, both duplicates would necessarily be retained by selection just to perform the original gene's function. This could equally

occur in a protein-coding part of the gene, for example with mutations interrupting different protein domains in each copy. Regardless, in both cases, the original function is compartmentalised between the two daughter genes; this is known as subfunctionalisation. The final category I explanation is the escape from adaptive conflict (EAC) model[232,233]. This model posits that, if the ancestral gene possessed two or more functions, positive selection can drive the specialisation of duplicates to one of these functions, resulting in improved performance of their respective roles.

Category II models assume that the initial duplication event is beneficial to fitness. The first model, again discussed by Ohno in 1970, considers a scenario where gene function is dosage-sensitive and the increased dosage caused by duplication is selectively advantageous[187,228,234,235]. Alternatively, duplications may be favoured by selection because they buffer against deleterious mutations[236,237], but the selective advantage of this is likely to be minimal, and it requires repeated, cyclical duplications to continuously shield from mutation events[228,238]. Both of these models result in an outcome of 'gene conservation', with both (or all) duplicates retaining the ancestral function[198]. The final category II model posits that a duplication event will increase fitness if it instantly creates a gene with a new function; this may happen through, for instance, partial copying of the coding sequence or regulatory regions, or acquisition of new regulatory properties in its new location[228,239,240].

Category III models work on the basis that there was pre-duplication polymorphism. One such model is permanent heterozygote advantage; if a gene for which there is heterozygote advantage (i.e. a higher fitness of heterozygotes) is duplicated, and recombination leads to one duplicate of each allele, then this will be selectively favoured as the heterozygote fitness benefit is achieved without the reduced fitness of homozygotes[187,228,241,242]. Finally, the one category IV model applies when a large duplication such as WGD is fixed. This 'dosage-balance' model posits that if two or more

genes have interdependent optimal dosages, such as proteins that form a part of larger complexes, and both are duplicated by WGD, then one cannot be lost without the other because this would lead to suboptimal dosages, leading to the necessary retention of both genes by selection[228,243,244].

1.4.3 Interlocus gene conversion

Tandem duplications create neighbouring gene copies with identical sequences. One consequence of this is that, if a double-strand break occurs in the region, the high sequence similarity means that the repair process may erroneously use a paralogous sequence as the template for repair. This process of unidirectional DNA transfer from a donor to an acceptor sequence, called interlocus (or non-allelic) gene conversion, can result in the 'copy and pasting' of the sequence of one paralogue over another[245]. It is most likely to occur between genes that are highly similar in sequence and close together in the genome[245,246], so recent tandem duplicates are prime candidates. Gene conversion drastically affects the evolution of duplicate gene copies; it can completely prevent divergence by repeated homogenisation of their sequences, driving a process called concerted evolution[200,247] and preventing neofunctionalisation and subfunctionalisation. Notably, it can also be a facilitator of the category II dosage-sensitivity selective benefit of gene duplication, because it prevents pseudogenisation of the genes[190,248]. Gene conversion is not an inescapable fate, and the process can be halted if a mutation causes a large divergence in the sequences[249–251]. Overall, interlocus gene conversion can occur after gene duplications and its impact on their evolution must be considered.

1.4.4 The importance of gene loss

The importance of gene loss to the evolution of ETCHbox genes should not be overlooked as part of this discussion. The two major molecular mechanisms by which a gene can be functionally lost from a genome are pseudogenisation by mutation or

instantaneous complete deletion, for example by an unequal crossover event[252]. Frameshift-inducing insertions or deletions, premature stop codons, splice site mutations or the insertion of a transposable element can all cause the pseudogenisation of a gene. Just like gene duplications, gene losses are 1) pervasive throughout the evolutionary history of metazoans and 2) able to have dramatic effects on genome evolution[252–255]. Tapeworms, for example, have undergone dramatic losses associated with their parasitic lifestyle and derived simplistic body plan, with the gene number of *Echinococcus granulosus* estimated at only 10,231[256]. Their loss of a through-gut, for instance, is correlated with the loss of the entire ParaHox homeobox gene cluster[256], which is involved in its specification[25,257,258].

Most gene losses are thought to occur by relaxed selection, with inactivating mutations not selected against if they do not reduce fitness. Interestingly, this can happen even if the gene has a beneficial function if distributed robustness or functional redundancy lead to mutational robustness, shielding a mutation from purifying selection[124,252]. Furthermore, just like duplications or sequence changes, gene losses can be adaptive and therefore actively favoured by selection. This is often associated with a substantial change to the organism's environment[252], and has been widely shown in bacteria[259–261] but also applies to multicellular organisms including mammals[262]. In recent human history, a pseudogenising mutation in the *CCR5* gene, which confers resistance to viral infections including HIV, has been under strong positive selection[263,264].

1.5 Mammalian preimplantation development

1.5.1 Characteristics and phylogenetics of Mammalia

Discussing the makings of mammals represents somewhat of a sharp change of direction from the role of gene duplications in evolution, but is equally important to understanding the evolution of ETCHbox genes. Mammals (class Mammalia) are a small but diverse clade of vertebrates defined by a distinct group of characteristics including the presence of hair; a jaw hinged directly to the skull; a unique architecture of the inner ear; a neocortex in the brain; a muscular diaphragm; enucleated mature red blood cells and endothermy[265–268]. However, their distinguishing feature, and that from which the taxon's name is drawn, is the milk-producing mammary glands from which mothers feed their offspring. Mammals first appeared in the fossil record approximately 220 million years ago and underwent a dramatic adaptive radiation into ecological niches vacated by dinosaurs following the Cretaceous–Paleogene (K-Pg) extinction[269,270]. Extant Mammalia consists of around 5,500 species, from 2 gram bats to 150 kilogram blue whales, exploiting marine, freshwater, terrestrial, arboreal, subterranean and aerial lifestyles, and inhabiting myriad ecological niches in all corners of the world[265,270,271]. Today, mammals are of critical ecological and agricultural importance.

Surviving mammals belong to one of three clades: Monotremata (monotremes), Marsupialia (marsupials) or Placentalia (placentals). Though mammals are associated with viviparity, the birth of live young, monotremes, the most basal extant mammalian clade, do not possess this feature. The two extant monotreme families, Ornithorhynchidae (platypus) and Tachyglossidae (echidna), possess a mix of typically mammalian and primitive synapsid properties; for instance, despite being egg-layers, they suckle their young and have bona fide hair[267]. The highly significant innovation which occurred in the ancestor of Theria (marsupials and placentals) was embryo implantation, during which trophectoderm cells adhere to the uterus' endometrium[272] and

form a placenta by which nutrient and gas exchange occurs[273]. Most marsupials use a relatively simple, non-invasive chorio-vitelline placenta developing from the trophectoderm and yolk sac[274]. After a short gestation period and the birth of altricial young at a comparatively early stage of development, marsupials are weaned for an extended period, completing their development attached to their mother's teat in a pouch (marsupium)[128,266,274,275].

Placentals (or eutherians) exhibit marked variation in their placentas[127,128], but all employ multi-layered chorio-allantoic placentas developing from the trophoblast and allantois, using this to develop for much longer periods inside the womb and resulting in birth at a later stage of development[266,274]. The Eutheria is by far the most speciose group of mammals[266] and includes the taxa containing the main mammalian model species, such as primates, rodents and ungulates. It is taxonomically divided into four superorders: Afrotheria, Xenarthra, Laurasiatheria and Euarchontoglires[276]. Together, the Laurasiatheria and Euarchontoglires constitute the Boreoeutheria, but there has been significant debate about the evolutionary relationships between the Boreoeutheria, Afrotheria, and Xenarthra[269,276–279]. The scenario presented in Figure 1.9, in which Afrotheria and Xenarthra together form a clade (Atlantogenata) which is sister to the Boreoeutheria, is the current prevailing opinion and therefore used in this thesis[266,276,280,281]. Substantial phenotypic diversity can be present within superorders; Laurasiatheria, for example, includes Carnivora (e.g. dogs, seals), Cetartiodactyla (e.g. whales, cows), Chiroptera (bats), Eulipotyphla (e.g. shrews), Perissodactyla (e.g. horses, rhinoceroses) and Pholidota (pangolins).

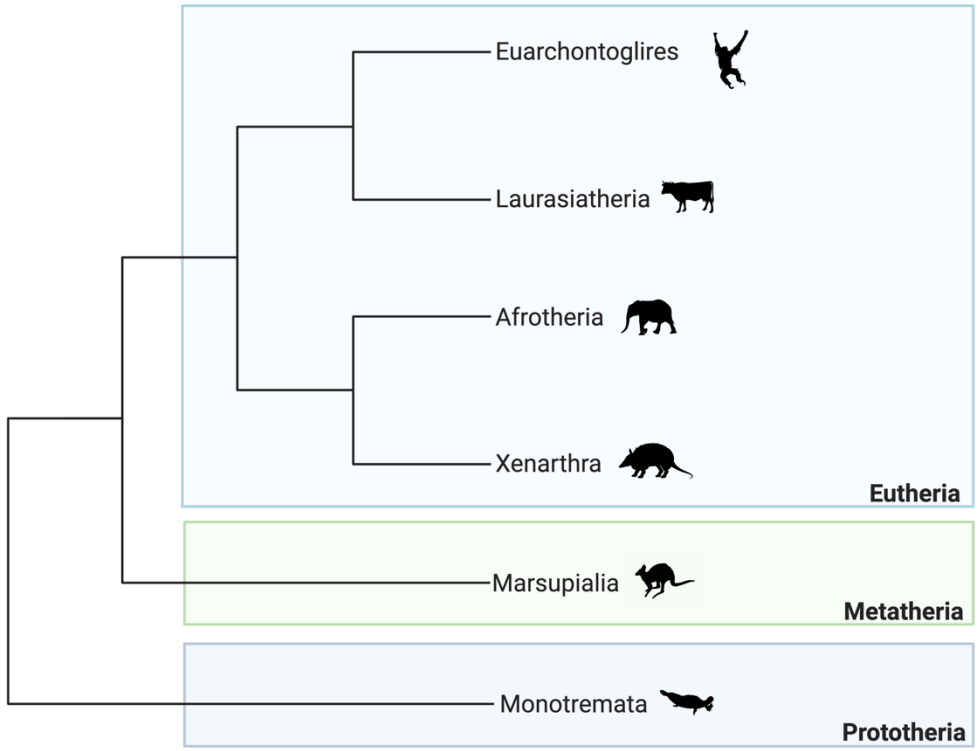


Figure 1.9 Phylogeny of Mammalia. Prototheria consists of monotremes and their extinct relatives, Metatheria of marsupials and their extinct relatives, and Eutheria of placentals and their extinct relatives. The Atlantogenata hypothesis for rooting the placental tree is currently supported, with Euarchontoglires and Laurasiatheria sister to Afrotheria and Xenarthra. Figure created with BioRender.

1.5.2 Mammalian preimplantation development: lessons from humans

Preimplantation development, the period from fertilisation to blastocyst implantation (Figure 1.10), encompasses only a small number of days but includes several dramatic changes, including genome reprogramming, the activation of transcription from the embryonic genome and the first lineage specification events in an individual's life. It begins with the fusion of two gametes, the spermatozoon and oocyte. This brings to an end the arrest of the oocyte at metaphase of meiosis II, and results in the formation of a fertilised, totipotent zygote, with two pronuclei each containing the haploid genome of one parent[282,283]. In this introduction, I focus on the events up to blastocyst formation and the second lineage specification decision, as further details are of little relevance to the following work. Fertilisation is followed by the initiation of several cleavage divisions, between which no cell growth occurs, resulting in progressively smaller blastomeres with

the total size of the embryo remaining constant[284,285]. A multiplicity of epigenetic changes occur throughout preimplantation development; the major wave of DNA demethylation, for instance, occurs after fertilisation and before the 2-cell stage[286–288]. Embryonic genome activation (EGA), also known as zygotic genome activation (ZGA) is the wave-like activation of transcription of genes from the genome of the embryo[289], prior to which the embryo is transcriptionally silent and relies on the translation of maternal mRNAs[290,291]. In humans, the major wave of EGA occurs between the 4 and 8-cell stages (embryonic day [E] 3), with minor waves occurring from the one-cell stage[291–294]. In addition to kick-starting transcription in the embryo, EGA also involves maternal RNA degradation and dramatic chromatin and epigenomic remodelling[137,292,295,296].

Cell polarisation and embryonic compaction occur after the third cleavage, with the latter involving the enlargement of cell-cell contacts and flattening and tighter attachment of cells, resulting in the formation of a compact cell cluster called the morula[297–301]. At this early point in development the first cell fate decision occurs: the internalised cells are specified to become the inner cell mass (ICM) and the outer cells trophectoderm (TED). The blastocoel cavity is then formed as TED cells pump fluid into the morula to create a fluid-filled lumen[297]. This culminates in the formation of the blastocyst, the structure which enables implantation, with an outer layer of TED cells surrounding the blastocoel with the ICM inside[302]. The second cell fate decision then occurs in the ICM at approximately E5 to E6, separating the epiblast (EPI), from which all lineages of the embryo proper are derived, from the primitive endoderm (PE, or hypoblast), which is destined to form the extra-embryonic yolk sac[283,303–305]. TED is a differentiated epithelial tissue and, after hatching from the zona pellucida, it facilitates implantation, which consists of attachment to the endometrial surface of the uterus and invasion into maternal epithelium[306,307]. It is worth noting that preimplantation development is remarkably error-prone; during its 7 - 9 day duration around two-thirds of pregnancies are lost[307–309].

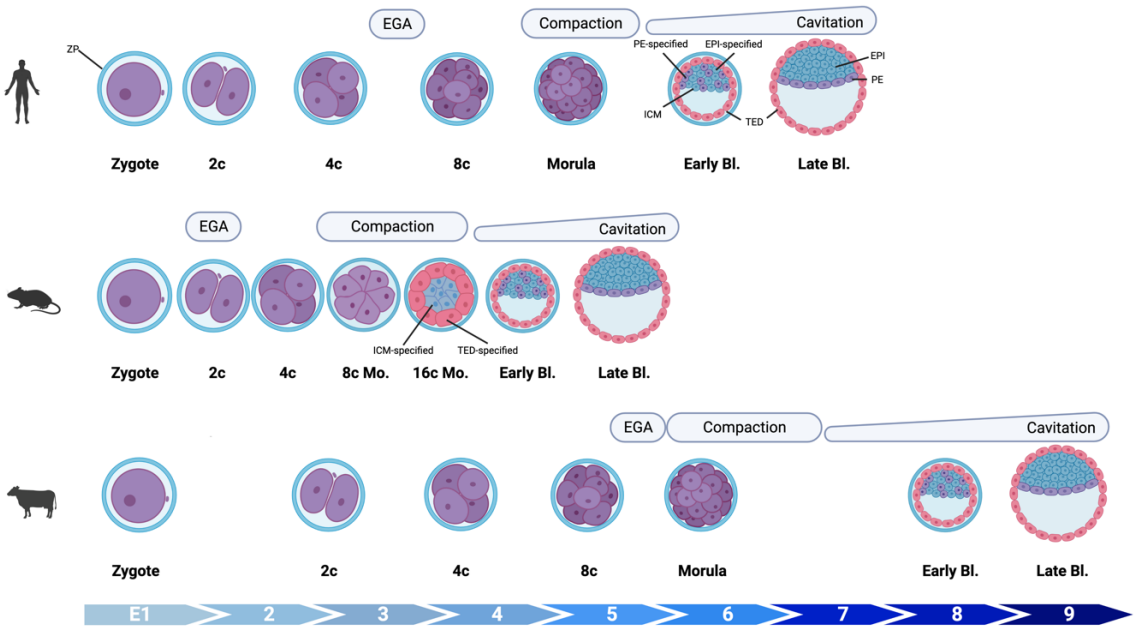


Figure 1.10 Preimplantation development in eutherians. The schematic shows the progression of development from fertilisation at E0 to the completion of the first two lineage specification events and blastocyst formation in humans, mice and cattle. Cells are coloured for their lineage: dark purple for unspecified cells, red for TED cells, blue for ICM cells and then EPI cells, and light purple for PE cells. The figure highlights key differences between species, such as the length of preimplantation development, the timing of EGA and the stage at which embryonic compaction occurs. The 'salt and pepper' mechanism of EPI and PE specification is now thought to occur across Mammalia. Abbreviations: Bl. = blastocyst; c = cell; E = embryonic day; EGA = embryonic genome activation; EPI = epiblast; ICM = inner cell mass; Mo. = morula; PE = primitive endoderm; TED = trophectoderm; ZP = zona pellucida. Figure created with BioRender.

1.5.3 Transcriptional control of lineage specification

Due to ethical considerations surrounding work with human embryos, the molecular mechanisms underlying these successive cell fate specification events are best understood in mouse. Mouse blastomeres are committed to becoming TED or ICM by E3.5 (32-cell stage)[285]; Cdx2 is the key driver of TED cell fate and is essential for TED formation[22], activating downstream TFs necessary for TED maintenance including Eomes and Elf5[283,310,311]. Oct4, Nanog and Sox2 are markers of future ICM cells; since Cdx2 represses these genes but is not expressed in inside cells, they become restricted to future ICM cells[22,24,312–315]. Reciprocal inhibition between TED and ICM-marking TFs subsequently reinforces this pattern, leading to mutually exclusive expression[316].

The second lineage specification event is initiated at E4.5, with the PE forming as a monolayer on the blastocoel surface of the ICM, with the EPI inside. Nanog promotes EPI fate and suppresses PE fate and Gata6 does the opposite[24,178,314,317,318], again exploiting reciprocal repression[180,319,320]. However, unlike the specification of ICM and TED which is based on position-dependent determination, mouse ICM cells exhibit mutually exclusive expression of Nanog or Gata6 in a 'salt and pepper' pattern (see Figure 1.10 stage 'Early Bl.'), followed by cell sorting by cellular migration of lineage-predisposed cells to their proper locations (Figure 1.10 stage 'Late Bl.')[285,321]. Both EPI and PE-fated cells migrate, and those remaining in an incorrect position undergo apoptosis[322–324]. This 'salt and pepper mechanism' appears to be conserved across eutherians[325–327]. Within each lineage, further TFs are activated to reinforce fate; for instance, a cascade of Gata6 then Sox17 then Gata4 then Sox7 activation occurs during PE specification[328].

1.5.4 Preimplantation development across Mammalia

While the major stages of preimplantation development are conserved across eutherians, there is significant variation in the timing of key events. For instance, milestones tend to occur at a lower cell number (i.e. earlier in development) in mouse than in human (Figure 1.10)[285,302,306]. EGA starts in the zygote and major EGA occurs at the 2-cell stage (4 to 8-cell stage in human); compaction occurs to form a morula by the 8-cell stage (16-cell in human); the blastocyst forms by E3.5 (E5 in human); and implantation occurs by E4.5 (E7 in human)[283,285,302,329–334]. Venturing further afield, bovine preimplantation development is further protracted, taking over 9 days to complete[326].

The molecular mechanisms underlying this period of development also vary more than one might expect. One dramatic difference between species at a molecular level is the identity of transcripts upregulated at EGA: only 40% of transcripts activated at human

EGA are also upregulated in mouse, and only 18.5% are conserved in cattle[335]. Remarkable variation also exists in the expression of lineage-specifying TFs between species. Key lineage-defining factors Cdx2 (TED)[336], Oct4 (ICM)[326,337], Sox2 (ICM and EPI)[325] and Sox17 (PE)[338] show divergent expression dynamics between human, mouse and bovine. Some embryonic lineage-specific TFs are also unique to one species; for instance, human orthologues of the mouse TED TFs Elf5 and Eomes remain completely inactive throughout preimplantation development[283,339]. The transcriptome of the EPI lineage, especially, is highly divergent between species and has been rewired by TE integration into GRNs[340]. Overall, there exists surprising variation in the networks underlying the events of preimplantation development across eutherians.

Differences in preimplantation development between eutherians and marsupials are even starker. Firstly, marsupial embryos do not undergo compaction to form a morula prior to blastocyst formation, with cells adhering to the zona pellucida rather than each other[127,341]. However, the salient difference is the lack of an ICM within the marsupial blastocyst, with the pluriblast (cells fated to become epiblast and PE)[342] forming as part of the continuous epithelium surrounding the unilaminar blastocyst[127] (Figure 1.11). These differences continue at the molecular level: while NANOG and GATA6 are conserved as markers of epiblast and PE respectively, key mouse markers for the pluriblast-TED specification, such as OCT4, SOX2, CDX2 and GATA3, while present in the genome, curiously do not show similar specificity of expression in marsupials[341,343].

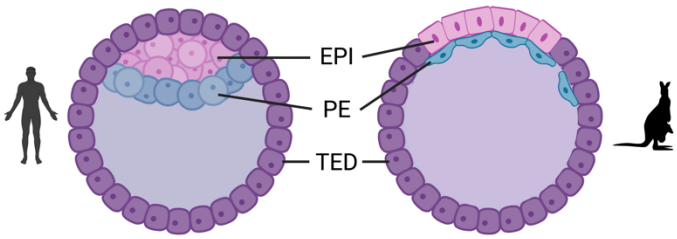


Figure 1.11 Blastocysts of eutherians and marsupials. Schematic comparing the blastocyst stages of eutherians (left, e.g. humans) and metatherians (right. e.g. kangaroo). The key difference is the lack of an ICM in metatherians, with the pluriblast forming as part of the unilaminar epithelium. Abbreviations: EPI = epiblast; PE = primitive endoderm; TED = trophectoderm. Figure created with BioRender.

1.6 Transposable elements

1.6.1 Introduction to transposable elements

Here we make another turn, away from the development of mammals and towards the composition of their genomes and, specifically, the repetitive sequences within them. TEs are mobile DNA sequences that possess the ability to change their position in the genome[344]. Though some TEs have been co-opted by their host genomes for functional use, they are broadly considered selfish elements, which are present in genomes because of their propensity for being replicated rather than because it is favoured by natural selection on the host[345–347]. Through this ability to replicate, they accumulate in genomes over time, and can constitute a considerable proportion of a given genome, including 37% and 45% of the mouse and human genomes respectively[122,348]. Though TEs are found across the tree of life, for the purpose of this thesis I take a mammal-centric approach to this introduction.

TEs are classified in a taxonomic system analogous to that used to classify species, and are divided into two broad classes (Figure 1.12). Class I transposons (retrotransposons) replicate via an RNA intermediate that is reverse transcribed to DNA and then integrated at another site in the genome in a ‘copy-and-paste’ style mechanism known as retrotransposition. Most class II transposons (DNA transposons) replicate with a DNA

intermediate and use a 'cut-and-paste' mechanism catalysed by an endogenously-encoded transposase[344,349]. DNA transposons are inactive in all mammals except bats[350–352], and comprise only around 3% of the human genome[122].

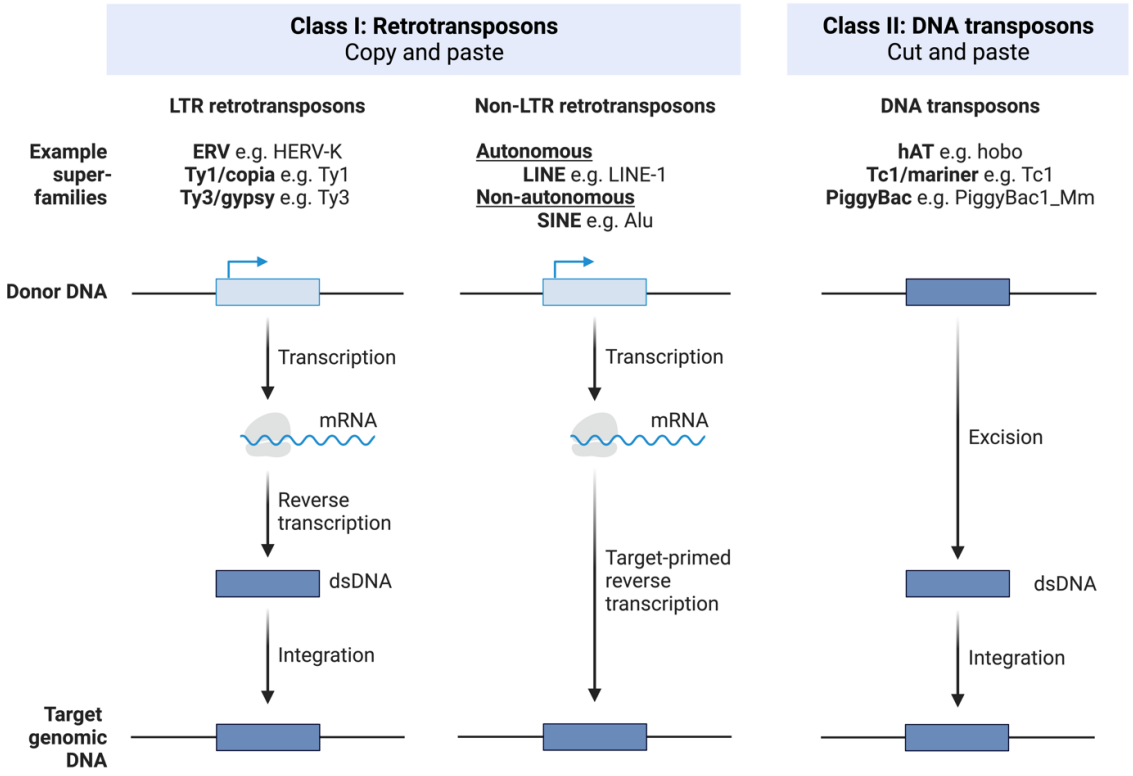


Figure 1.12 Transposable element classification and insertion mechanisms. Transposable elements are split into two classes based on their integration mechanisms. Class I elements (retrotransposons) integrate via an RNA intermediate in a copy-and-paste style mechanism, while most class II elements (DNA transposons) integrate via a dsDNA intermediate using a cut-and-paste style mechanism. Class I elements are split into LTR elements such as ERVs and non-LTR elements. Non-LTR retrotransposons can be autonomous (e.g. LINEs) or non-autonomous (e.g. SINEs). Figure produced with BioRender, data from Piégu et al. (2015)[353] and Bourque et al. (2018)[344].

1.6.2 Retrotransposons

Retrotransposons (class I) are the dominant TEs in mammals and are divided into long terminal repeat (LTR) and non-LTR retrotransposons based on their mechanism of integration. LTR retrotransposons (or endogenous retroviruses, ERVs) are derived from retroviruses and utilise an integrase-mediated genomic integration mechanism similar to that of their viral ancestors, while non-LTR elements use target-primed reverse transcription[354–357]. Even closely related genomes can show diversity in both their overall

TE repertoires and their complement of TEs that retain competence to transpose. In humans, all ERVs are now likely to be inactive (i.e. incapable of retrotransposition)[356,358,359] but in mice, ERVs remain remarkably active and are the cause of up to 10% of de novo mutations[360]. ERVs occupy approximately 8% and 10% of human and mouse genomes respectively[122,348].

Non-LTR elements can be divided into long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs) and Penelope-like elements (PLEs). LINEs are autonomous, encoding two open reading frames (ORF1 and ORF2) which mediate their retrotransposition, while SINEs are non-autonomous, and require LINE proteins acting *in trans* for retrotransposition[361–364]. PLEs are absent in humans[365], but LINEs and SINEs are highly abundant, comprising 21% and 13% of the human genome respectively[122]. The vast majority of human LINE elements are of the LINE-1 family, and Alu elements are the dominant SINE[122]. In humans, the only transposons that retain competence to retrotranspose are non-autonomous Alus and SINE-VNTR-Alu (SVA) elements (both SINEs), and 80 to 100 autonomous LINE-1 copies[359,366–369]. Active Alus predominantly belong to the youngest Alu subfamily, AluY, but a smaller number of members of the AluS subfamily also remain competent[370,371].

1.6.3 The good and the bad of TEs

TEs pose a significant threat to the integrity of their host genomes. Insertion into the genome is a mutagenic event; it can occur in both the germline and soma, but is of greater importance in the germline because such mutations are heritable[344]. While many TE insertions are selectively neutral, insertions can be deleterious if they disrupt functional sequences, for instance directly interrupting coding sequence, introducing intragenic polyadenylation signals, causing aberrant splicing, or creating deletions at the insertion site[372]. The first known *de novo* disease-causing TE insertion was the insertion of a LINE-1 element into Factor VIII, resulting in Haemophilia A[373] and, as of 2016, there

were 124 known TE insertions that cause genetic disease in humans[374]. As a result of this, mammalian genomes house an arsenal of proteins that restrict TE activity at various stages, including translational repression and the inhibition of retrotransposition itself[375].

Despite this negative reputation, the oft-used label of TEs as 'junk DNA' overlooks a vast diversity of functional roles. TEs are significant contributors to the diversity of *cis*-regulatory elements, providing enhancers, repressive elements, promoters and protein binding sites, and are thought to play key roles in large-scale rewiring of genetic networks[65,344,376]. Many TEs are also transcribed, and function as non-coding RNAs. Indeed, LINE-1 expression is an essential part of mouse preimplantation development; both HERVK and HERVH are part of the pluripotency network in humans, and developmentally important long non-coding RNAs like linc-ROR include TEs[129,130,377–382]. TEs are also the source of many microRNAs[383–385], and parts of some TEs have even been domesticated to function as coding RNAs[344]; one such example is syncytins, which function in placentation and are derived from the *env* gene of an ERV[386,387]. TEs can also be incorporated into existing genes as exons or part of exons; indeed the ETCHbox gene *ARGFX* has an exon derived entirely from an Alu element in humans[90]. Overall, TEs are major drivers of genome evolution in mammals.

1.6.4 TEs in preimplantation development

Given the potential threats of TE expression and the comprehensiveness of genomes' TE-regulation mechanisms, it is remarkable that during normal mammalian preimplantation development many families of TEs are briefly highly expressed in stage-specific pulses[129–131,388]. For instance, each stage of human preimplantation development is characterised by the expression of specific families of ERVs, with transcription driven by their LTRs[129]; HERVK expression is activated at EGA and transcription continues into the epiblast[130], and the young, human-specific ERV LTR5-Hs is derepressed from the 8-cell stage for one cell division[389]. The observed large

swathes of TE activation are thought not to be solely a passive, indirect effect of heterochromatin reprogramming[65], and are instead a product of both selective pressures on selfish TEs to replicate within the germline, and the co-option of TEs into embryonic regulatory networks[59].

Intriguingly, these derepressed ERVs are not simple parasites and appear to play active roles in the process of development, for example acting as stage-specific transcription initiation sites. In mice, MERVL (and its LTR, MT2) is expressed highly specifically at the 2-cell stage[57,390]. MERVL LTRs act as powerful 2-cell stage-specific promoters, activating the transcription of a large suite of 2-cell genes as chimaeric transcripts[57–59]. This is because they contain binding sites for the key TF driver of the 2-cell state, Dux, which strongly activates MERVL expression as well as that of 2-cell genes[60,61]. Overall, this TE has been recruited to the rewiring of an early developmental GRN, coordinating the expression of 2-cell stage-specific genes. This is unlikely to be unique; the human orthologue DUX4 and HERVL act similarly at human EGA[61] and other TEs (namely LTR70, LTR5_Hs and SVA_D) have been suggested to promote stage-specific gene transcription later in human development[131].

LINE-1 elements also show significant peaks of expression in both human and mouse preimplantation development[287,378,388,391,392], and pose another remarkable example of TE co-option in development. In mice, LINE-1 is required for the progression of development past the 2-cell stage, and the vast majority of LINE-1 knockdown embryos fail to reach the blastocyst[378,393]. This is because LINE-1 represses the 2-cell master regulator gene *Dux*, facilitating an exit from the 2-cell state, and also activates rRNA synthesis to allow cell proliferation[378]. In addition to this function, LINE-1 is an important regulator of chromatin accessibility during mouse preimplantation development[391]. The exaptation of TEs for developmental processes is not unique to mice. For example, the human ERV HERVH is part of the pluripotency network: it is specific to the pluripotent

epiblast[129,394]; is a direct target of the pluripotency factors NANOG and OCT4; and is required to maintain hESC identity[377,380]. Indeed, 25% of the binding sites for NANOG and OCT4 fall within TEs, suggesting that TEs have contributed to a significant rewiring of the pluripotency network. In summary, TEs are now closely intertwined with the GRNs regulating preimplantation development in eutherians.

1.7 Transcriptomics and RNA-sequencing

1.7.1 Introduction to transcriptomics

A central theme running throughout this thesis is the use of transcriptomic data arising from RNA-sequencing (RNA-seq) experiments, and it is therefore worthwhile taking some time to briefly explore the nature of such data. For a cell or tissue in a specific state, its transcriptome is the identity and respective quantities of coding and non-coding RNA transcripts present[395–397]. The field of transcriptomics is critical to our understanding of the genome, gene function and regulation, and more broadly of cellular and organismal biology and disease. While hybridisation-based technologies such as microarrays[398] were previously common, their drawbacks, most notably the requirement for existing sequence knowledge, have meant that they have been largely superseded by sequencing-based approaches in the last 10 years.

RNA-seq is a technology that uses next-generation sequencing (NGS) to reveal the sequences and quantities of RNA molecules in a sample, and its advent has revolutionised our ability to study cells' transcriptomes. Its capacity to characterise complete transcriptomes quantitatively, at single-base resolution, with high sensitivity, reduced biases, high throughput and if necessary without a reference genome sequence, vastly exceeds the capabilities of previously used sequence-based technologies like serial analysis of gene expression (SAGE) and expressed sequence tag (EST) sequencing[395,399]. Since it was first developed around 15 years ago[400,401], RNA-

seq has become a fundamental and universal tool used routinely by molecular biologists. While differential gene expression analyses in which the quantitative expression level of genes is compared between conditions remain the most widely used application of RNA-seq, it is also frequently utilised for other purposes, for example to identify novel genes or transcripts or characterise splice variants[402].

1.7.2 Computational analysis of transcriptomic data

After RNA extraction from a sample and sequencing on a high-throughput platform (usually Illumina), raw RNA-seq data are distributed in fastq format and processed using established bioinformatics pipelines. At each stage of this process, several different programmes are available, each with unique strengths and drawbacks. Pipelines must begin with quality control (QC)[403]; for Illumina data, FastQC[404] is often used to measure many QC metrics. QC should result in the discarding of poor-quality reads, the trimming of sequencing adapters and, debatably, also the trimming of poor-quality bases[405,406]. Post-QC, there are three main analysis pipelines (Figure 1.13). Conventionally, reads are mapped to the genome, assembled into transcripts and then transcripts quantified (pipeline 1). More recently, software such as Kallisto[407] and Salmon[408] has employed a lightweight pseudoalignment approach, mapping reads directly to the transcriptome rather than to the genome (pipeline 2); such programmes are significantly faster than traditional alignment, but a drawback is the lack of an alignment score to validate mappings, meaning reads where even the best mapping is spurious are assigned to that locus[409]. Both of the above methods rely on the availability of genome or transcriptome data from the species being sequenced; should this not be available, programmes such as Trinity[410] can be used to conduct de novo transcriptome assemblies without a reference sequence (pipeline 3). Reads are then mapped back to the transcriptome and quantified as in pipeline 2. Once quantification is complete, downstream analyses, such as differential expression analyses using DESeq2[411] or edgeR[412], can be performed to answer questions of interest.

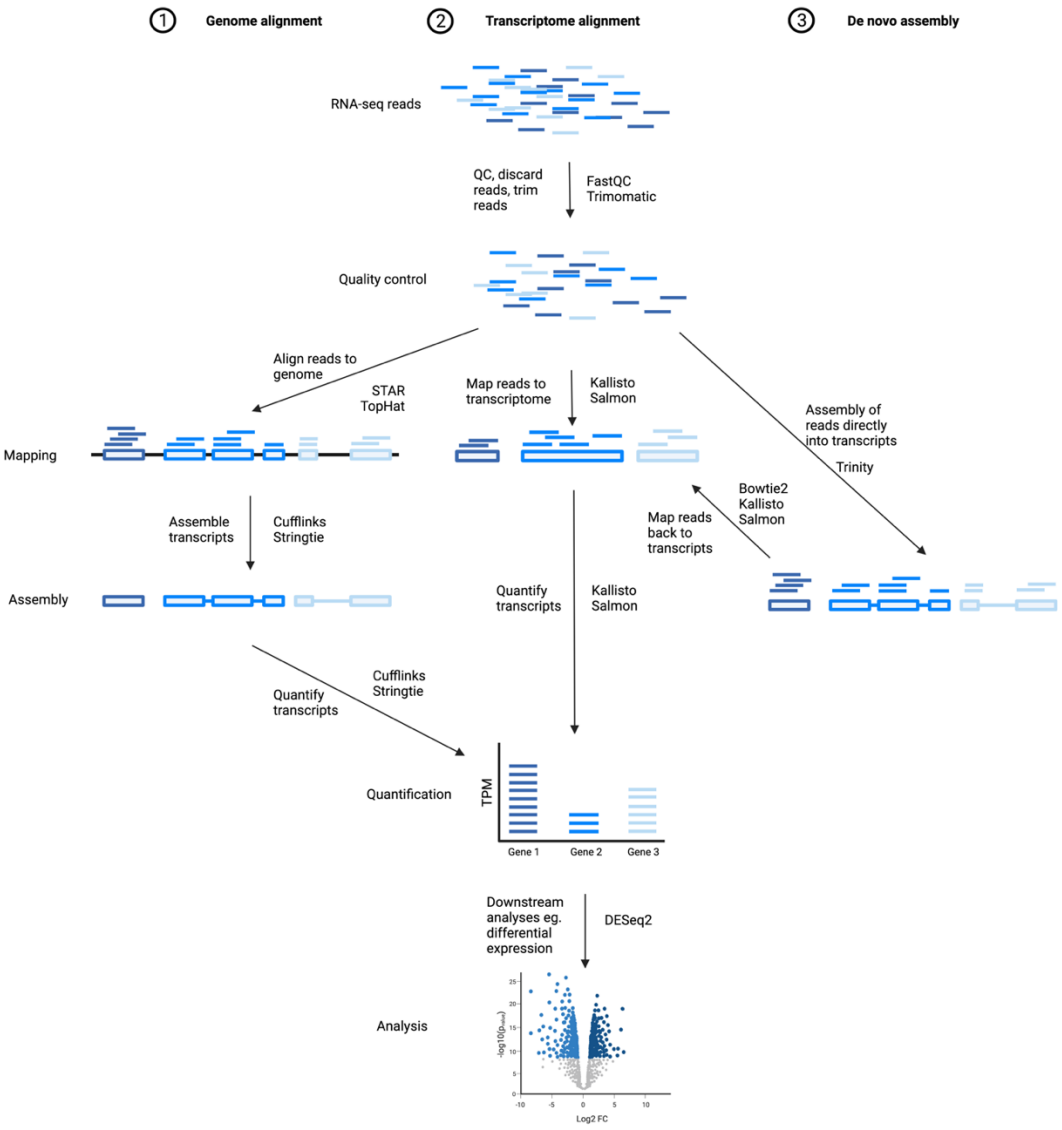


Figure 1.13 Bulk RNA-seq data analysis pipelines. In general, one of three main pipelines are used: 1) genome alignment, 2) transcriptome alignment, or 3) de novo assembly. Figure produced with BioRender.

1.7.3 Advances in transcriptomics

Developments in RNA-seq technologies have continued apace. For instance, long-read Pacific Biosciences (PacBio) Iso-Seq increases the read lengths from around 50 to 200 bp to approximately 10,000 bp, allowing the sequencing of entire transcripts and permitting better isoform-specific mapping[413]; direct sequencing of RNAs (direct RNA-seq) using Oxford Nanopore Technology retains epigenetic information lost during standard workflows and eliminates biases associated with library preparation steps[414]; and unique molecular identifiers (UMIs), random DNA molecules tagged to each cDNA

before sequencing, reduce amplification biases and improve the accuracy of quantitative estimates of transcript abundance[415–417]. Spatial approaches and single-cell RNA-sequencing (scRNA-seq) are now further expanding the boundaries of what questions RNA-seq can be used to answer[418,419].

The development of scRNA-seq has been arguably the most significant advancement to RNA-sequencing technologies, allowing individual analysis of the transcriptomes of single cells. Following gene quantification, cell populations are clustered based on the similarity of their transcriptomes; clusters are then identified by marker genes. In this way, the similarities and differences of cells in a population can be explored at an extremely high resolution. With the continual development of higher throughput lower cost scRNA-seq technologies, (such as droplet-based protocols like Drop-seq[420] and Chromium[421]) datasets are becoming increasingly large; experiments have passed the million cell mark[422], and 'atlas' studies have been undertaken with the aim of characterising all cells in a given species[423]. These rapid advances present a challenge in extracting biological meaning from such vast volumes of data.

1.8 Thesis aims

ETCHbox genes offer a promising opportunity to study the recruitment of rapidly-evolving, lineage-specific homeobox genes to roles in early development. They possess an unusual collection of characteristics, with highly specific and conserved expression patterns at odds with their rapidly-evolving sequences and elevated rate of gene turnover. In recent years, data on the functional role of ETCHbox genes has begun to accumulate, but publications invariably neglect to take an evolutionary perspective. The broad goal of this thesis is not only to advance our understanding of ETCHbox genes' embryonic function but, importantly, to elucidate the forces driving their evolution.

Chapter 2: Comparative genomics of ETCHbox genes

Previous comparative genomic analyses aiming to explain ETCHbox genes' curious evolutionary patterns suffered from limited data availability and low-quality genome assemblies, but recent advances in low-cost long-read sequencing have dramatically increased the availability of high-quality chromosome-level assemblies. In this chapter, I ask whether expanded sampling made available by advances in genome sequencing supports the idea of ETCHbox genes as fast-evolving homeobox genes. I also question whether further cases of large lineage-specific ETCHbox tandem duplications can be identified, and ask what evolutionary processes explain the current variation in ETCHbox repertoires observed across eutherians.

Chapter 3: ETCHbox function across large phylogenetic distances

Currently, all functional work on ETCHbox genes has used the model species of humans and mice, but the divergent ETCHbox repertoire of mice impedes meaningful comparisons between these species. In this chapter, I seek to understand the function of ETCHbox genes in a eutherian species from a different superorder (cattle, *Bos taurus*, Laurasiatheria). I explore the functional role of ETCHbox proteins in cattle and ask whether the rapid evolution and divergence of ETCHbox genes between species has been accompanied by changes in this function.

Chapter 4: ETCHbox function in primates

Up to this point, comparisons of ETCHbox genes between species have only concerned species separated by a considerable phylogenetic distance[96]. In this chapter, I use the most densely sampled clade of eutherians, primates, as a model taxon. I ask firstly whether rapid evolution has continued within this clade, and then whether the observed sequence divergence translates into differences in protein function.

Chapter 5: ETCHbox proteins and transposable elements

While numerous studies have provided informative data on the functional role ETCHbox genes play in embryonic development[75,90,96,104], it is clear that our current understanding is far from complete. No association between ETCHbox genes and TEs has been previously investigated, and given that the timing of ETCHbox expression in the preimplantation embryo coincides with that of TE reactivation[129–131], it stands to reason that these could be linked. In this chapter, I hypothesise that either indirectly or directly, ETCHbox proteins may regulate the expression of TEs. I first ask whether genes in the TE repression armoury are regulated by ETCHbox proteins. I then ask whether ETCHbox ectopic expression causes differential expression of TEs and finally I test whether ETCHbox proteins affect the rate of TE retrotransposition.

Chapter 6: Expression of ETCHbox genes at a single-cell level

Previous work has investigated the timing of ETCHbox expression in RNA-seq datasets, concluding that human, bovine and mouse ETCHbox genes are expressed only in the preimplantation embryo[75,90,96]. However, no work has so far been able to ask whether ETCHbox genes show lineage-specificity during their blastocyst expression, which may point to a role in pluripotency, cell fate determination or the development of eutherian novelties like the inner cell mass. In this chapter, I test for lineage-specificity of ETCHbox genes of humans, pigs and mice using publicly-available scRNA-seq data. In addition, previous work has placed minimal focus on the genes which act upstream of ETCHbox genes in the preimplantation embryo; I use GRN reconstruction based on scRNA-seq data to elucidate these factors.

Chapter 2: Dynamic molecular evolution of mammalian homeobox genes: duplication, loss, divergence and gene conversion sculpt PRD class repertoires

Thomas D. Lewin, Amy H. Royall and Peter W. H. Holland

Introduction to manuscript

This chapter has been published as a first author original research article in the *Journal of Molecular Evolution* (doi: 10.1007/s00239-021-10012-6). It was first published online on 7th June 2021, and is reproduced here with minor changes as permitted as part of an integrated thesis. The authors credited on this manuscript, in order, are Thomas D. Lewin, Amy H. Royall and Peter W. H. Holland (corresponding author).

Author contributions

All authors contributed to study conception and design. Material preparation, data collection and analysis were performed by Thomas D. Lewin with input from Peter W. H. Holland. TDL and PWHH wrote the manuscript with input from Amy H. Royall. All authors read and approved the final manuscript.

2.1 Abstract

The majority of homeobox genes are highly conserved across animals, but the eutherian-specific ETCHbox group, embryonically-expressed and highly divergent duplicates of *CRX*, are a notable exception. Here we compare the ETCHbox genes of 34 mammalian species, uncovering dynamic patterns of gene loss and tandem duplication, including the presence of a large tandem array of *LEUTX* loci in the genome of the European rabbit (*Oryctolagus cuniculus*). Despite extensive gene gain and loss, all sampled species possess at least two ETCHbox genes, suggesting their collective role is indispensable. We find evidence for positive selection and show that *TPRX1* and *TPRX2* have been the subject of repeated gene conversion across the Boreoeutheria, homogenising their sequences and preventing divergence, especially in the homeobox region. Together, these results are consistent with a model where mammalian ETCHbox genes are dynamic in evolution due to functional overlap, yet have collective indispensable roles.

2.2 Introduction

Homeobox genes encode a diverse set of transcription factors found across the Eukaryota, each of which has a characteristic DNA-binding homeodomain of around 60 amino acids[12,78,424]. Many homeobox genes play critical roles in early embryo patterning and cell fate specification[14,425,426] and, as fundamental components of developmental gene regulatory networks, are generally highly conserved over large phylogenetic distances[106]. Indeed, most research on homeobox genes has focused on highly conserved examples, such as *HOX*[32,425,427,428], *PAX*[429–431], *POU*[432,433] and *LIM*[434–436].

In contrast, there are a smaller number of fast-evolving, taxon-specific homeobox genes found in some animals, including genes expressed during nematode[437,438], lepidopteran[39,40], spiralian[47,439] and mammalian[90] embryonic development. Within mammals, the clearest examples of fast-evolving homeobox genes are *NANOGNB*, a member of the ANTP class[52], and several loci classified within the PRD class, although they lack a PAIRED box. These include the *CPHX*, *DUX* and *RHOX* groups, and the Eutherian Totipotent Cell homeobox (ETCHbox) genes[68,75,90,100]. Six paralogous groups make up the ETCHbox genes – *ARGFX*, *DPRX*, *LEUTX*, *PARGFX*, *TPRX1* and *TPRX2* – all derived by duplication and extensive sequence divergence from the OTX-family member *CRX*[90,440]. These genes are a synapomorphy of the Eutheria. They are absent from monotremes and marsupials, having arisen in the lineage leading to the eutherians, after which they underwent rapid asymmetric evolution and diverged extensively from *CRX*[90].

ETCHbox genes are notable for their remarkably specific temporal expression patterns. Though there are slight variations, human and cow orthologues are expressed between the 4-cell stage and early blastocyst in the preimplantation embryo and then never expressed again[90]. Despite being extensively duplicated, mouse ETCHbox genes are

also expressed in the preimplantation embryo[90,96,112–114]. Furthermore, data from ectopic expression experiments suggest that ETCHbox proteins regulate preimplantation embryo-expressed genes and that LEUTX has a role in embryonic genome activation[90,104].

Maeso et al. (2016)[90] published the most extensive characterisation of ETCHbox complements to date, comparing nine eutherian species. These data suggested that the group is more dynamic than typical homeobox genes, with a high prevalence of gene duplication and loss, contrasting with their conserved and highly specific expression pattern. However, this analysis was limited by sparse phylogenetic coverage and the low-quality genome assemblies available at the time. Katayama et al. (2018)[95] performed a deeper analysis of *LEUTX* evolution, but this study was restricted to the one orthogroup and was without analysis of gene loss. Recent improvements to long-read DNA sequencing mean that many additional mammalian genome sequences are now available, assembled to high contiguity and accuracy. These data give a timely opportunity to describe the number and organisation of ETCHbox genes across the eutherian phylogeny, which is a necessary step towards understanding the factors underpinning their unusual pattern of evolution.

The causes and consequences of ETCHbox genes' dynamic evolution are yet to be elucidated. It has been proposed that their dynamism may be driven by a possible role in the evolution of reproductive traits in mammals, such as placentation, which is highly variable between eutherians[90], or due to selection in some lineages for shorter gestation times[95]. Alternatively, it may be a consequence of partial functional redundancy, which would cause relaxed selection on loss-of-function mutations, or distributed robustness, where the perturbation of one part of a system (loss of a gene) is compensated by non-redundant parts (other genes)[96,124]. Finally, the genes may lack an important function,

and therefore their high rates of pseudogenisation and loss would be a consequence of relaxed selection.

Here we analyse publicly-available genome sequences to produce a dataset of ETCHbox repertoires for 34 mammals. We focus particularly on assemblies made with long-read DNA sequencing technology and species chosen to optimise phylogenetic coverage, allowing us to deduce with confidence the underlying patterns and pathways of gene duplication and loss. We uncover large arrays of tandem ETCHbox duplicates across multiple species and show that ETCHbox genes have been the subject of positive selection and concerted evolution.

2.3 Materials and methods

2.3.1 Comparative genomics

Genome sequences for 32 eutherian species were downloaded from the NCBI webpage, focussing on taxa with high contiguity genome assemblies (Supplementary Table S2.1); this includes re-analysis of species previously analysed using lower quality genome data[90]. Where possible, genomes sequenced using long-read technologies were utilised, as such data facilitate assessment of tandem gene duplication and gene loss. To include species from every order of the Boreoeutheria (Laurasiatheria and Euarchontoglires), three taxa were included despite lacking long-read assemblies: *Galeopterus variegatus* (Sunda flying lemur, Dermoptera), *Condylura cristata* (star-nosed mole, Eulipotyphla) and *Manis javanica* (Sunda pangolin, Pholidota). High-quality human and mouse genome assemblies were analysed by Maeso et al. (2016)[90] and Royall et al. (2018)[96] respectively and are used but not recharacterised here, giving a total dataset of 34 species.

Homo sapiens (human)[90] and *Bos taurus* (cattle; this work) ETCHbox gene structures were verified using transcriptome data. Briefly, for *B. taurus*, RNA-seq reads[441–444] (Supplementary Table S2.2) were obtained from the NCBI Sequence Read Archive (SRA), aligned to the *B. taurus* reference genome ARS-UCD1.2 using STAR version 2.7.0[445] and assembled into transcripts using StringTie version 1.3.6[446]. Genes of interest were identified in each transcriptome using the Basic Local Alignment Search Tool (BLAST)[447,448], and intron/exon boundaries refined by inspection of raw reads using the Integrative Genomics Viewer (IGV)[449].

ETCHbox genes of other species were identified and annotated using sequence similarity searches of genome assemblies (blastn, blastp, megablast and tblastn[447,448,450]). To identify divergent sequences blastn searches were used with low stringency reward/penalty values (command line options -reward 1 -penalty -1). For short exons, the option -blastn-short was used. Gene identities were assigned using a combination of reciprocal BLAST, synteny and phylogenetic analysis using MrBayes[451,452]. Intron/exon boundaries were refined manually using (a) retrogene sequences, (b) comparison to human and cow sequences validated by transcriptome data, and (c) mammalian consensus splice sites[453]. The first coding exon of ETCHbox genes is very short and highly variable and was therefore not always identified. Genes were considered probable pseudogenes when there were stop codons, splice site mutations or frameshifts upstream of (or within) the homeobox. Genes with frameshifts or premature stop codons immediately downstream of the homeobox are of unknown functional status. If no gene was identified by BLAST and the expected syntenic region surrounding the gene was split over two or more scaffolds we do not conclude certain gene loss.

For phylogenetic analysis, amino acid sequence alignments were made using Clustal Omega in Seaview version 4.7[454,455] and phylogenies inferred using MrBayes version 3.2.7a[451,452] and rendered using iTOL[456].

2.3.2 Estimating gene gain and loss

Two methods were used to assess gene gain and loss. First, ETCHbox genes were grouped into gene families and the stochastic birth and death model in CAFE version 4.2.1[457,458] used to calculate maximum likelihood values of λ and μ (rates of gain or loss, respectively, per gene per million years) and estimate gene numbers at internal nodes. Second, the event-inference parsimony algorithm in Notung version 2.9[459,460] was used. Gene trees were generated for each ETCHbox gene as above and Notung run with a duplication-loss model and default parameters (weights: duplications = 1.5, co-divergences = 0.0, losses = 1.0) to reconcile gene and species trees and estimate the timing and minimum weighted number of independent duplication and loss events. To prevent weakly supported branches causing overestimation of gene turnover, gene trees were rearranged using the 'Rearrange' function, allowing branches with posterior probabilities <95% to be reconfigured to minimise duplications and losses. The species tree used was generated using TimeTree[461].

CAFE was also used to test for an acceleration in the rate of gene duplication of each ETCHbox gene compared to other homeobox genes present in mammals using the Monte Carlo sampling procedure described previously[462,463]. The Viterbi assignment method[458] was used to establish which branches contributed to such accelerations. For the purpose of gene duplication analyses, Cetartiodactyla *TPRX3* genes were assigned as *TPRX2* duplicates, as by Maeso et al. (2016)[90].

2.3.3 Detecting gene conversion

We defined *TPRX1* and *TPRX2* using neighbouring genes and orientation, not sequence: *TPRX1* is upstream of *CRX* and in inverse orientation, *TPRX2* is downstream of *CRX* on the same strand. To test for interlocus gene conversion between *TPRX1* and *TPRX2*, four methods were used, following the guidelines of Mansai and Innan (2010)[464]. First, the expected *TPRX* duplication history was compared to Bayesian gene trees to search for phylogenetic incompatibilities. Second, sequence similarity was assessed by running Biostrings version 2.57.1[465] in R version 4.0.0[466] to compute all versus all Needleman-Wunsch[467] global pairwise alignments. Percent nucleotide identities were calculated and plotted using gplots version 3.0.3[468]. To understand whether sequence similarity is constant across the length of the genes, a sliding window analysis was performed using Spider version 1.5.0[469], measuring Kimura 2-parameter (K2P) distance[470] between the *TPRX1* and *TPRX2* genes of a given species in 50 base pair (bp) overlapping windows with increments of 1 bp. Only species with at least one putatively functional copy of both *TPRX1* and *TPRX2* were used.

Third, we tested for incompatibilities between phylogenies built using different partitions of the genes. The HyPhy[471,472] programme GARD[473,474] was run using Datamonkey[475] with default parameters on codon alignments of all *TPRX1* and *TPRX2* genes (GARD was also run on alignments of *Oryctolagus cuniculus* and *Microcebus murinus LEUTX* tandem duplicates and *Peromyscus leucopus TPRX* and *LEUTX* genes). GARD uses an aggressive population-based hill climber to search multiple sequence alignments for phylogenetic incongruity and identify putative gene conversion and recombination breakpoints. The AIC_C (small sample Akaike Information Criterion) was used to select the model with the best fit to the data, with Akaike weights (w_i) calculated using the R package qpcR version 1.4.1[476–481]. The alignment was split into two partitions based on the breakpoint identified by GARD, and Bayesian phylogenies of each partition built as above. To measure tree dissimilarity, the tqDist algorithm[482] was implemented in the R

package Quartet version 1.2.0[483] to calculate quartet distance[484] and quartet divergence[485].

Finally, GENECONV version 1.81a[486] was used to identify putative gene conversion events by searching for fragments of sequences with sufficient nucleotide similarity to suggest gene conversion. GENECONV was run with default parameters apart from /lp (implements pairwise comparisons), /w123 (creates reproducible results by initiating at the same random seed number) and -gscale=2 (allows mismatches in the conversion tracts with a penalty of 2). GENECONV returns multiple comparison-corrected *p*-values based on 10,000 permutations for fragments found with global and pairwise sequence comparisons. Given a significance threshold of 0.05, it is expected that if there was no gene conversion in the dataset 69 of the 1378 pairwise comparisons would produce false positives. We identified 613 events, suggesting that the majority are not false positives. Furthermore, as a negative control, GENECONV was run with option -Randomize_sites; this permutes sites once and therefore removes any gene conversion signal. This identified just seven gene conversion events, again suggesting that the events detected above are not false positives. Events identified in the negative control analysis were discarded from the results. Only putatively functional genes were included in the gene conversion analysis, and *Mus musculus* and *Peromyscus leucopus Obox* (*TPRX2*) genes were also omitted because they show extreme lineage-specific sequence divergence and their inclusion may disrupt analysis.

2.3.4 Tests for accelerated divergence and positive selection

Tajima's relative rate test[487] (α = 0.05) was used to test for changes in the rate of homeodomain sequence evolution. Each ETCHbox homeodomain was compared to its conspecific CRX protein using a marsupial CRX sequence (*Monodelphis domestica*) as an outgroup. Where there are lineage-specific duplications, only one duplicate was used.

The Benjamini-Yekutieli[488] false discovery rate method (false discovery rate = 0.05) was used to correct for multiple testing as it does not require independence of tests.

Episodic positive selection in ETCHbox homeodomains was detected using the HyPhy[471,472] Branch-Site Unrestricted Statistical Test for Episodic Diversification (BUSTED)[489] via Datamonkey[475] with default parameters using codon alignments generated with PAL2NAL[490] and phylogenies reflecting known species relationships. CODEML (PAML version 4.8)[491,492] was used to test for pervasive positive selection (Model = 0, NSsites = 0, 1, 2, 7, 8) with the option cleanData = 0. Likelihood ratio tests (LRTs) were used to compare model 2 (M2, positive selection model) to model 1 (M1, nearly neutral model) and model 8 (M8, beta and ω model – positive selection) to model 7 (M7, beta model – no positive selection).

The HyPhy Mixed Effects Model of Evolution (MEME)[493], which uses mixed-effects branch-site models, was used to detect specific codon sites evolving under episodic positive selection. MEME is preferred to the branch-site mode of CODEML because it does not require *a priori* specification of branches to be tested but retains good statistical power[494]. The position of residues in relation to homeodomain structure was deduced by comparative structural modelling to the PRD-class homeodomain of *Drosophila melanogaster* Aristaless in complex with DNA (RCSB Protein Data Bank entry 3LNQ)[495,496] using Modeller[497] implemented in UCSF Chimera 1.15[498]. RELAX[499] was run with default parameters using codon alignments of ETCHbox and *CRX* homeoboxes to test for relaxed selection in each ETCHbox gene versus a reference group of six *CRX* genes (*Canis lupus familiaris*, *Condylura cristata*, *Equus caballus*, *Homo sapiens*, *Mus musculus* and *Ovis aries*).

Sites with a gap in more than 50% of sequences were removed from selection analyses. *Mus musculus* and *Peromyscus leucopus Obox* (*TPRX2*) genes and genes with

frameshifts or early stop codons downstream of the homeodomain were also omitted from selection analysis. Furthermore, the phylogenetic incongruity caused by gene conversion could lead to inaccurate results when testing for selection[474,500,501]. To account for gene conversion in the *TPRX* genes, we used the gene conversion breakpoint identified by GARD[473,474] to partition the alignment into two sections. MrBayes[451,452] was used as above to calculate gene trees for each partition. The above methods were then performed separately for each of the two partitions.

2.4 Results

2.4.1 Identification of ETCHbox genes in eutherian genomes

We first characterised the ETCHbox genes of *Bos taurus* (cattle) using transcriptome data (Figure 2.1). *B. taurus* possesses putatively functional *ARGFX*, *LEUTX*, *TPRX1* and *TPRX2* genes, but *DPRX* is a putative pseudogene due to a 2 bp insertion in the homeobox and the loss of exon 1; *PARGFX* has been lost. *B. taurus* also possesses a *TPRX* duplicate, which we refer to as *TPRX3*. Compared with *Homo sapiens* (human), *B. taurus ARGFX* has an additional 5' coding exon, which extends the reading frame. All genes are located in the same syntenic position as in humans, with *LEUTX*, *TPRX1*, *TPRX2* and *DPRX* (and *TPRX3*) in a loose cluster on chromosome 18 (human chromosome 19), and *ARGFX* separate from the cluster on chromosome 1 (human chromosome 3).

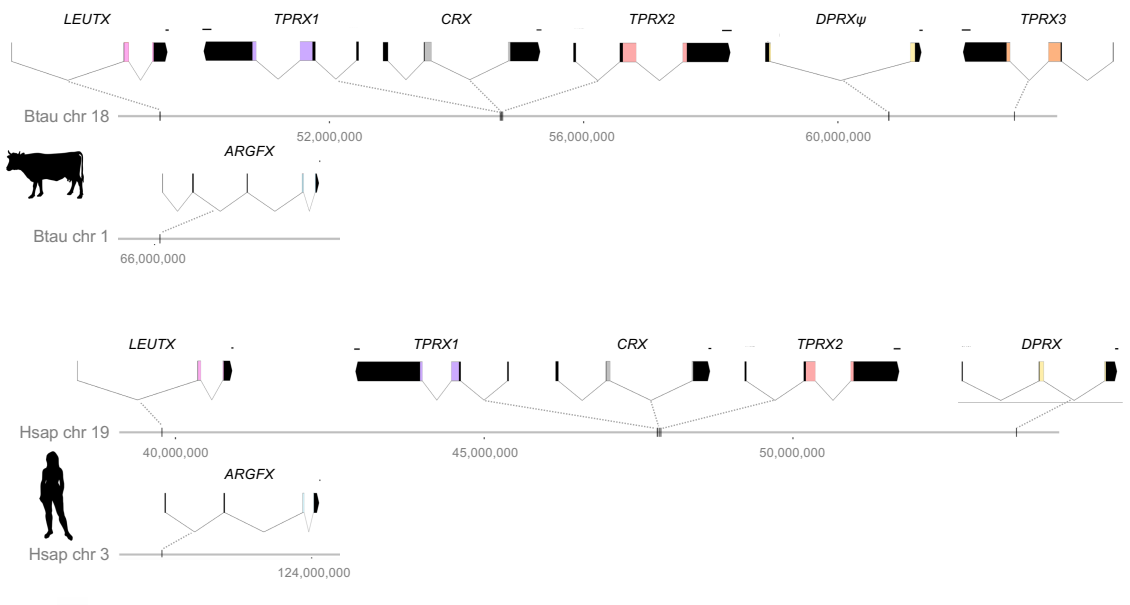


Figure 2.1 ETCHbox repertoires of *Homo sapiens* (humans) and *Bos taurus* (cattle). Gene structures are determined using transcriptome assemblies. Horizontal grey bars represent chromosomes, vertical black bars represent the genomic position of ETCHbox genes. For gene structure representations, coding regions are shown in black, homeoboxes in colour. Untranslated regions (UTRs) are not shown. Black scale bars at 3' end of genes = 100 bp. *DPRX*, *LEUTX*, *TPRX1*, *TPRX2* (and *B. taurus TPRX3*) form a loose cluster on a single chromosome (*B. taurus* chromosome 18, *H. sapiens* chromosome 19); *ARGFX* has translocated to another chromosome (*B. taurus* chromosome 1, *H. sapiens* chromosome 3). *TPRX1* and *TPRX2* are located either side of the ETCHbox 'ancestor' *CRX*.

We then characterised the ETCHbox gene repertoires in the genomes of 31 further eutherian species, using a combination of phylogenetics, synteny and reciprocal BLAST searches to assign gene identities (Figure 2.2). These methods concurred in almost all cases except for *TPRX1* and *TPRX2* genes, where sequence-based methods disagreed with genomic position; for these genes we use genomic position to assign gene name and assess below whether incongruence is due to gene conversion. The only other discordance occurs in *Peromyscus leucopus* (white-footed mouse), where the two genes at the *LEUTX* locus cluster with rodent *TPRX1* genes, although we find no evidence of gene conversion in this case (GENECONV analysis, no gene conversion fragments identified). In phylogenetic analyses, branch lengths are longer for ETCHbox genes than for their paralogues *CRX* and *OTX1*, indicating a higher amino acid substitution rate. Particularly long branches are observed for *Oryctolagus cuniculus* (European rabbit)

TPRX2 and *LEUTX1*, *Microcebus murinus* (gray mouse lemur) *PARGFX*, and *Mus musculus* (house mouse) and *P. leucopus TPRX1* (= *Crxos*) and *TPRX2* (= *Obox*). Loci with a stop codon, frameshift or splice site disruption in, or upstream from, the homeodomain are inferred to be pseudogenes. The ETCHbox genes frequently spawn retrocopies; these were also characterised, with every sampled species possessing at least one ETCHbox retrogene (Supplementary Table S2.3a); retrocopies are not clustered, and are found dispersed around the genome (e.g. *Homo sapiens* and *Bos taurus;* Supplementary Table S2.3b).

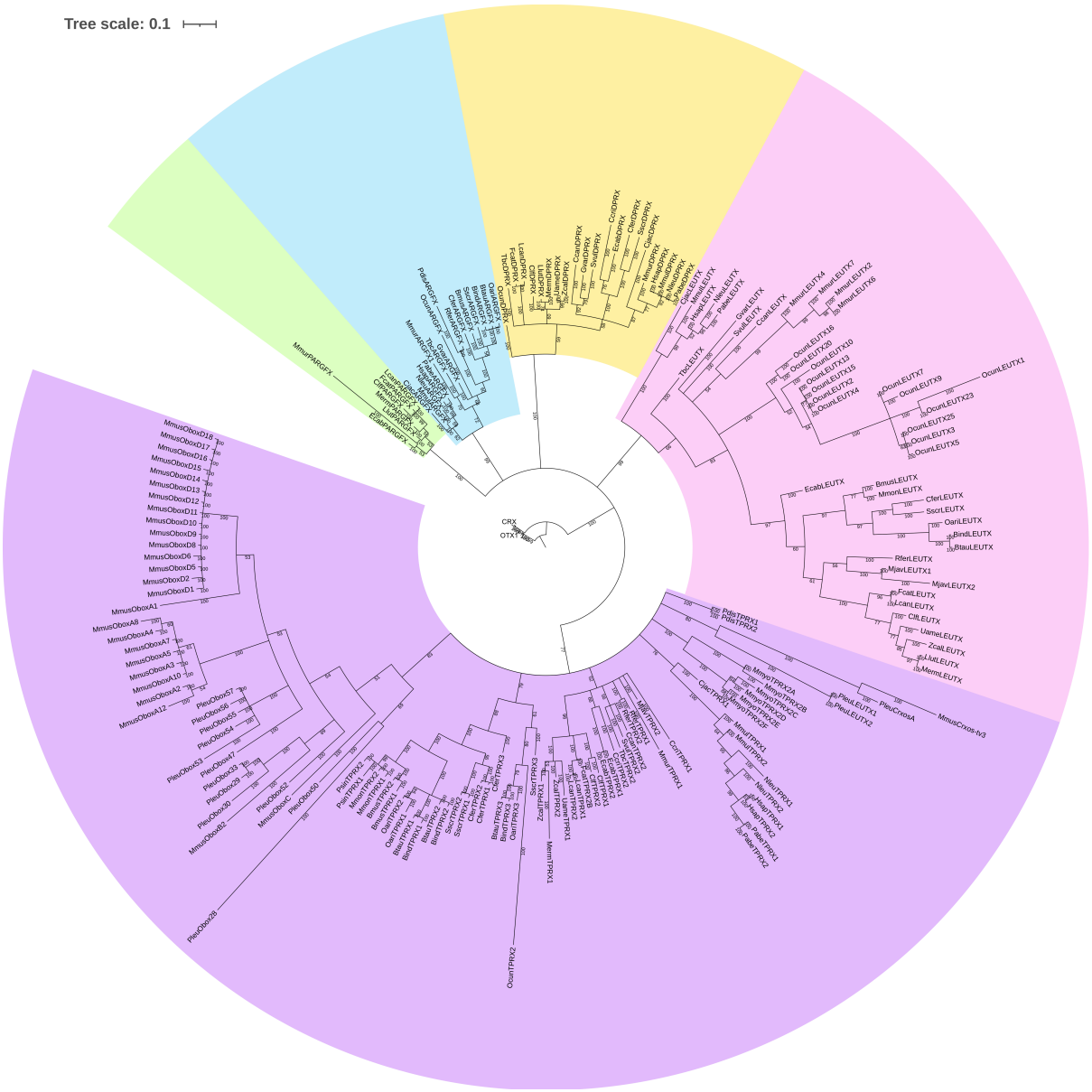


Figure 2.2 Bayesian gene tree of putatively functional ETCHbox genes identified in this work. Colours highlight ETCHbox gene families; labels show posterior probabilities. The *ARGFX*, *DPRX*, *LEUTX* and *PARGFX* clades are supported by ≥99% probabilities. Due to the limited length of the homeodomain (60 amino acids), gene trees do not always recapitulate known relationships between species. *TPRX* duplicates in Cetartiodactyla are referred to as *TPRX3*. The *TPRX1* and *TPRX2* genes of *Mus musculus* and *Peromyscus leucopus* are referred to as *Crxos* and *Obox* respectively, reflecting their extensive sequence change compared to the ancestral *TPRX* genes. Abbreviations: Bind = *Bos indicus*, Bmus = *Balaenoptera musculus*, Btau = *Bos taurus*, Ccan = *Castor canadensis*, Ccri = *Condylura cristata*, Cfer = *Camelus ferus*, Cjac = *Callithrix jacchus*, Clf = *Canis lupus familiaris*, Ecab = *Equus caballus*, Fcat = *Felis catus*, Gvar = *Galeopterus variegatus*, Hsap = *Homo sapiens*, Lcan = *Lynx canadensis*, Llut = *Lutra lutra*, Mjav = *Manis javanica*, Merm = *Mustela erminea*, Mmon = *Monodon monoceros*, Mmul = *Macaca mulatta*, Mmur = *Microcebus murinus*, Mmus = *Mus musculus*, Mmyo = *Myotis myotis*, Nleu = *Nomascus leucogenys*, Oari = *Ovis aries*, Ocun = *Oryctolagus cuniculus*, Pabe = *Pongo abelii*, Pdis = *Phyllostomus discolor*, Pleu =

Peromyscus leucopus, Psin = *Phocoena sinus*, Rfer = *Rhinolophus ferrumequinum*, Sscr = *Sus scrofa*, Svul = *Sciurus vulgaris*, Tbc = *Tupaia belangeri chinensis*, Uame = *Ursus americanus*, Zcal = *Zalophus californianus.*

ETCHbox gene repertoires are highly variable between species, with additional tandem gene duplication, pseudogenisation and gene loss occurring repeatedly across eutherians (Figure 2.3). Previous work showed that all six ETCHbox genes were present in the ancestor of the Boreoeutheria[90] so absence at a terminal node implies gene loss. All sampled species have lost at least one ETCHbox gene, and each gene has been lost in at least one sampled species. Some gene losses are inferred to have occurred in the ancestors of large clades (e.g. *ARGFX* in the Carnivora and *DPRX* in the Cetruminantia); many other losses are more recent (e.g*. LEUTX* is lost in *Phocoena sinus* [vaquita] but present in other sampled Cetacea species).

In *Mi. murinus*, we note the first putatively functional *PARGFX* gene reported for any member of the Euarchontoglires. *Mi. murinus PARGFX* is in the expected syntenic position and groups phylogenetically with other *PARGFX* genes, albeit on a long branch (Figure 2.2). *Galeopterus variegatus* (Sunda flying lemur) also has a detectable *PARGFX* locus, but it is inferred to be a pseudogene.

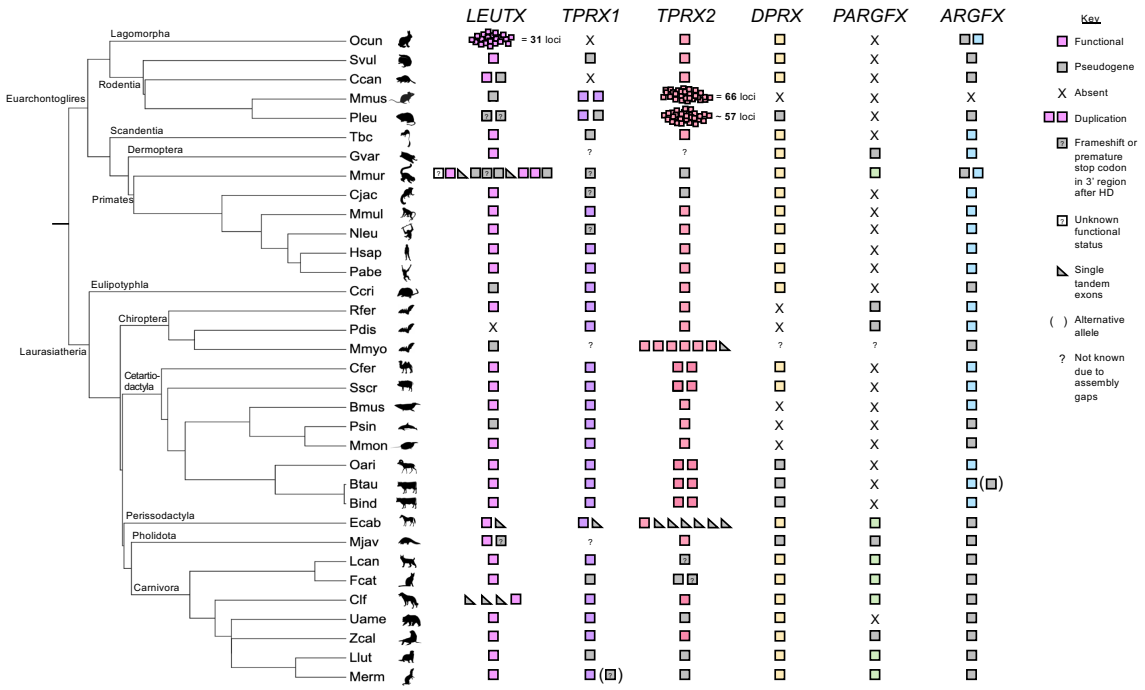


Figure 2.3 ETCHbox repertoires of 34 eutherians. Phylogenetic relationships are based on TimeTree[461]. Coloured boxes = putatively functional gene. Multiple boxes = gene duplicates. Black X = no gene remnants (complete gene loss). Grey boxes = putative pseudogene; grey boxes with a black question mark = complete homeodomain but subsequent frameshift or premature stop codon. White boxes with a question mark = unclear functional status due to incomplete assembly in the region. Grey triangles = tandem single exons. Brackets = polymorphism; question marks = assembly gap such that gene presence or absence cannot be determined. HD = homeodomain. Species abbreviations as in Figure 2.2.

2.4.2 Giant arrays of ETCHbox genes

We identify several arrays of tandem ETCHbox duplicates, one of the largest of which is an array of *LEUTX* loci in *O. cuniculus*. Previous work detected six loci[95], whereas we detect 27 gene copies in the assembly analysed, of which 14 are putatively functional, 11 are putative pseudogenes and two are of uncertain functional status (Supplementary Figure S2.1). We also find four single exons in the cluster, giving a total of 31 loci. This is the largest *LEUTX* expansion discovered and one of the largest ETCHbox expansions, smaller than only those of *Mu. musculus*[96] and *P. leucopus Obox* (*TPRX2*) genes. An inversion on *O. cuniculus* chromosome 5 has split the array in two, with *LEUTX1* to *LEUTX5* approximately 9 Mb from *LEUTX6* to *LEUTX27*. *Mi. murinus* also has a tandem *LEUTX* expansion of 10 loci, at least three of which are putatively functional, and *Mi.*

murinus and *O. cuniculus* are both also notable because they have an *ARGFX* duplication. We find several cases of tandem duplication of single exons, including at *Equus caballus* (domestic horse) *LEUTX*, *TPRX1* and *TPRX2* loci.

It was shown previously that *Mu. musculus* has lost *ARGFX*, *DPRX*, *LEUTX* and *PARGFX* and possesses two *TPRX1* copies (called *Crxos*) and 66 *TPRX2* loci (called *Obox*), all of which are highly divergent in sequence[90,96]. We asked when the transition to this highly derived state occurred. Our results indicate that this evolved within the rodents. *Sciurus vulgaris* (red squirrel, Sciuridae) and *Castor canadensis* (American beaver, Castoridae) possess putatively functional *DPRX* and *LEUTX* genes, and neither have *TPRX1* or *TPRX2* duplicates (Figure 2.3). However, *Peromyscus leucopus* (white-footed mouse, Cricetidae) has two *TPRX1* loci, and no functional *ARGFX*, *DPRX* or *PARGFX*, as in *Mu. musculus.* Furthermore, we detect 57 *P. leucopus TPRX2* (*Obox*) loci, of which 12 are putatively functional. Seven of these loci have escaped the *TPRX2* cluster on chromosome 1 and form a separate cluster on chromosome 12. The observation that *P. leucopus TPRX1* and *TPRX2* genes group phylogenetically with *Mu. musculus Crxos* and *Obox* respectively (Figure 2.2) is consistent with the Notung result that the *TPRX1* duplication and several of the *TPRX2* duplications occurred before the split of *Mu. musculus* and *P. leucopus* (below). This suggests that the transition from *TPRX1* and *TPRX2* to *Crxos* and *Obox*-like states occurred before the split of the Muridae and Cricetidae.

2.4.3 Rates of gene duplication

We compared rates of gene duplication and loss for each gene by modelling a stochastic birth-death process using CAFE[457,458], giving maximum likelihood estimates for the rates of ETCHbox gene gain and loss (events per million years; λ and μ respectively; Table 2.1). CAFE was also used to infer likely ancestral gene numbers (Supplementary Figure S2.2). Rates of gain (λ) and loss (μ) are highly variable between ETCHbox families.

TPRX2 is the gene most prone to duplication (λ = 0.016) and *PARGFX* is most prone to gene loss (μ = 0.019). *ARGFX*, *DPRX* and *PARGFX* have very low rates of gene gain but relatively high rates of loss.

We tested for an acceleration in the rate of duplication of ETCHbox genes versus all other homeobox genes, and find evidence that *LEUTX* (p = 0.003) and *TPRX2* (p = 0.000) have indeed duplicated at a higher rate within the eutherians. The Viterbi assignment method[458] reveals that the high overall duplication rate of *LEUTX* is primarily a result of changes along the *O. cuniculus* ($p = 1.503 \times 10^{-8}$) and *Mi. murinus* (p = 0.028) branches; the high duplication rate of *TPRX2* is influenced largely by the branches leading to Cetartiodactyla (p = 0.027), *Myotis myotis* (greater mouse-eared bat, p = 0.001), Muroidea ($p = 2.774 \times 10^{-10}$), *Mu. musculus* ($p = 1.774 \times 10^{-36}$) and *P. leucopus* ($p$ = 0.003).

To validate the results of the probabilistic approach of CAFE, the parsimony algorithm of Notung[459,460] was used to estimate the number of ETCHbox duplication and loss events and infer the timings of such events (Table 2.1 and Supplementary Figure S2.3). A high duplication rate for *LEUTX* and *TPRX2* was supported by analysis incorporating gene trees. Gene loss is expected to have most functional relevance when a single copy gene transitions to total absence of a functional gene; we find this occurred most for *PARGFX,* in accordance with CAFE results. There are two cases of apparent gene turnover overestimation by the Noting algorithm: Notung reports the *Camelus ferus* (Bactrian camel) *TPRX3* duplication as independent of other Cetartiodactyla *TPRX3* duplicates, and a *DPRX* duplication at the base of the Caniformia followed by multiple losses. These are likely artefacts caused by the rapidly-evolving nature of ETCHbox sequences but do not distort the overall inferences from the analysis.

Gene	λ (gains per gene per million years, CAFE)	μ (losses per gene per million years, CAFE)	Estimated number of gene duplication events (Notung)	Estimated number of gene loss events (Notung)	Number of species with at least one putatively functional gene
ARGFX	3.37E-10	6.70E-03	0	5	17
DPRX	4.26E-11	3.42E-03	1	8	22
LEUTX	8.88E-03	3.32E-03	18	4	29
PARGFX	5.84E-11	1.93E-02	0	10	7
TPRX1	8.49E-04	6.15E-03	1	10	25
TPRX2	1.61E-02	6.33E-03	41	7	28

Table 2.1 Duplication and loss of ETCHbox genes. Probability of duplication or loss (λ and μ) for each ETCHbox gene, estimated by CAFE's stochastic birth and death model[457,458], together with estimates of numbers of gene duplication and gene loss events, calculated by Notung[459,460]. Pseudogenes are excluded as duplication events.

2.4.4 Polymorphism in ETCHbox genes

We find two cases of ETCHbox intraspecific polymorphism where one allele has a frameshift mutation. In *Mustela erminea* (stoat), we identify a putatively functional *TPRX1* in one haplotype of the phased genome assembly while the alternate haplotype has a 'CC' dinucleotide insertion causing a frameshift in exon 3. In the *B. taurus* reference genome (ARS-UCD1.2), we find a 13 bp deletion in *ARGFX* exon 2 that causes a frameshift and a premature stop codon before the homeobox, making it a putative pseudogene. We do not find this deletion in several other *B. taurus* datasets (Supplementary Table S2.2) or in the genome of other Bovidae species (Supplementary Table S2.4).

2.4.5 *TPRX1* and *TPRX2* have been the subject of repeated gene conversion

Interlocus gene conversion is a naturally occurring 'copy-and-paste' process that can take place during double-strand break repair, where DNA sequence from one locus is erroneously used as the template for the repair of DNA sequence at a different locus in the same genome[245]. The incongruence between gene identity inferred from phylogenetics versus gene position for *TPRX1* and *TPRX2* suggests that gene

conversion may have occurred between these loci, as suggested previously[90]. However, this hypothesis needs further testing, and it is currently unclear whether the complete loci were affected, when it occurred or how often it occurred in evolution.

We first investigated these questions using a phylogenetic approach, searching for incompatibilities between the known species tree and the inferred gene tree. Under the null hypothesis of no gene conversion, *TPRX1* and *TPRX2* genes would form separate clades diverging since the base of eutherians; gene conversion would result in paralogues grouping more closely together. Bayesian nucleotide phylogenies of putatively functional *TPRX* genes reveal eight cases where the *TPRX1* and *TPRX2* genes from a given species group together as pairs of sister sequences, suggesting recent gene conversion events in these lineages (Figure 2.4a, blue boxes). There are also indications of further gene conversion events deeper in the phylogeny, notably in the stem lineages of Cetacea, Bovidae, Carnivora and Primates (Figure 2.4a, blue dots). Intriguingly, we found evidence for additional gene conversion events when phylogenetic analysis was restricted to the homeobox sequence only. This revealed 13 recent conversion events between *TPRX* loci, with five new cases identified in addition to the eight above (Figure 2.4b, pink boxes). Several of the additional events are nested within the clades that showed evidence of older gene conversion (Primates, Cetacea), suggesting successive gene conversion events in evolution. The occurrence of successive gene conversion events is also supported by analysis of pairwise nucleotide identity (Supplementary Figure S2.4) which, for example, suggests gene conversion at the base of the Cetartiodactyla, then further events within the Bovidae, Cetacea, *Sus scrofa* (domestic pig) and *C. ferus*.

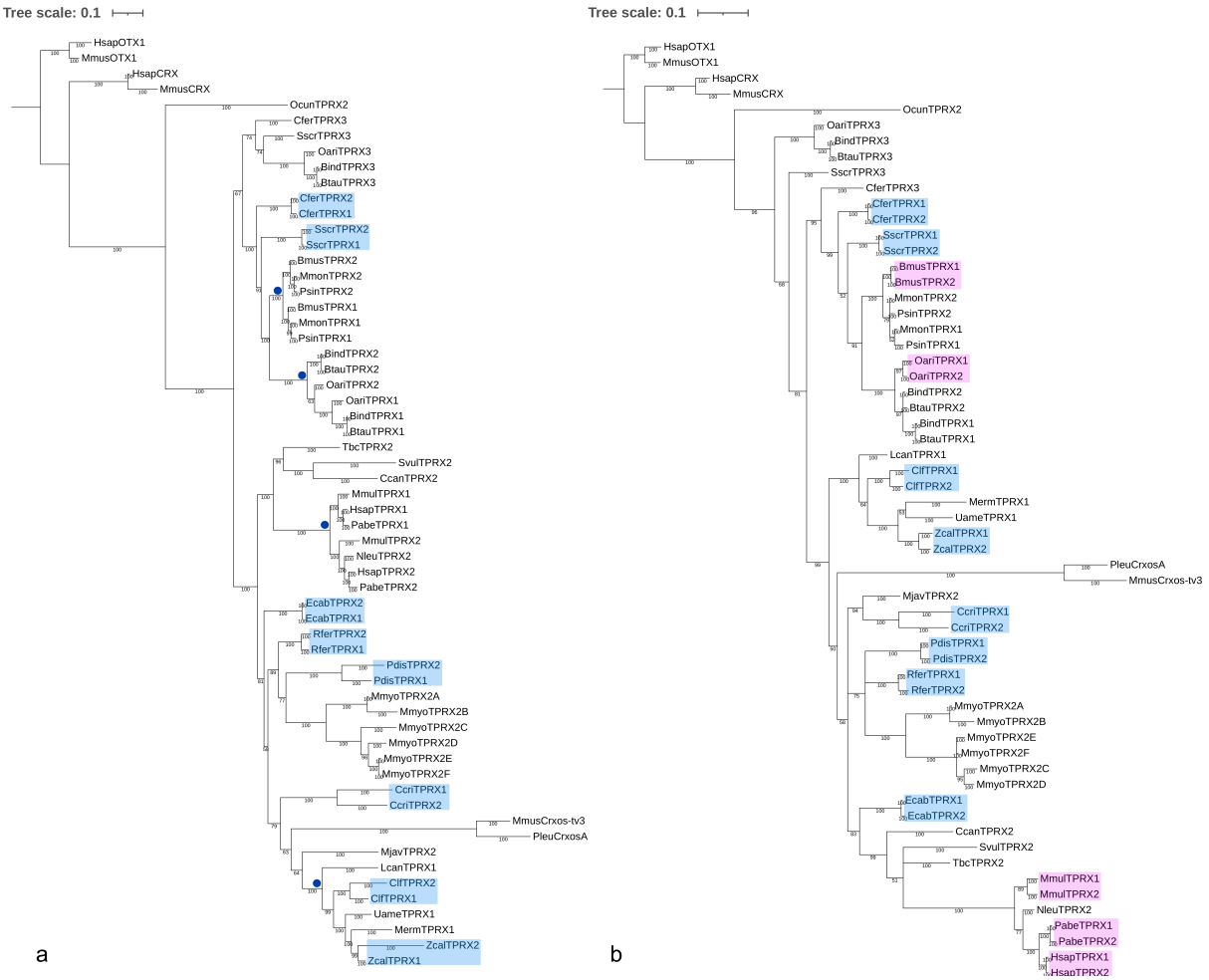


Figure 2.4 Bayesian phylogenies of putatively functional *TPRX1*, *TPRX2* and *TPRX3* full gene sequences (A) and homeoboxes (B). Blue boxes highlight cases where conspecific *TPRX1* and *TPRX2* pairs are more closely related to each other than to other sequences. Pink boxes highlight cases that appear on tree b but not tree a. Blue dots mark putative gene conversion events that occurred deeper in the phylogeny. Putative pseudogenes were excluded. Labels show posterior probabilities. Gene trees created in MrBayes using general time reversible (GTR) model. Alignment lengths 560 bp and 180 bp for trees A and B, respectively. Species abbreviations as in Figure 2.2.

Since a homeobox-only tree suggests additional episodes of gene conversion, we hypothesised that the 5' region of *TPRX* genes is more prone to gene conversion than the 3' region. To test this, we conducted a sliding window analysis calculating pairwise Kimura 2-parameter (K2P) distances[470] between *TPRX1* and *TPRX2* sequences (Figure 2.5). In *Bos indicus* (Zebu cattle), *B. taurus*, *Balaenoptera musculus* (blue whale), *C. ferus*, *Felis catus* (domestic cat), *H. sapiens*, *Lynx canadensis* (Canada lynx), *Monodon monoceros* (narwhal), *Nomascus leucogenys* (northern white-cheeked gibbon), *Ovis*

aries (sheep), *P. sinus* and *S. scrofa*, the lowest sequence distances (highest similarities) are at the 5' end, suggesting this region is more prone to homogenisation via gene conversion.

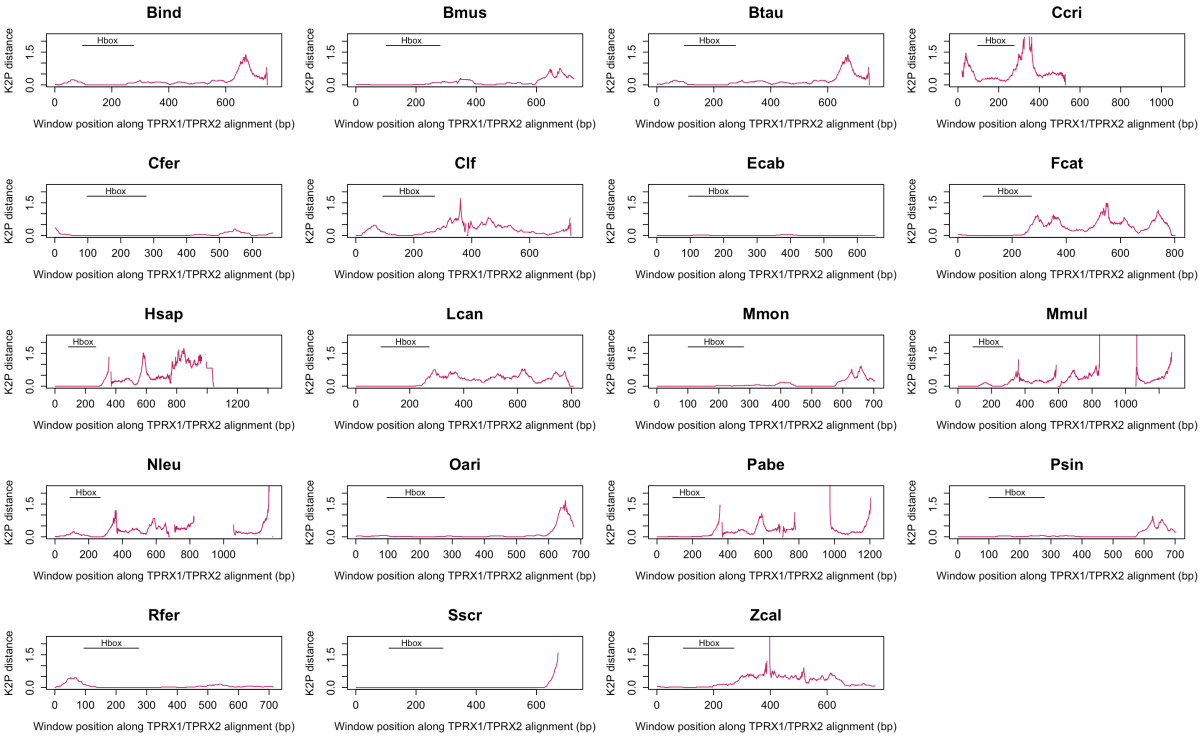


Figure 2.5 Sequence similarity between *TPRX1* and *TPRX2* genes within a species. Plots show the Kimura 2-parameter (K2P) distance in 50 bp sliding windows between conspecific *TPRX1* and *TPRX2* genes. Higher K2P values indicate more divergent sequences. Gaps in the trace indicate indels in the alignment. The black bar marked 'Hbox' demarcates the position of the homeobox in each alignment. For many species, the K2P values increase towards the 3' end of the gene, suggesting that the *TPRX* genes have been homogenised by gene conversion less at their 3' ends. Putative pseudogenes were excluded. Species abbreviations as in Figure 2.2.

Gene conversion occurring repeatedly in one region of a gene pair is predicted to result in differences between phylogenetic trees built from different sub-regions of the genes. Using GARD[473,474], which tests for phylogenetic incongruence within a gene, we identified a consistent putative gene conversion breakpoint, located immediately downstream of the homeobox, dividing the gene into two regions with different phylogenetic histories (null model AIC_C = 54094.1, breakpoint model AIC_C = 52467.8, ΔAIC_C = 1626.3; null model Akaike weight (w_i) = 0, breakpoint model Akaike weight (w_i)

= 1; breakpoint model receives 100% of the weight of the models compared). Bayesian nucleotide sequence phylogenies of partition 1 (including the homeobox) and partition 2 (downstream) show different topologies (Figure 2.6; quartet distance = 41,506; quartet divergence = 0.102). Partition 1 trees show more gene conversion events than partition 2 (e.g. Bovidae, Cetacea and Primates in Figure 2.6), reinforcing the hypothesis that the 5' region is more prone to gene conversion.

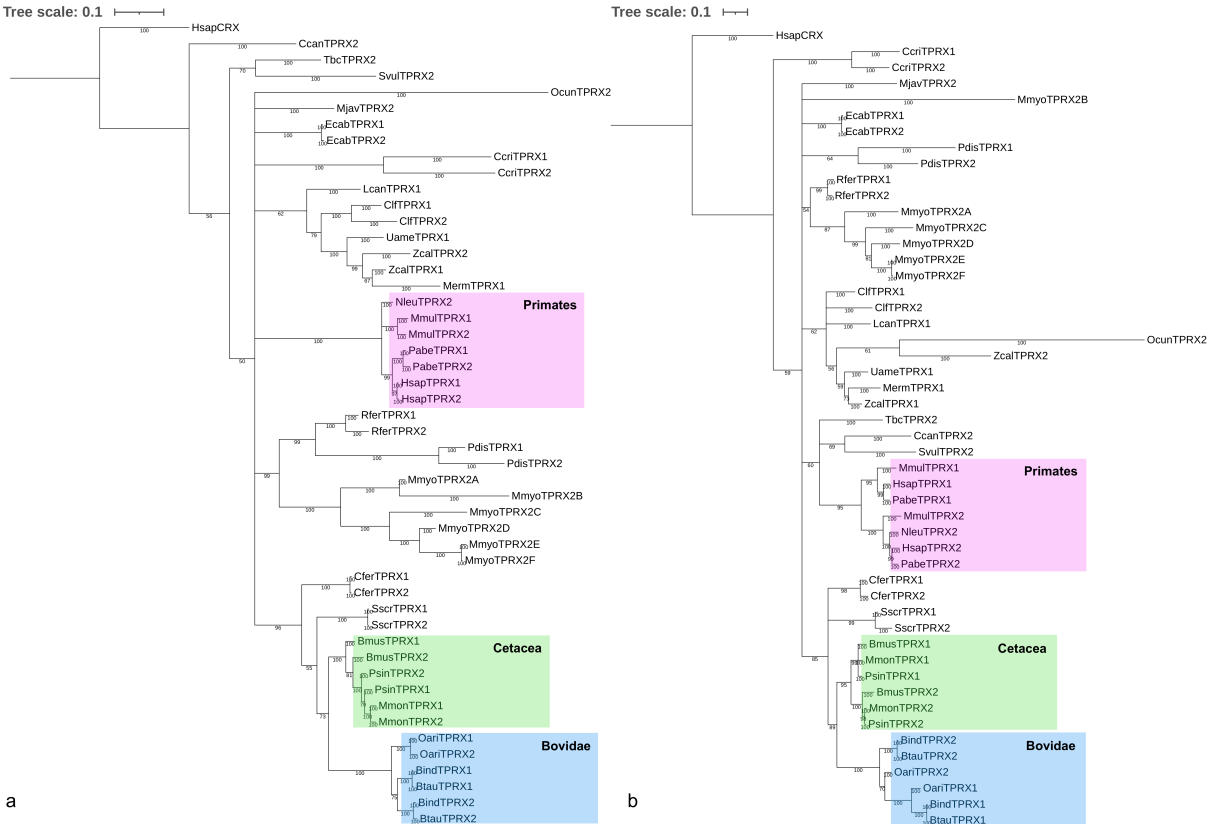


Figure 2.6 Bayesian phylogenies inferred using partition 1 (A) and partition 2 (B) of putatively functional *TPRX1* and *TPRX2* genes split at the gene conversion breakpoint identified by GARD. Boxes highlight the Bovidae, Cetacea and Primates, where topology differs markedly between the two trees. For example, Tree b is consistent with a gene conversion event at the base of the Primates; Tree a has conspecific pairs of *TPRX* genes consistent with additional more recent gene conversion events in the ancestors of these species within the Primates. Species abbreviations as in Figure 2.2.

Finally, we used GENECONV[486] to search for long regions of unusually high sequence identity in multiple sequence alignments as further evidence of gene conversion. In the 53 TPRX sequences analysed, GENECONV identifies eight gene conversion events by

global comparisons (Supplementary Table S2.5) and 613 fragments by pairwise comparisons (Supplementary Table S2.6), ranging from 9 to 492 bp in length. Pairwise comparisons are particularly powerful for detecting very recent gene conversion. For example, we find evidence for 10 'species-specific' gene conversion events (Supplementary Table S2.7), eight of which were also detected by phylogenetic methods tree (Figure 2.4b). Notably, GENECONV and GARD[473,474] give similar locations for the position of gene conversion breakpoints between 5' and 3' regions, and both show that the upstream region is subject to more frequent gene conversion than the downstream region. Across all species, no fragments have GENECONV breakpoints downstream of position 492 of the 2331 bp multiple sequence alignment (Supplementary Figure S2.5); in the majority of species, position 492 is very close to the putative gene conversion breakpoint identified by GARD, and in human they are only nine nucleotides apart (a large insertion in bats means that they are further apart in the full alignment, Supplementary Figure S2.6). This corroboration by two methods lends strong support to this partition, which is within exon 3, downstream of the homeobox.

We note that gene conversion continued to occur between the six *My. myotis TPRX2* duplicates following tandem duplication, with 13 fragments identified by GENECONV (Supplementary Table S2.6). Furthermore, gene conversion in the ETCHbox genes is not limited to *TPRX*. Pairwise analysis using GENECONV identifies 15 gene conversion events between *Mi. murinus LEUTX* tandem duplicates and 367 events between *O. cuniculus LEUTX* duplicates (Supplementary Table S2.8). Both results are reinforced by GARD (Supplementary Table S2.9).

2.4.6 Positive selection in ETCHbox genes

Using Tajima's relative rate test[487], we find that all ETCHbox sequences have a faster evolutionary rate than their sister gene *CRX* (124 genes analysed; Supplementary Table S2.10). To investigate if the elevated evolutionary rates are due to positive selection, we

used BUSTED[489] to test for episodes of selection that may vary over time and between lineages, and MEME[493] to identify specific sites under selection. These analyses were performed on complete coding sequences of *ARGFX*, *DPRX*, *LEUTX*, *PARGFX* and *TPRX*. However, since gene conversion leads to phylogenetic incongruity, which interferes with detecting positive selection[474,500,501], we divided *TPRX* genes at the gene conversion breakpoint identified by GARD into 5' and 3' regions and performed analyses separately on the two regions.

Using BUSTED[489], we detect evidence of episodic positive selection during the evolution of *ARGFX*, *DPRX*, *LEUTX*, *PARGFX* and both partitions of the *TPRX* genes (LRT $p < 0.05$ for all genes). We also find strong evidence for positive selection in *ARGFX*, *DPRX*, *LEUTX*, *TPRX* partition 1 and *TPRX* partition 2, but not *PARGFX*, using CODEML[491,492] (Supplementary Table S2.11), supporting the BUSTED result. Using MEME[493], we find evidence for positive selection acting on between 3 (*PARGFX*) and 31 (*TPRX*) codons in each gene (Supplementary Table S2.12). The sites deduced to have undergone positive selection are spread across the encoded proteins and include codons within homeodomains (*ARGFX* 3 sites; *DPRX* 1 site; *LEUTX* 8 sites; *TPRX* 4 sites; Supplementary Figure S2.7). The spatial locations of sites under positive selection within homeodomains were inferred by comparative modelling of human ETCHbox homeodomains to a known PRD class structure using Modeller[497] (Figure 2.7, Supplementary Figure S2.8). Sites under positive selection include those within the N-terminal arm of ARGFX (E4), LEUTX (Y1, P4, R7) and TPRX1/2 (Q1), and the recognition helix of ARGFX (S43).

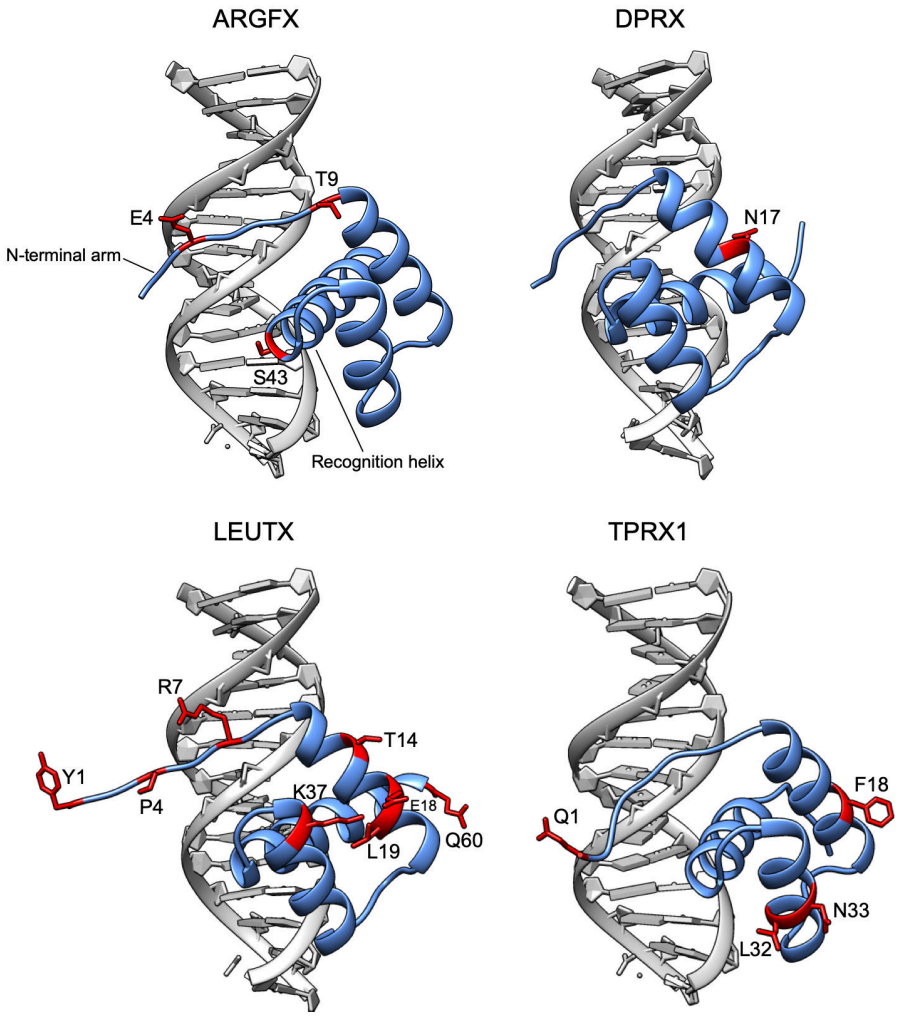


Figure 2.7 Models of ETCHbox homeodomain structures including sites under positive selection. Homeodomains of human ETCHbox proteins (blue) are modelled in complex with DNA (grey). Residues under positive selection are coloured red. Amino acid side chains are shown for sites under positive selection only. Letters show the identity of positively selected residues in human, numbers show their position within the homeodomain. TPRX1 and TPRX2 homeodomains are identical due to gene conversion so only one is shown.

Using RELAX[499], we also find evidence for relaxed selection in all ETCHbox genes compared with their sister gene *CRX* (Supplementary Table S2.13), suggesting that a combination of relaxed and positive selection is required to explain their fast evolutionary rate.

2.5 Discussion

After duplication from *CRX* in the lineage leading to eutherians, the ETCHbox genes underwent asymmetric evolution and continued to be duplicated and lost[90]. The genes are suspected to have important roles in preimplantation development and embryonic genome activation[90,104], making observed variability of ETCHbox gene sets a mystery. Here, we compared the ETCHbox complements of 34 species with the aim of illuminating the processes that have sculpted such varied repertoires. Restricting the analyses to genomes sequenced using long-read technologies allowed us to establish with confidence clear examples of gene duplication and secondary loss, something that was challenging in previous work based on lower-quality genome assemblies.

We find that, despite extensive and frequent gene loss, all sampled species possess at least two putatively functional ETCHbox genes. This retention suggests that the genes, collectively, are indispensable for eutherian development, and that fluctuations in gene number and rapid sequence evolution are not due to the lack of a function and neutrality. Previous work has shown that some ETCHbox genes can act in an antagonistic fashion, with gene sets upregulated by one gene overlapping with those downregulated by another[90,104]. This antagonism could explain why at least two different genes are always retained.

A second line of evidence supporting functionality is that all of the genes have been under recent positive selection, including at residues within the homeodomain. Residues in the N-terminal arm of the ARGFX, LEUTX and TPRX1 homeodomains, identified as having amino acid change driven by positive selection, are suggested by comparative modelling to interact with the minor groove of DNA. Residue 7 specifically, deduced to be under selection in LEUTX, is involved in sequence-specific contact in other homeodomain proteins and therefore may affect binding specificity[502,503]. ARGFX

homeodomain residue S43, also deduced to have been under positive selection, sits within the DNA-binding and specificity-determining recognition helix of PRD-class homeodomains[107] and is key for target site recognition[504]. These results suggest that there has been selection for altered DNA-binding properties in ETCHbox homeodomains. In addition, residues in homeodomain helices 1 and 2 are deduced to have been under selection in DPRX, LEUTX and TPRX proteins; modelling suggests that most of these residues are on the outer surface of the homeodomain. Since both of these helices have been proposed to mediate protein-protein interactions in some homeodomains, including those of the PRD class[505–509], we propose that this selection has altered ETCHbox protein-protein binding properties. Madissoon et al. (2016)[75] found that homeodomain differences are not sufficient to explain the differing transcriptional effects of PRD-like genes, implying that other protein domains also contribute to specificity; we therefore suggest that sites outside of the homeodomain that are under selection also influence specificity. Overall, our results suggest that there has been on-going and divergent selection for altered DNA-binding specificity and/or co-factor interactions in ETCHbox proteins, implying that functions have been modified as part of their rapid evolution during mammalian radiation. Experimental evidence supports these conclusions. For example, Royall et al. (2018)[96] found that, at some point during rodent evolution, *Crxos* (*TPRX1*) likely underwent a change in function to take on part of the role of *ARGFX.*

Though positive selection has contributed to changes in ETCHbox protein sequences, their timing of expression during development has remained relatively stable. Consistent with the results of Maeso et al. (2016)[90], we find that all sampled species possess at least one 'processed pseudogene' derived from an ETCHbox gene; these are generated by retrotransposition exclusively from genes expressed in the germline, including uncommitted early embryonic cells[97,98]. This suggests that across large phylogenetic distances the ETCHbox genes retain expression in the very early embryo. This is

corroborated by transcriptome data which showed that ETCHbox genes are expressed in preimplantation development in both humans (Euarchontoglires) and cattle (Laurasiatheria)[90].

Despite all eutherian mammals possessing at least two ETCHbox genes, there has been extensive gene loss. We find that of the six ETCHbox genes (*ARGFX*, *DPRX*, *LEUTX*, *PARGFX*, *TPRX1*, *TPRX2*), each has been lost in at least one sampled species, with *PARGFX* lost at the highest rate; furthermore, all sampled species have lost at least one ETCHbox gene. This pattern could be explained through a degree of genetic functional redundancy, whereby multiple genes perform similar functions and can partially substitute for each other, a pattern common after gene duplication[125,126,510]. Functional overlap could lead to relaxed selection, allowing repertoires to vary while an overall indispensable function is maintained. This suggestion is consistent with the finding of Maeso et al. (2016)[90] that gene sets regulated by *LEUTX* and *TPRX1* in human cells have a large degree of overlap. Partial redundancy between ETCHbox genes would not be without precedent: it is a common component of biological systems and is known to be a feature of other homeobox duplicates, including members of different *HOX* clusters in mammals[126,511–513]. Genetic redundancy can be evolutionarily stable and may be maintained by selection when, for example, one of the genes occasionally fails to perform a function successfully, or when genes possess other, non-redundant functions which are co-selected with redundant ones[126,514,515].

Gene duplication is a potential driver of functional innovation. Here we identify large tandem arrays of ETCHbox duplicates in several species, including *O. cuniculus* and *Mi. murinus LEUTX* and *P. leucopus* and *My. myotis TPRX2.* Further arrays have been previously described, such as the 66 Obox (*TPRX2*) loci of *Mu. musculus*[90,96]. It is likely that the propensity for tandem duplication stems from the position of these genes in a dynamic and unstable genomic region (chromosome 19 in human, 18 in *B. taurus*), in

which there is a high density of repetitive sequences, low density of recombination hotspots and elevated gene duplication rates[90,120,123,516], but the selective forces favouring retention of these duplicates are currently unclear. There are three main mechanisms by which duplications could be advantageous in the short term[228]: (1) by increasing gene dosage where function is dosage sensitive[234]; (2) by buffering against deleterious mutations[236,237]; and (3) the immediate emergence of a new function, for example due to the partial duplication of regulatory elements, or alteration of genomic location[239,240,517]. None of these explanations appears sufficient to explain the giant arrays observed for ETCHbox genes. The alternative is that the initial duplication event is selectively neutral[228], but duplicates are retained following either neofunctionalisation[187] or duplication-degeneration-complementation (DDC)[231]. Current data support this model. The high rates of pseudogenisation in the tandem arrays (45% for *O. cuniculus LEUTX*, 79% for *P. leucopus TPRX2*) suggest that some duplicates are selectively neutral and not actively retained, and previous studies have uncovered functional differences between *Mu. musculus Obox* duplicates, implying that sub- or neo-functionalisation has occurred following expansion of the tandem array[96].

Tandem gene duplicates can be subject to gene conversion, and we find overwhelming support that gene conversion has been a major force affecting *TPRX1* and *TPRX2* molecular evolution throughout the Boreoeutheria. Interestingly, these two genes are not directly adjacent to each other, but lie either side of the *CRX* locus. Gene conversion is expected to cause concerted evolution, meaning that instead of gene duplicates accumulating mutations independently they evolve in parallel, maintaining a higher than expected level of sequence similarity[200,247,251,518,519]. Gene conversion thus restricts the ability of duplicates to neofunctionalise, because their sequence is repeatedly homogenised and divergence is lost[250,251,520,521]. As genes diverge, the accumulation of many small mutations or fewer large sequence changes (e.g. transposable element insertion) can cause a threshold to be reached, at which point sequences differ enough

that gene conversion no longer occurs; at this stage, independent evolution commences and neofunctionalisation may take place[249–251]. The recent gene conversion events and high sequence similarities detected in this work suggest that this threshold is yet to be reached in the *TPRX* genes of most sampled lineages.

It is interesting to consider why *TPRX1* and *TPRX2* seem subject to such frequent gene conversion events, and why this has continued over long time periods across diverse lineages. One possibility is that the genes are dosage-sensitive with a beneficial effect if dosage is increased, as this can cause gene conversion to be favoured by selection[248]. Regardless, we suggest that gene conversion will affect the strength of selection on *TPRX* genes[251]. For example, gene conversion can lead to faster adaptation because alleles can be transferred between paralogues, enabling the spread of beneficial mutations and elimination of deleterious ones[245,520,522,523]. It is also expected to lead to faster adaptation through increasing effective population size, which enhances the efficiency of selection occurring within a gene family[523]. Overall, gene conversion has been a critical factor driving *TPRX1* and *TPRX2* evolution and is predicted to influence their functional role.

2.6 Conclusions

The ETCHbox genes represent an example of the recruitment of eutherian mammal-specific homeobox genes to a very early developmental stage, making them a promising model to study the evolution of young, lineage-specific homeobox genes. Our data show that, unlike the vast majority of homeobox genes, they have been subject to frequent tandem duplications and gene losses over relatively short evolutionary timescales, leading to varied ETCHbox repertoires even amongst closely related species. This includes newly discovered large tandem arrays of homeobox genes. The data also suggest that the ETCHbox genes are indispensable to eutherian preimplantation

development and that positive selection has continued to modify their functions. Finally, we show that gene conversion between *TPRX1* and *TPRX2* has occurred on a striking number of occasions and prevented divergence of their homeodomains; the consequences of this for function are currently unclear. Overall, high rates of gene duplication and loss, extensive divergence, concerted evolution and positive selection have sculpted the varied ETCHbox repertoires that are observed across eutherians; our results support the idea that antagonism and redundancy are key factors in determining these unusual evolutionary patterns.

2.7 Acknowledgements and declarations

We thank Yichen Dai, Ignacio Maeso, Peter Mulhair, Rodrigo Pracana, Sebastian Shimeld and Sonia Trigueros for helpful discussions and advice, and Carlos Herrera-Úbeda and Cecy Price for contributions to preliminary work. We also thank two anonymous reviewers for valuable comments that improved the manuscript.

Funding

This work was supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) [grant number BB/M011224/1] and an Oxford-Wolfson Marriott BBSRC Graduate Scholarship.

Competing interests

The authors have no conflicts of interest to declare.

Data, material and code availability

The datasets supporting the conclusions of this article are included within the article and its Electronic Supplementary Material.

2.8 Supplementary figures

Figure S2.1



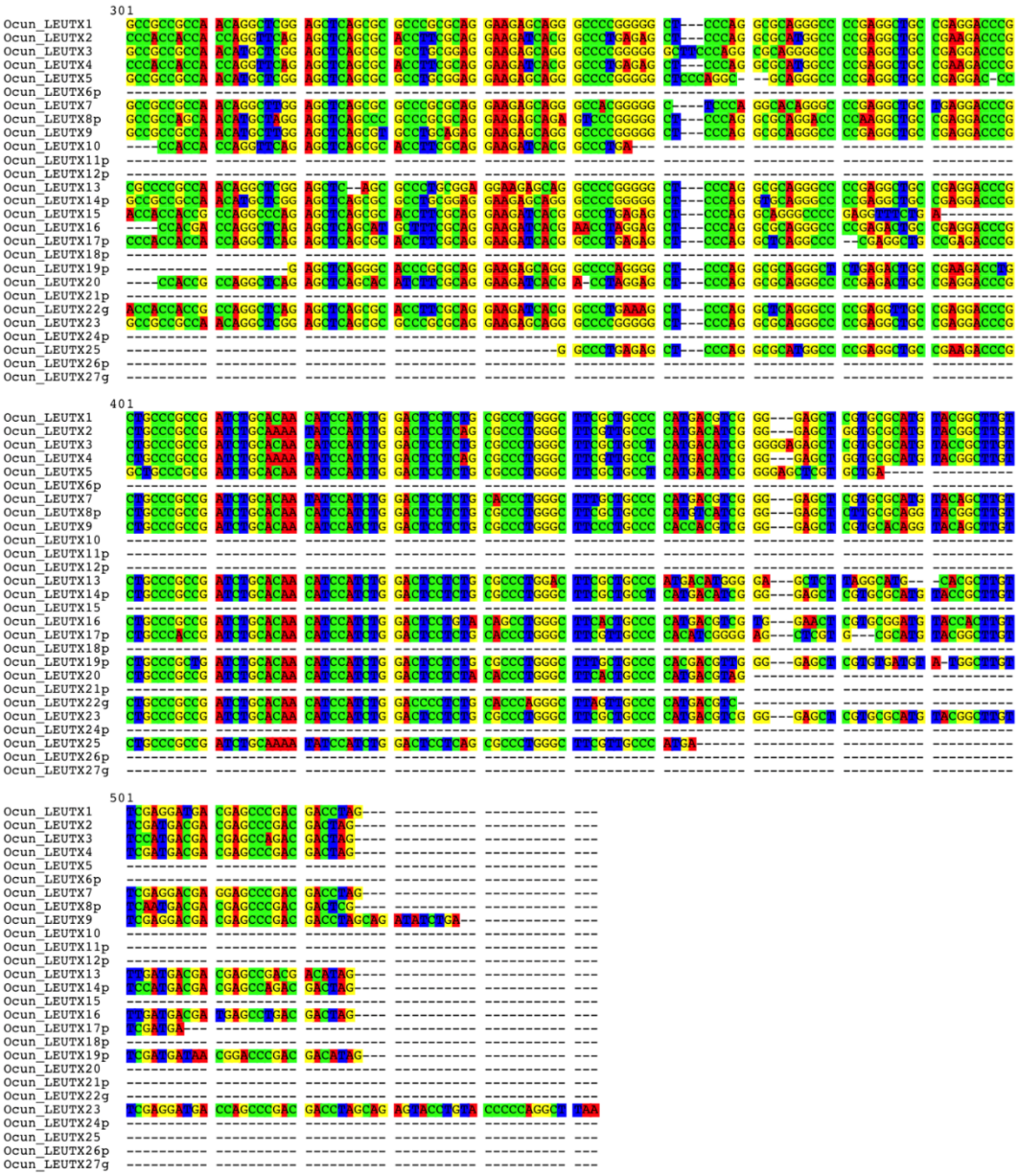


Figure S2.1 Multiple sequence alignment of *Oryctolagus cuniculus LEUTX* genes. Homeobox-containing exons (coding exons 2 and 3) shown; coding sequence shown only. The following genes are designated pseudogenes (no complete homeobox): *LEUTX6p* (frameshift), *LEUTX8p* (missing part of exon 2), *LEUTX11p* (frameshift; splice site lost), *LEUTX12p* (stop codon in homeobox), *LEUTX14p* (splice site lost), *LEUTX17p* (frameshift), *LEUTX18p* (frameshift), *LEUTX19p* (missing part of exon 3), *LEUTX21p* (frameshift), *LEUTX24p* (frameshift), *LEUTX26p* (frameshift). The status of *LEUTX22g* and *LEUTX27g* is unknown due to assembly gaps. p = pseudogene, g = assembly gaps at locus.

Figure S2.2

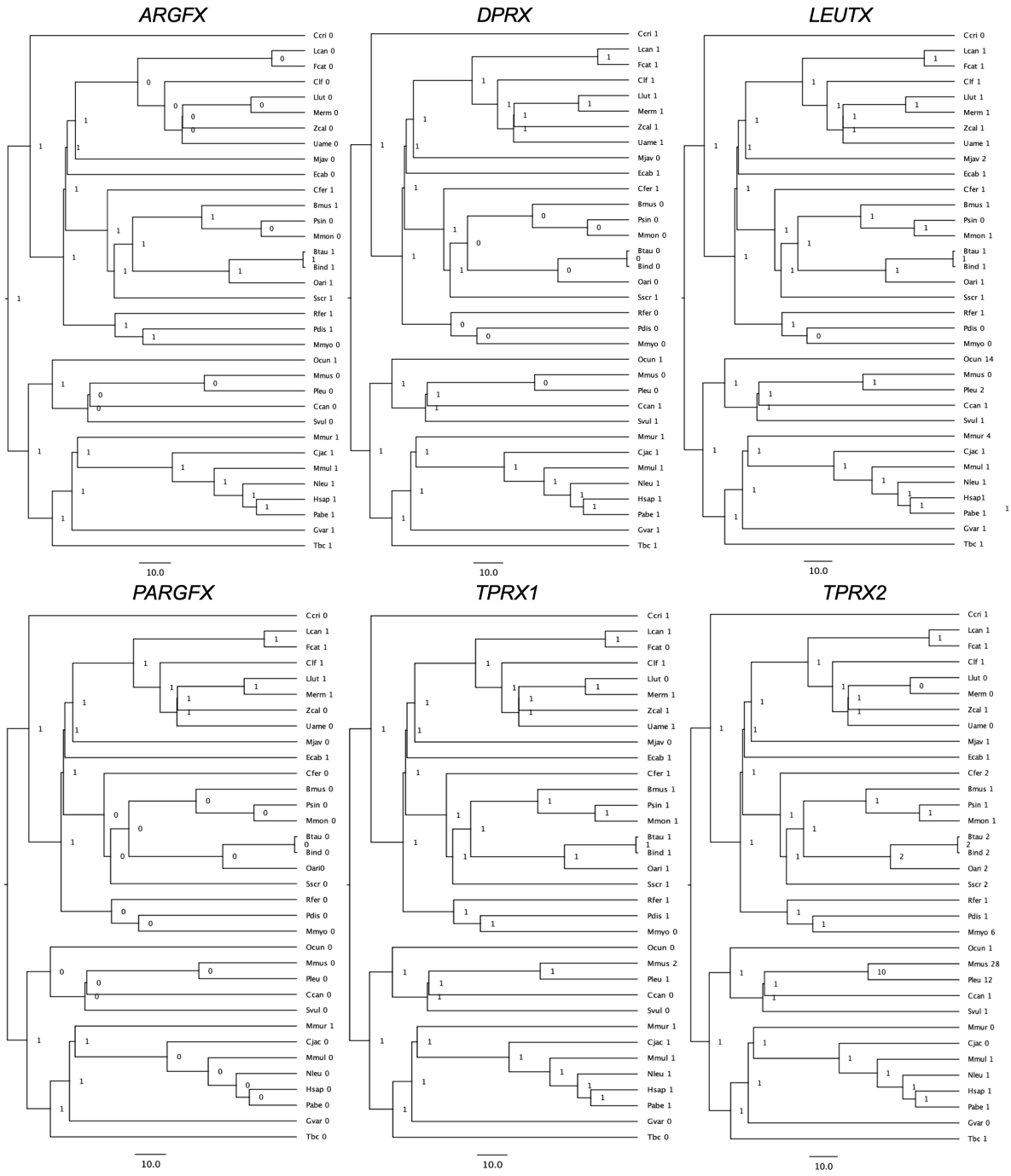


Figure S2.2 Inference of ETCHbox copy number. Phylogenies of 34 eutherian species with values indicating the number of gene copies present. Gene copy numbers at internal nodes were estimated using CAFE[457,458]. Putative pseudogenes are excluded. The phylogeny for this analysis was calculated using TimeTree[461]. Species abbreviations as in Figure 2 of main text.

Figure S2.3

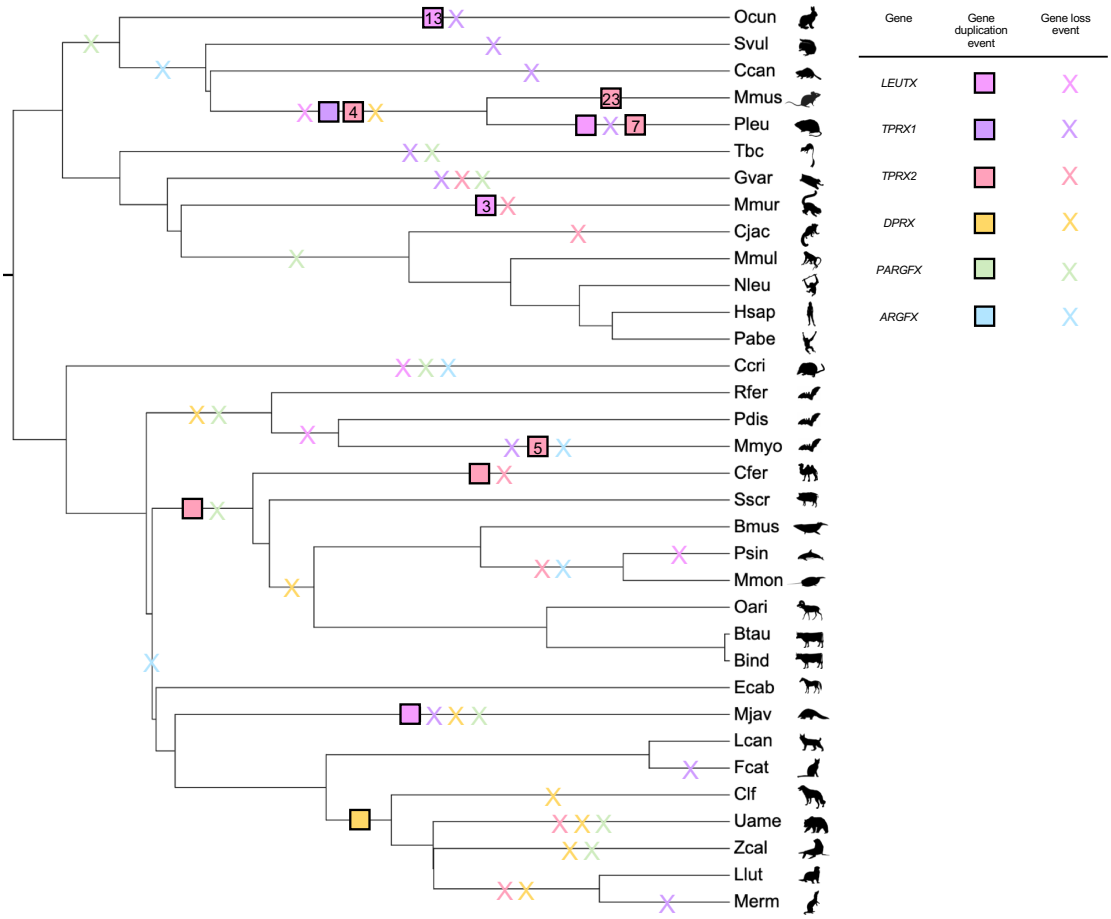


Figure S2.3 Duplication and loss of ETCHbox genes. Phylogeny of 34 eutherian species including gene duplication and gene loss events. Lineages where a gene duplication has occurred are marked with a coloured square, lineages where a gene loss event has occurred are marked with a coloured X. Pink = *LEUTX*, purple = *TPRX1*, red = *TPRX2*, yellow = *DPRX*, green = *PARGFX*, blue = *ARGFX*. Gene duplication and loss events were estimated using Notung[459,460]. Putative pseudogenes are excluded. The phylogeny for this analysis was calculated using TimeTree[461]. Species abbreviations as in Figure 2 of main text.

Figure S2.4

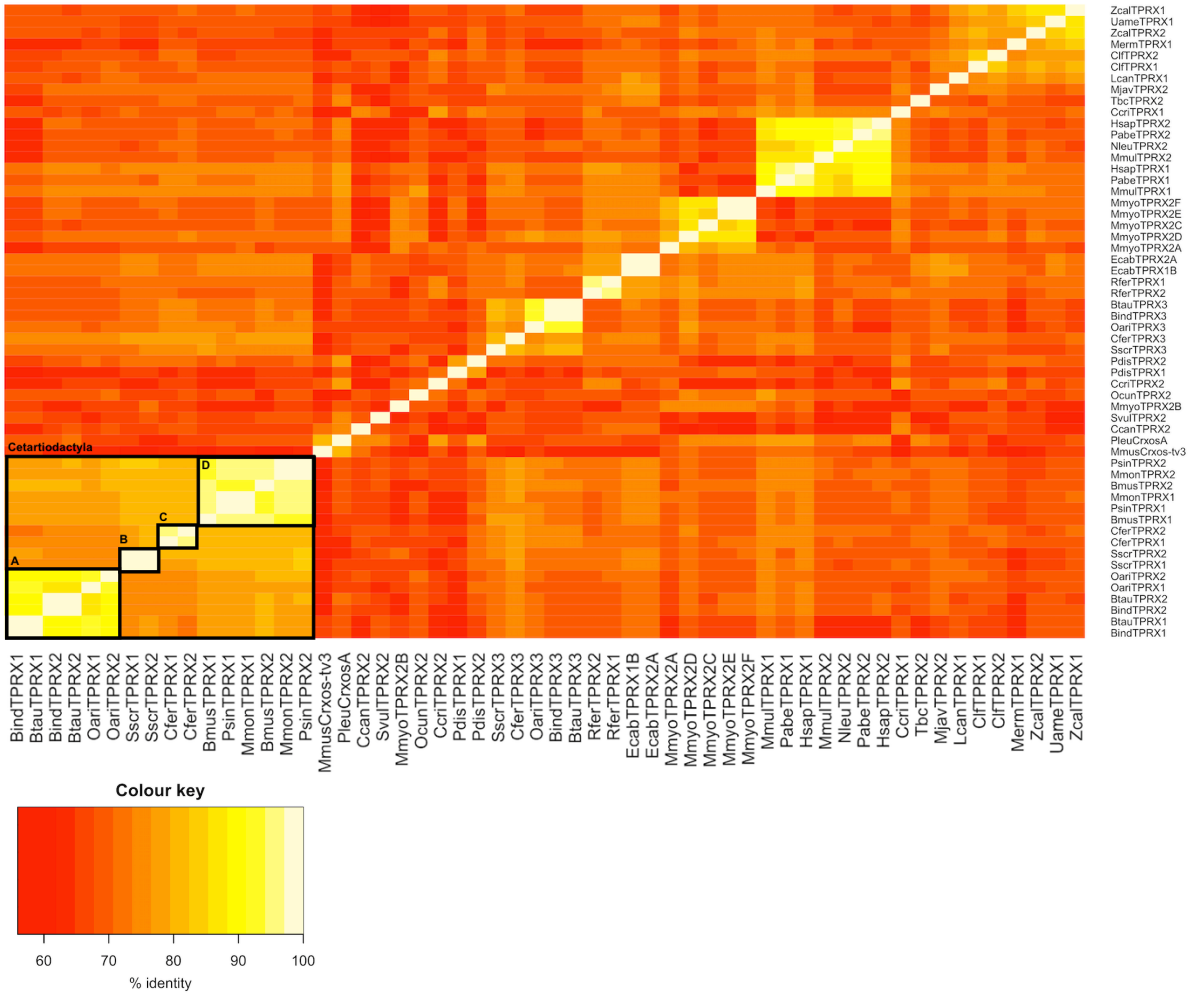
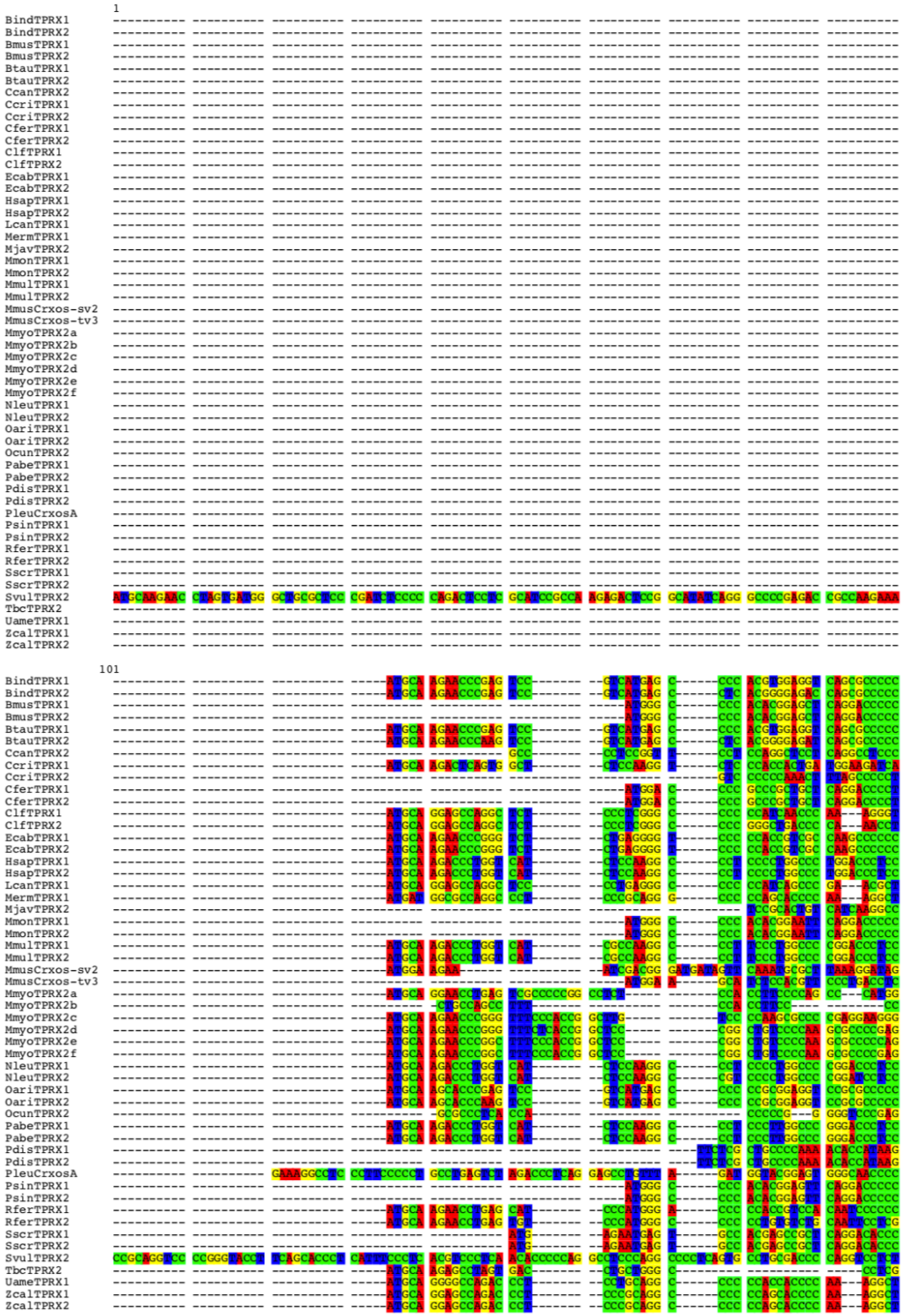
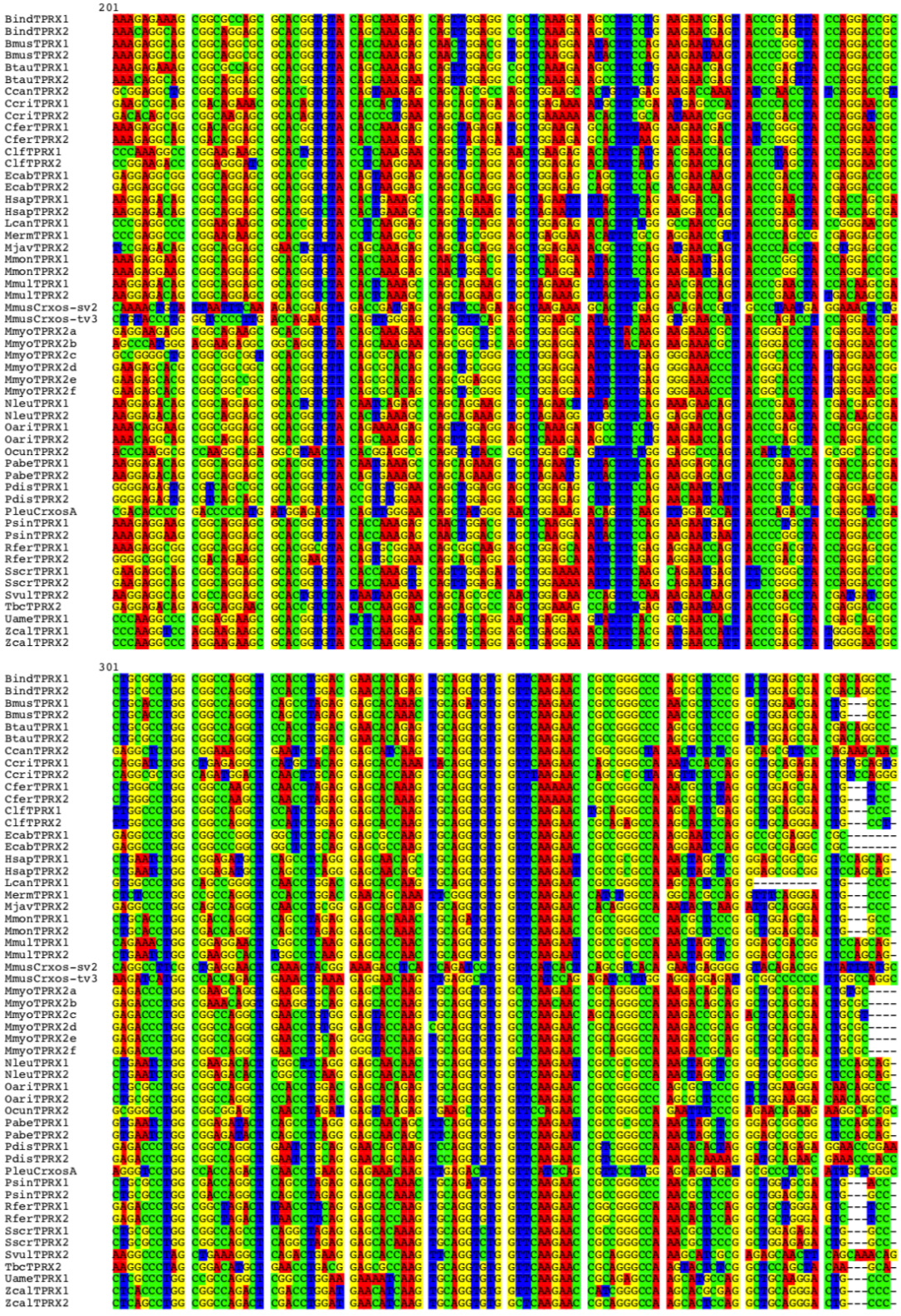
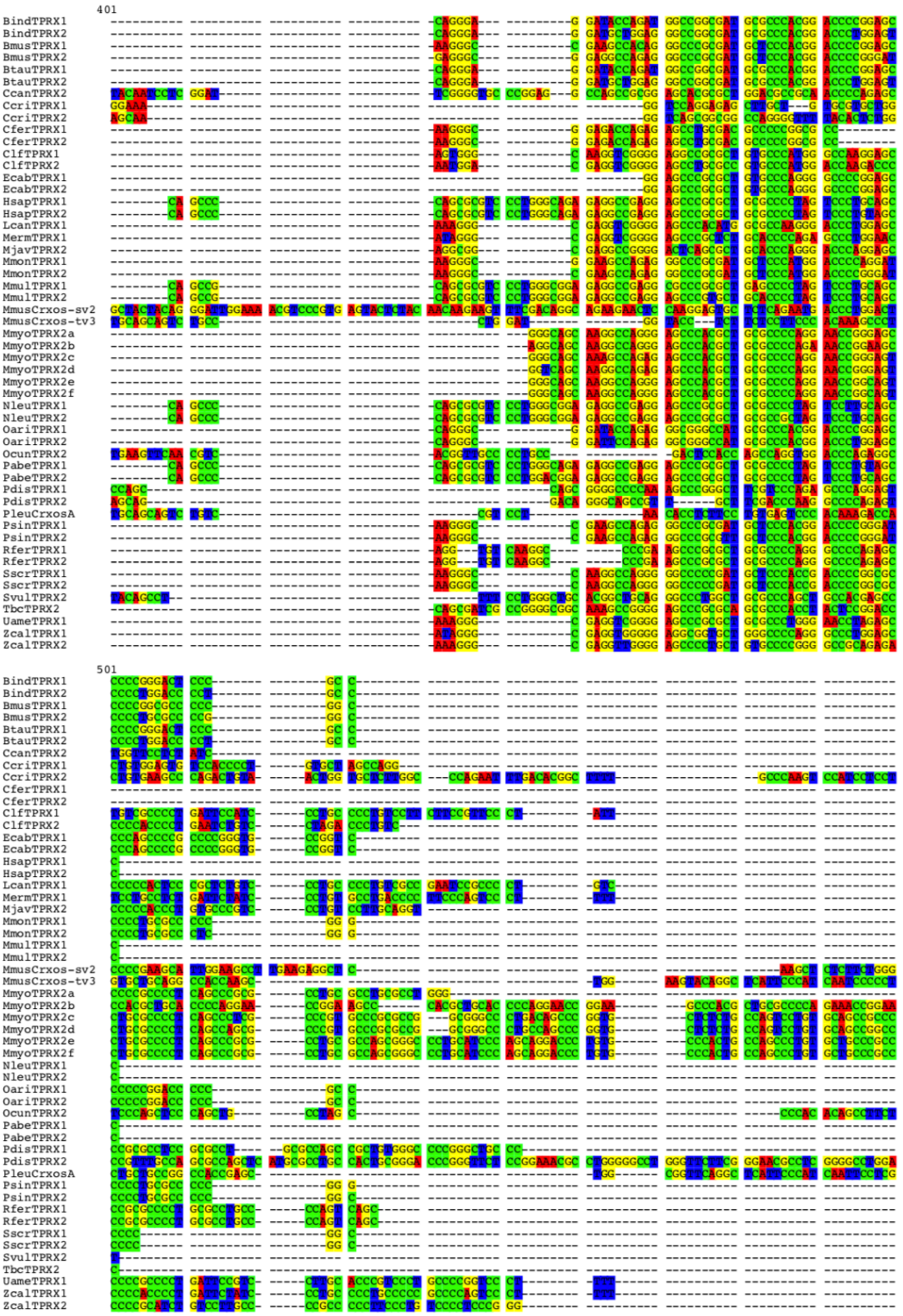


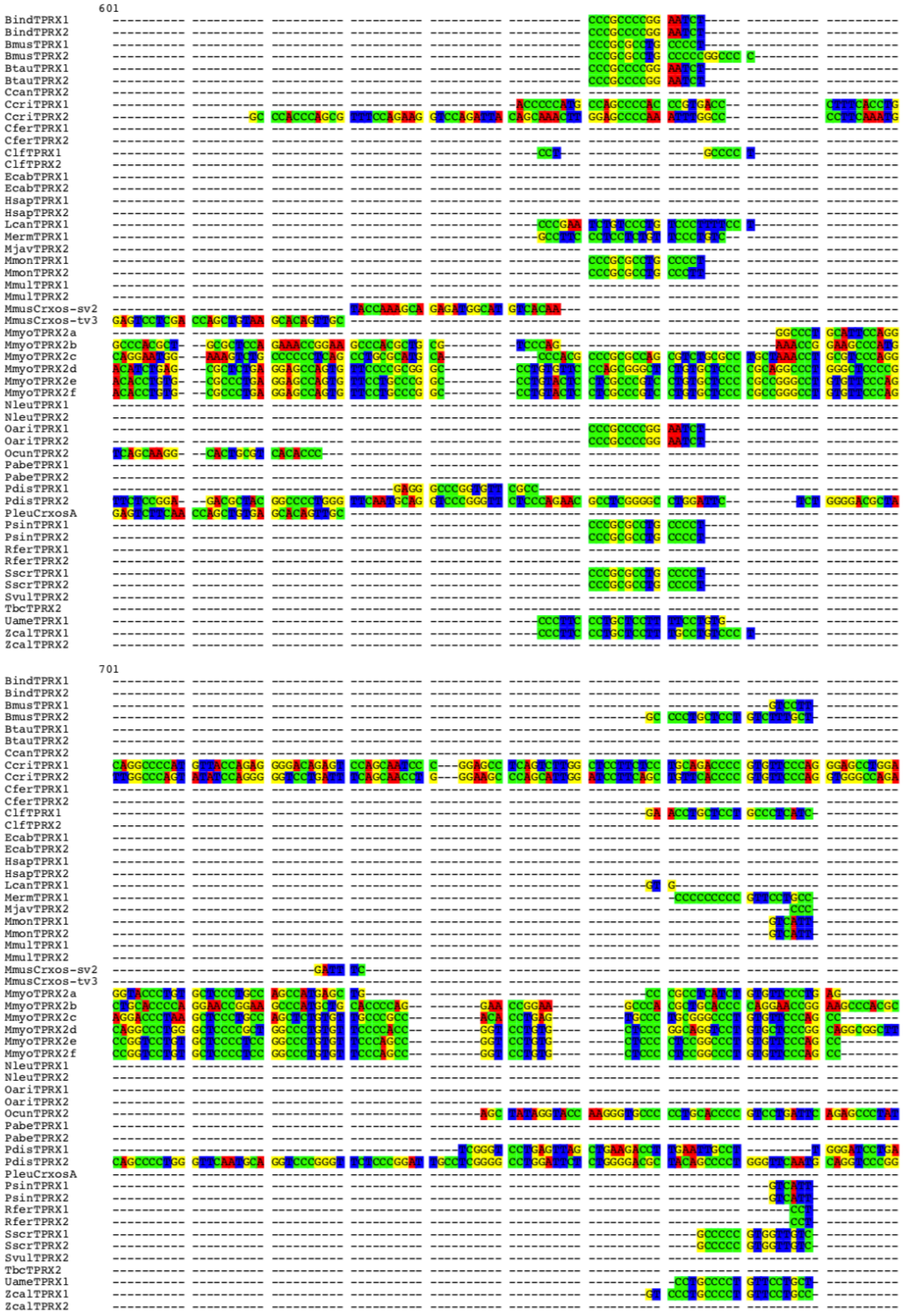
Figure S2.4 Gene conversion of TPRX genes. Heatmap shows percent nucleotide sequence identity between pairwise alignments of *TPRX1*, *TPRX2* and *TPRX3* (even-toed ungulate-specific) coding regions, with sequences clustered by similarity. Lighter colours indicate higher levels of identity; red indicates lowest identity. High similarity between *TPRX1* and *TPRX2* genes of one species is indicative of recent gene conversion events (e.g *Rhinolophus ferrumequinum*, Rfer and *Equus caballus*, Ecab). Box 'Cetartiodactyla' highlights the Cetartiodactyla genes. All Cetartiodactyla *TPRX1* and *TPRX2* genes are more similar to each other than to any other species' genes (indicated by brighter colours in the heatmap), suggesting a gene conversion event at the base of the Cetartiodactyla. Inside this Cetartiodactyla box, there are further bright boxes representing the *TPRX1* and *TPRX2* genes of the Bovidae (box A), *Sus scrofa* (Sscr; B), *Camelus ferus* (Cfer; C) and Cetacea (D), suggesting that further gene conversion events have occurred in these lineages within the Cetartiodactyla. Species abbreviations as in Figure 2 of main text.

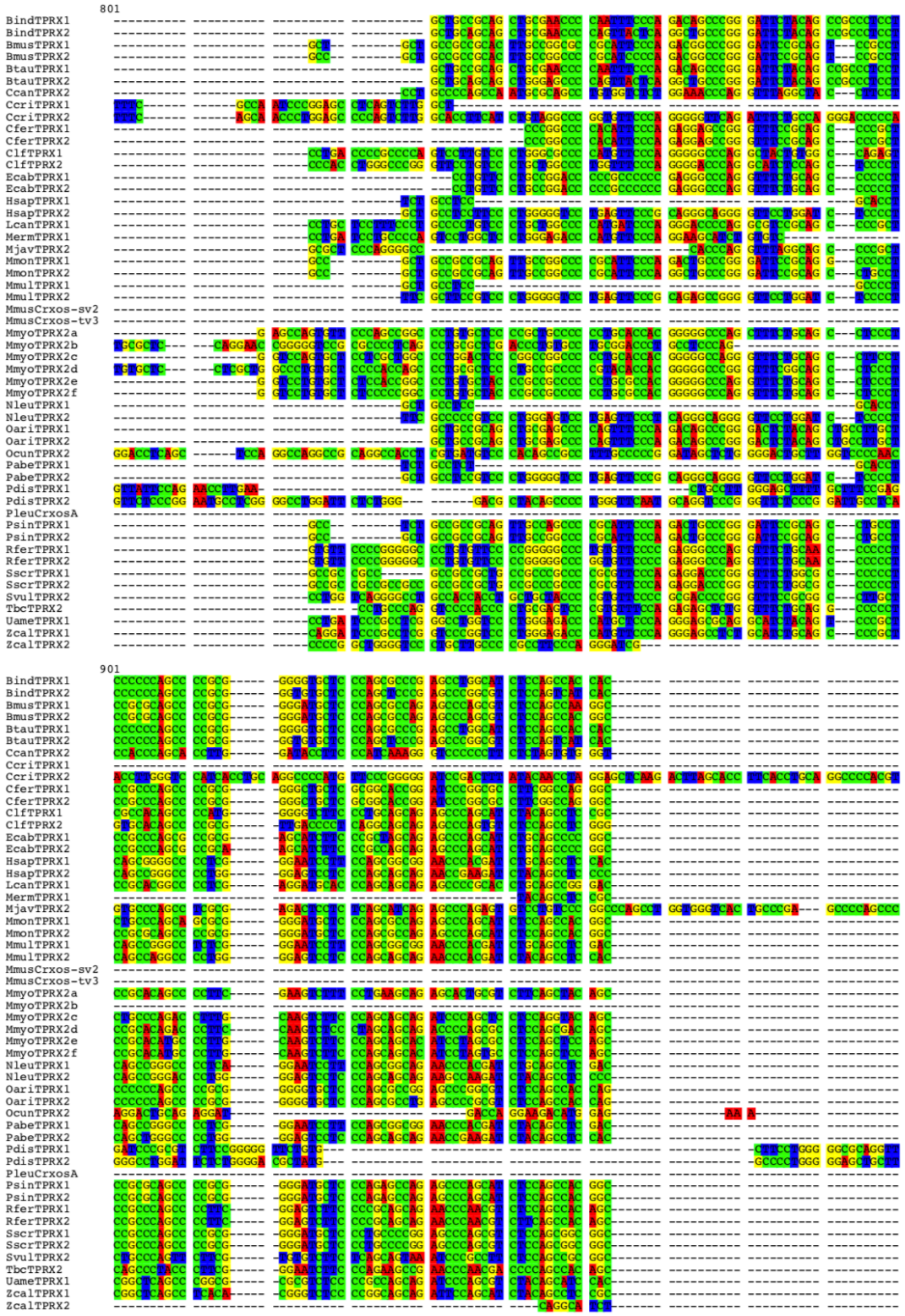
Figure S2.5

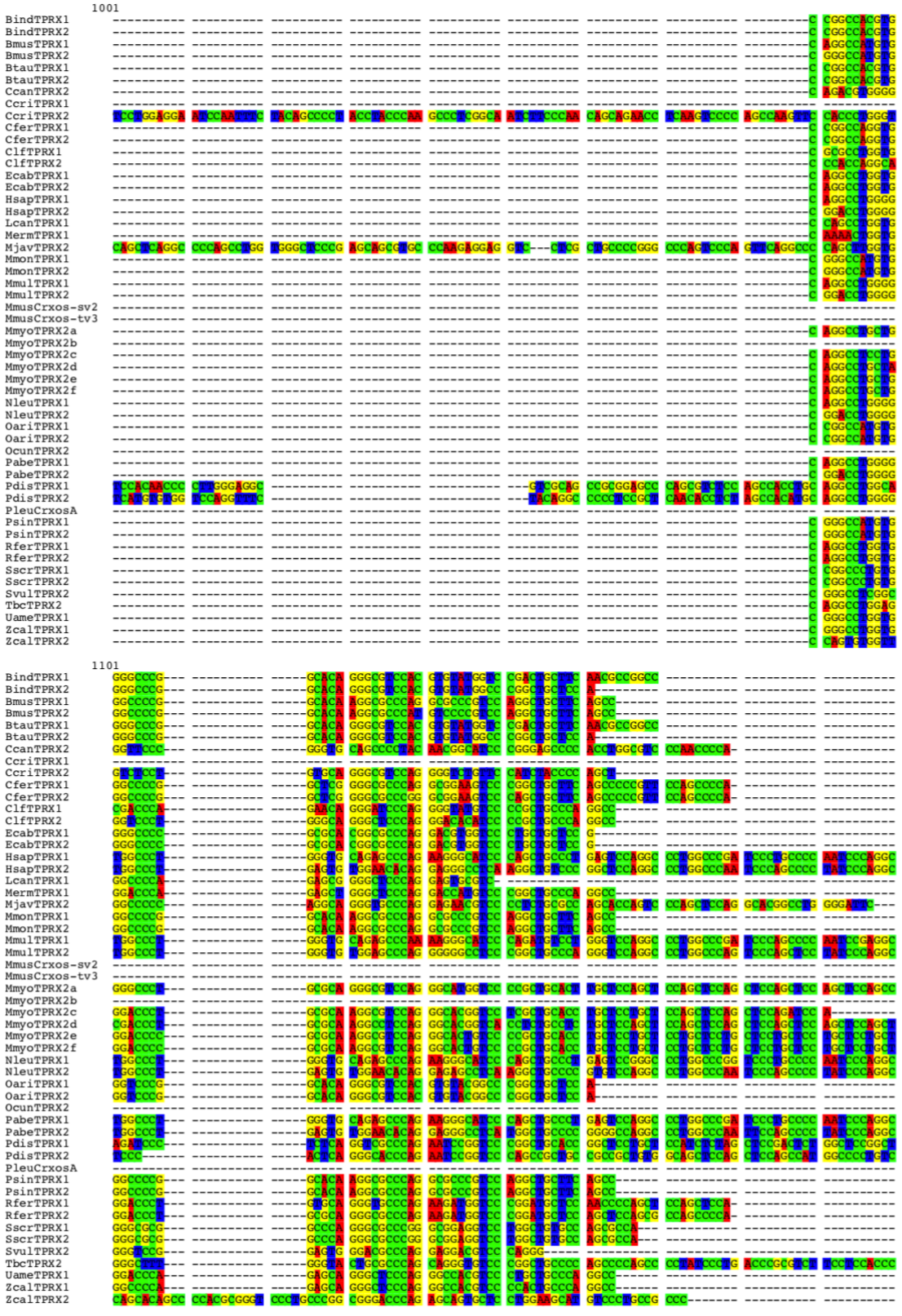

1
101
BindTPRX1
BindTPRX2
BmusTPRX1
BmusTPRX2
BtauTPRX1
BtauTPRX2
CcanTPRX2
CcriTPRX1
CcriTPRX2
CferTPRX1
CferTPRX2
ClfTPRX1
ClfTPRX2
EcabTPRX1
EcabTPRX2
HsapTPRX1
HsapTPRX2
LcanTPRX1
MermTPRX1
MjavTPRX2
MmonTPRX1
MmonTPRX2
MmulTPRX1
MmulTPRX2
MmusCrxos-sv2
MmusCrxos-tv3
MmyoTPRX2a
MmyoTPRX2b
MmyoTPRX2c
MmyoTPRX2d
MmyoTPRX2e
MmyoTPRX2f
NleuTPRX1
NleuTPRX2
OariTPRX1
OariTPRX2
OcunTPRX2
PabeTPRX1
PabeTPRX2
PdisTPRX1
PdisTPRX2
PleuCrxosA
PsinTPRX1
PsinTPRX2
RferTPRX1
RferTPRX2
SscrTPRX1
SscrTPRX2
SvulTPRX2
TbcTPRX2
UameTPRX1
ZcalTPRX1
ZcalTPRX2

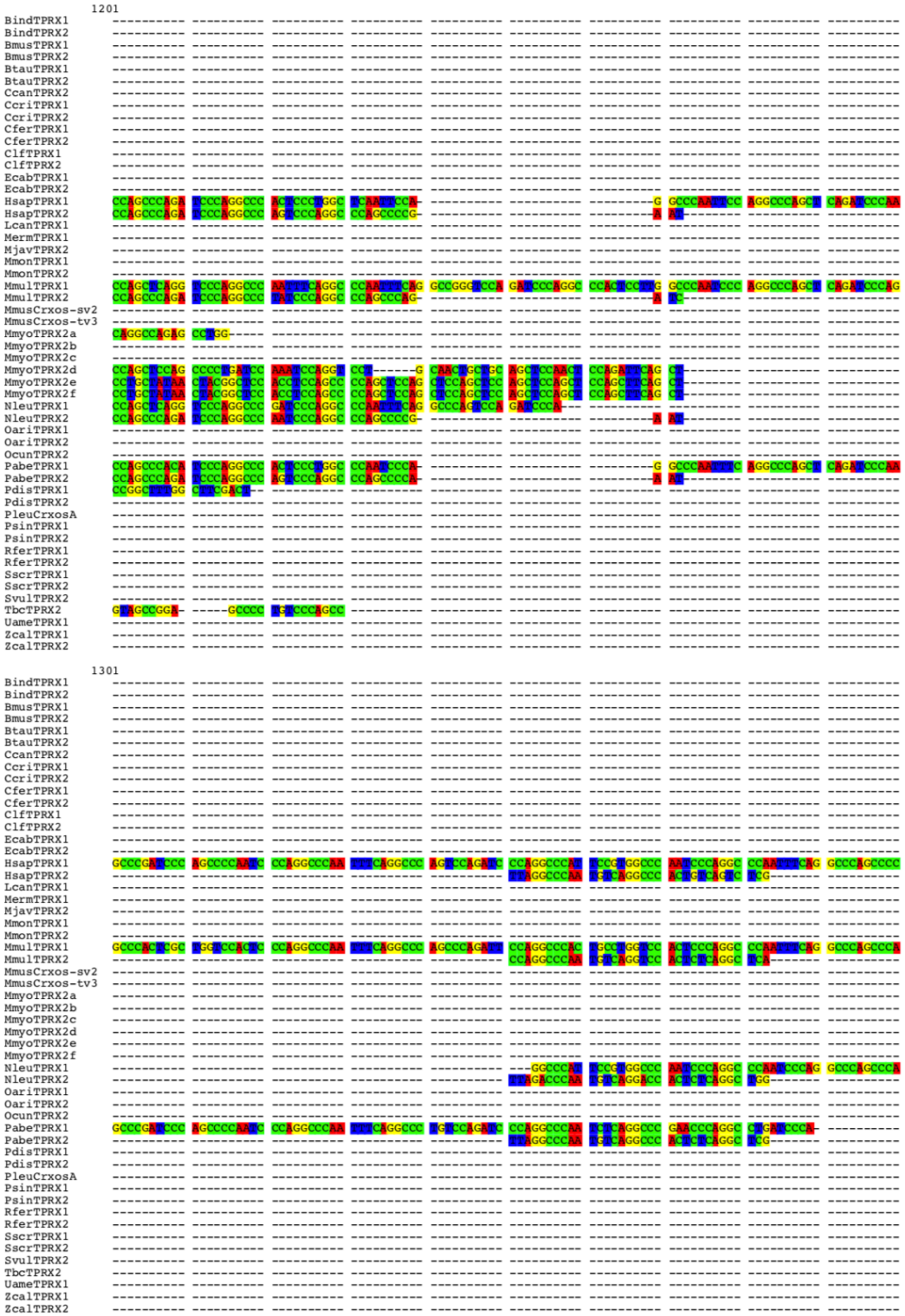

201
BindTPRX1
BindTPRX2
BmusTPRX1
BmusTPRX2
BtauTPRX1
BtauTPRX2
CcanTPRX2
CcriTPRX1
CcriTPRX2
CferTPRX1
CferTPRX2
ClfTPRX1
ClfTPRX2
EcabTPRX1
EcabTPRX2
HsapTPRX1
HsapTPRX2
LcanTPRX1
MermTPRX1
MjavTPRX2
MmonTPRX1
MmonTPRX2
MmulTPRX1
MmulTPRX2
MmusCrxos-sv2
MmusCrxos-tv3
MmyoTPRX2a
MmyoTPRX2b
MmyoTPRX2c
MmyoTPRX2d
MmyoTPRX2e
MmyoTPRX2f
NleuTPRX1
NleuTPRX2
OariTPRX1
OariTPRX2
OcunTPRX2
PabeTPRX1
PabeTPRX2
PdisTPRX1
PdisTPRX2
PleuCrxosA
PsinTPRX1
PsinTPRX2
RferTPRX1
RferTPRX2
SscrTPRX1
SscrTPRX2
SvulTPRX2
TbcTPRX2
UameTPRX1
ZcalTPRX1
ZcalTPRX2
301
BindTPRX1
BindTPRX2
BmusTPRX1
BmusTPRX2
BtauTPRX1
BtauTPRX2
CcanTPRX2
CcriTPRX1
CcriTPRX2
CferTPRX1
CferTPRX2
ClfTPRX1
ClfTPRX2
EcabTPRX1
EcabTPRX2
HsapTPRX1
HsapTPRX2
LcanTPRX1
MermTPRX1
MjavTPRX2
MmonTPRX1
MmonTPRX2
MmulTPRX1
MmulTPRX2
MmusCrxos-sv2
MmusCrxos-tv3
MmyoTPRX2a
MmyoTPRX2b
MmyoTPRX2c
MmyoTPRX2d
MmyoTPRX2e
MmyoTPRX2f
NleuTPRX1
NleuTPRX2
OariTPRX1
OariTPRX2
OcunTPRX2
PabeTPRX1
PabeTPRX2
PdisTPRX1
PdisTPRX2
PleuCrxosA
PsinTPRX1
PsinTPRX2
RferTPRX1
RferTPRX2
SscrTPRX1
SscrTPRX2
SvulTPRX2
TbcTPRX2
UameTPRX1
ZcalTPRX1
ZcalTPRX2

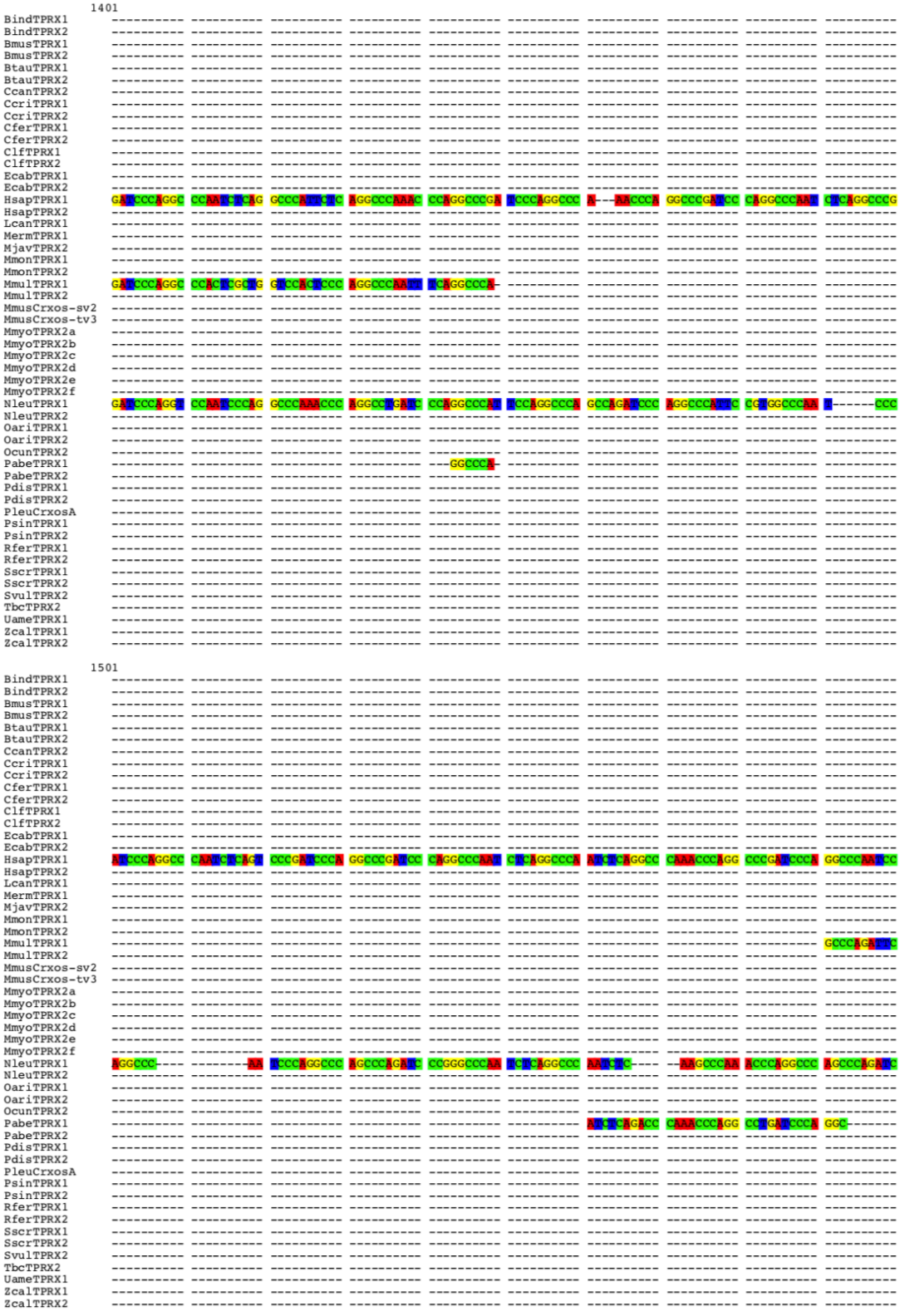


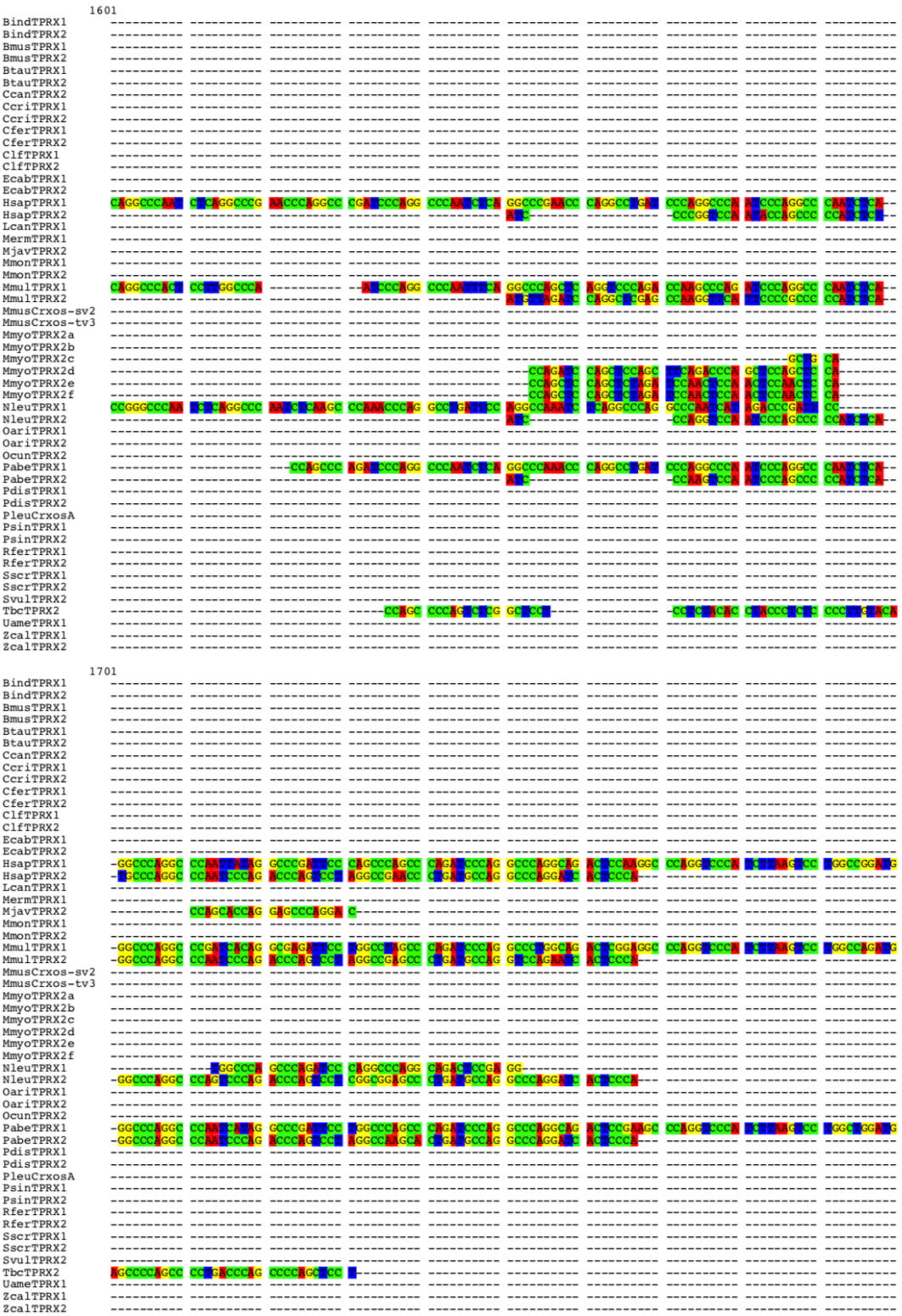

601
701
BindTPRX1
BindTPRX2
BmusTPRX1
BmusTPRX2
BtauTPRX1
BtauTPRX2
CcanTPRX2
CcriTPRX1
CcriTPRX2
CferTPRX1
CferTPRX2
ClfTPRX1
ClfTPRX2
EcabTPRX1
EcabTPRX2
HsapTPRX1
HsapTPRX2
LcanTPRX1
MermTPRX1
MjavTPRX2
MmonTPRX1
MmonTPRX2
MmulTPRX1
MmulTPRX2
MmusCrxos-sv2
MmusCrxos-tv3
MmyoTPRX2a
MmyoTPRX2b
MmyoTPRX2c
MmyoTPRX2d
MmyoTPRX2e
MmyoTPRX2f
NleuTPRX1
NleuTPRX2
OariTPRX1
OariTPRX2
OcunTPRX2
PabeTPRX1
PabeTPRX2
PdisTPRX1
PdisTPRX2
PleuCrxosA
PsinTPRX1
PsinTPRX2
RferTPRX1
RferTPRX2
SscrTPRX1
SscrTPRX2
SvulTPRX2
TbcTPRX2
UameTPRX1
ZcalTPRX1
ZcalTPRX2

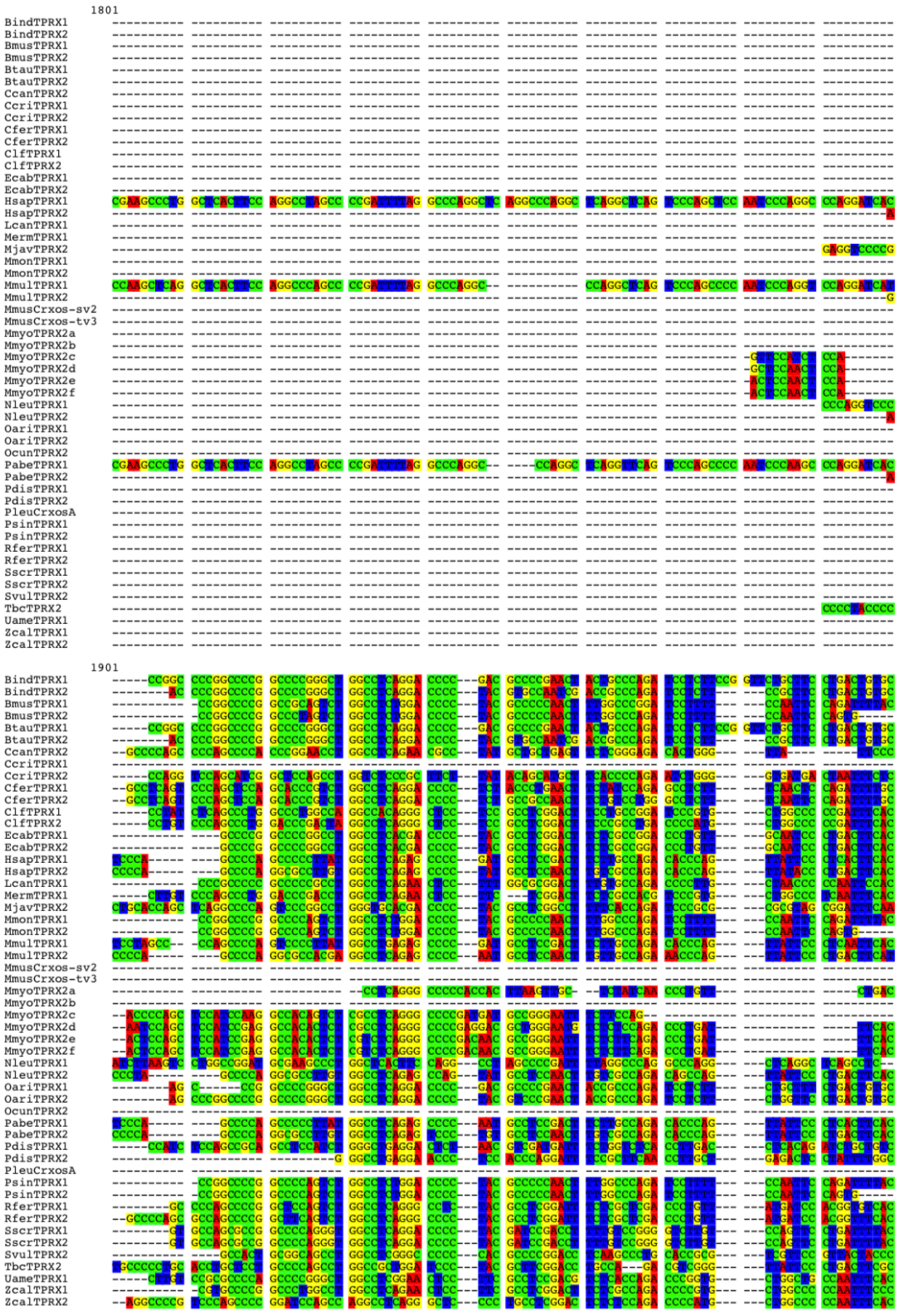

801
BindTPRX1
BindTPRX2
BmusTPRX1
BmusTPRX2
BtauTPRX1
BtauTPRX2
CcanTPRX2
CcriTPRX1
CcriTPRX2
CferTPRX1
CferTPRX2
ClfTPRX1
ClfTPRX2
EcabTPRX1
EcabTPRX2
HsapTPRX1
HsapTPRX2
LcanTPRX1
MermTPRX1
MjavTPRX2
MmonTPRX1
MmonTPRX2
MmulTPRX1
MmulTPRX2
MmusCrxos-sv2
MmusCrxos-tv3
MmyoTPRX2a
MmyoTPRX2b
MmyoTPRX2c
MmyoTPRX2d
MmyoTPRX2e
MmyoTPRX2f
NleuTPRX1
NleuTPRX2
OariTPRX1
OariTPRX2
OcunTPRX2
PabeTPRX1
PabeTPRX2
PdisTPRX1
PdisTPRX2
PleuCrxosA
PsinTPRX1
PsinTPRX2
RferTPRX1
RferTPRX2
SscrTPRX1
SscrTPRX2
SvulTPRX2
TbcTPRX2
UameTPRX1
ZcalTPRX1
ZcalTPRX2
901
BindTPRX1
BindTPRX2
BmusTPRX1
BmusTPRX2
BtauTPRX1
BtauTPRX2
CcanTPRX2
CcriTPRX1
CcriTPRX2
CferTPRX1
CferTPRX2
ClfTPRX1
ClfTPRX2
EcabTPRX1
EcabTPRX2
HsapTPRX1
HsapTPRX2
LcanTPRX1
MermTPRX1
MjavTPRX2
MmonTPRX1
MmonTPRX2
MmulTPRX1
MmulTPRX2
MmusCrxos-sv2
MmusCrxos-tv3
MmyoTPRX2a
MmyoTPRX2b
MmyoTPRX2c
MmyoTPRX2d
MmyoTPRX2e
MmyoTPRX2f
NleuTPRX1
NleuTPRX2
OariTPRX1
OariTPRX2
OcunTPRX2
PabeTPRX1
PabeTPRX2
PdisTPRX1
PdisTPRX2
PleuCrxosA
PsinTPRX1
PsinTPRX2
RferTPRX1
RferTPRX2
SscrTPRX1
SscrTPRX2
SvulTPRX2
TbcTPRX2
UameTPRX1
ZcalTPRX1
ZcalTPRX2

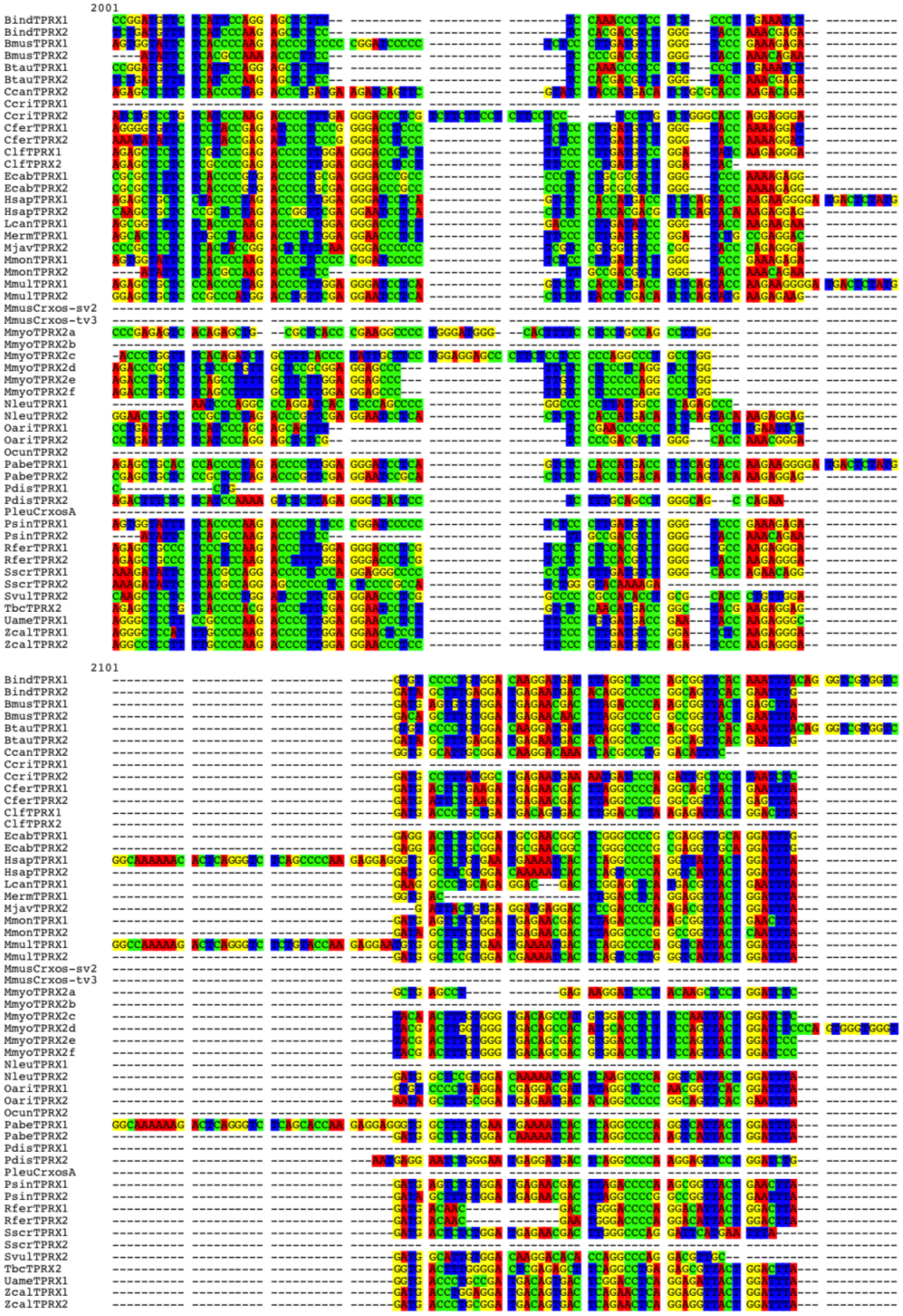





1401
BindTPRX1
BindTPRX2
BmusTPRX1
BmusTPRX2
BtauTPRX1
BtauTPRX2
CcanTPRX2
CcriTPRX1
CcriTPRX2
CferTPRX1
CferTPRX2
ClfTPRX1
ClfTPRX2
EcabTPRX1
EcabTPRX2
HsapTPRX1
HsapTPRX2
LcanTPRX1
MermTPRX1
MjavTPRX2
MmonTPRX1
MmonTPRX2
MmulTPRX1
MmulTPRX2
MmusCrxos-sv2
MmusCrxos-tv3
MmyoTPRX2a
MmyoTPRX2b
MmyoTPRX2c
MmyoTPRX2d
MmyoTPRX2e
MmyoTPRX2f
NleuTPRX1
NleuTPRX2
OariTPRX1
OariTPRX2
OcunTPRX2
PabeTPRX1
PabeTPRX2
PdisTPRX1
PdisTPRX2
PleuCrxosA
PsinTPRX1
PsinTPRX2
RferTPRX1
RferTPRX2
SscrTPRX1
SscrTPRX2
SvulTPRX2
TbcTPRX2
UameTPRX1
ZcalTPRX1
ZcalTPRX2
1501
BindTPRX1
BindTPRX2
BmusTPRX1
BmusTPRX2
BtauTPRX1
BtauTPRX2
CcanTPRX2
CcriTPRX1
CcriTPRX2
CferTPRX1
CferTPRX2
ClfTPRX1
ClfTPRX2
EcabTPRX1
EcabTPRX2
HsapTPRX1
HsapTPRX2
LcanTPRX1
MermTPRX1
MjavTPRX2
MmonTPRX1
MmonTPRX2
MmulTPRX1
MmulTPRX2
MmusCrxos-sv2
MmusCrxos-tv3
MmyoTPRX2a
MmyoTPRX2b
MmyoTPRX2c
MmyoTPRX2d
MmyoTPRX2e
MmyoTPRX2f
NleuTPRX1
NleuTPRX2
OariTPRX1
OariTPRX2
OcunTPRX2
PabeTPRX1
PabeTPRX2
PdisTPRX1
PdisTPRX2
PleuCrxosA
PsinTPRX1
PsinTPRX2
RferTPRX1
RferTPRX2
SscrTPRX1
SscrTPRX2
SvulTPRX2
TbcTPRX2
UameTPRX1
ZcalTPRX1
ZcalTPRX2




1801
1901
BindTPRX1
BindTPRX2
BmusTPRX1
BmusTPRX2
BtauTPRX1
BtauTPRX2
CcanTPRX2
CcriTPRX1
CcriTPRX2
CferTPRX1
CferTPRX2
ClfTPRX1
ClfTPRX2
EcabTPRX1
EcabTPRX2
HsapTPRX1
HsapTPRX2
LcanTPRX1
MermTPRX1
MjavTPRX2
MmonTPRX1
MmonTPRX2
MmulTPRX1
MmulTPRX2
MmusCrxos-sv2
MmusCrxos-tv3
MmyoTPRX2a
MmyoTPRX2b
MmyoTPRX2c
MmyoTPRX2d
MmyoTPRX2e
MmyoTPRX2f
NleuTPRX1
NleuTPRX2
OariTPRX1
OariTPRX2
OcunTPRX2
PabeTPRX1
PabeTPRX2
PdisTPRX1
PdisTPRX2
PleuCrxosA
PsinTPRX1
PsinTPRX2
RferTPRX1
RferTPRX2
SscrTPRX1
SscrTPRX2
SvulTPRX2
TbcTPRX2
UameTPRX1
ZcalTPRX1
ZcalTPRX2


2001
2101
BindTPRX1
BindTPRX2
BmusTPRX1
BmusTPRX2
BtauTPRX1
BtauTPRX2
CcanTPRX2
CcriTPRX1
CcriTPRX2
CferTPRX1
CferTPRX2
ClfTPRX1
ClfTPRX2
EcabTPRX1
EcabTPRX2
HsapTPRX1
HsapTPRX2
LcanTPRX1
MermTPRX1
MjavTPRX2
MmonTPRX1
MmonTPRX2
MmulTPRX1
MmulTPRX2
MmusCrxos-sv2
MmusCrxos-tv3
MmyoTPRX2a
MmyoTPRX2b
MmyoTPRX2c
MmyoTPRX2d
MmyoTPRX2e
MmyoTPRX2f
NleuTPRX1
NleuTPRX2
OariTPRX1
OariTPRX2
OcunTPRX2
PabeTPRX1
PabeTPRX2
PdisTPRX1
PdisTPRX2
PleuCrxosA
PsinTPRX1
PsinTPRX2
RferTPRX1
RferTPRX2
SscrTPRX1
SscrTPRX2
SvulTPRX2
TbcTPRX2
UameTPRX1
ZcalTPRX1
ZcalTPRX2

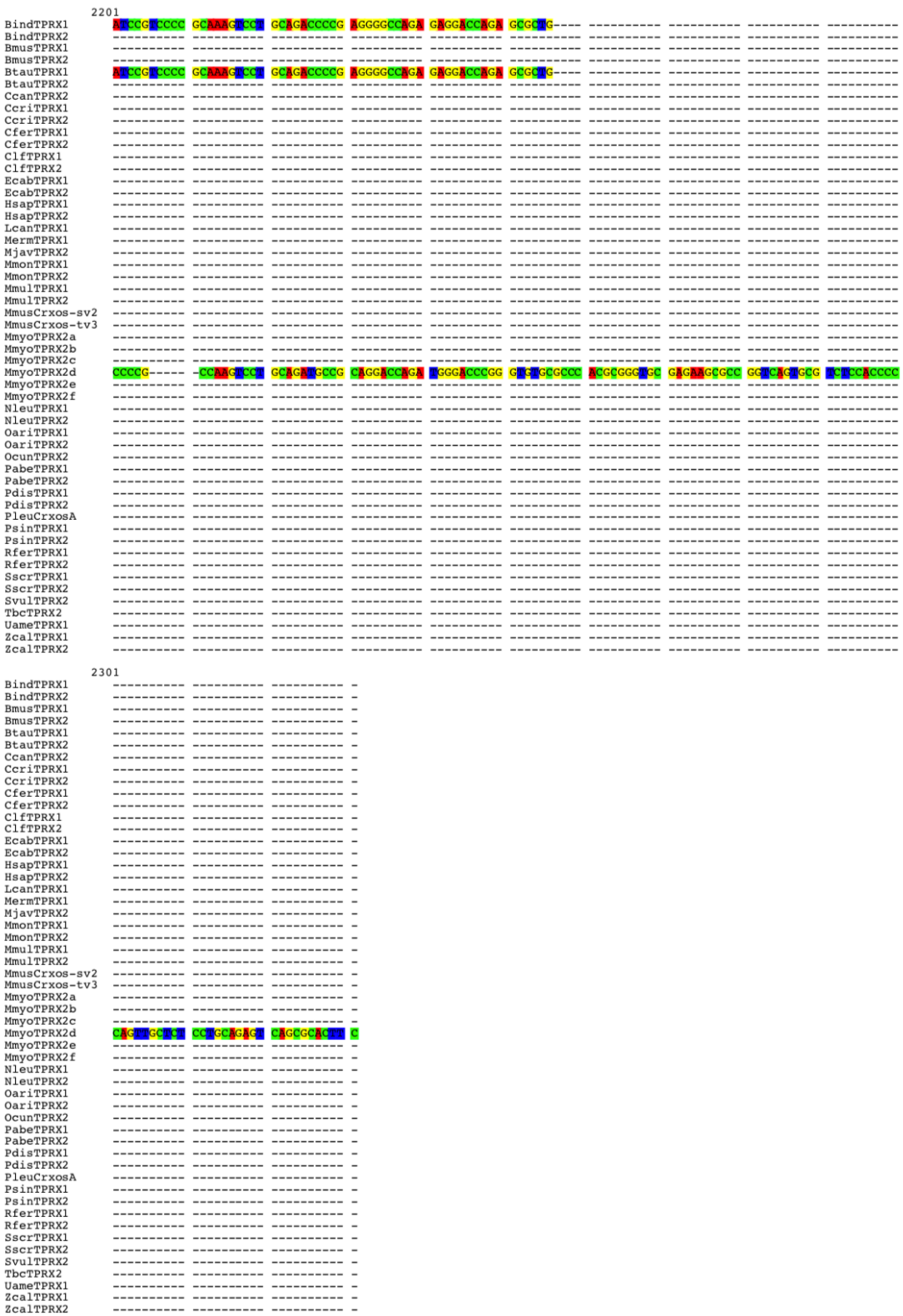


Figure S2.5 Eutherian TPRX genes. Multiple sequence alignment of coding regions of *TPRX* genes used for gene conversion analysis. Stop codons were removed. Species abbreviations as in Figure 2 of main text.

Figure S2.6

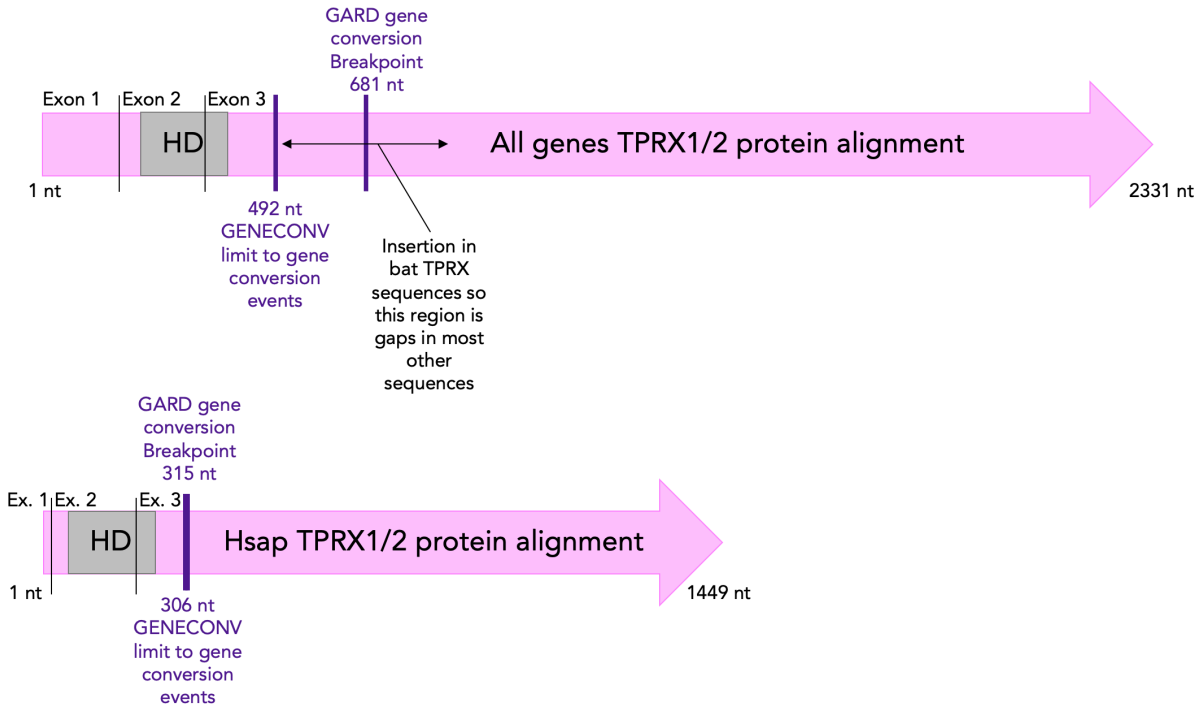


Figure S2.6 Gene conversion in TPRX genes. Location of breakpoint identified by GARD[473,474] and limit to gene conversion identified by GENECONV[486] in multiple sequence alignments of the *TPRX1* and *TPRX2* genes of all sampled species (top) and just *Homo sapiens* (bottom). GARD breakpoint and GENECONV gene conversion limit are 189 nucleotides apart in the 'all species' alignment primarily because of an insertion in bats. In most species they are much closer, and in humans the deduced breakpoints are only 9 nucleotides apart. Multiple sequence alignment of all *TPRX* genes is available as Supplementary Figure S2.5. Abbreviations: nt = nucleotide; Ex. = exon; HD = homeodomain, Hsap = *Homo sapiens*.

Figure S2.7A

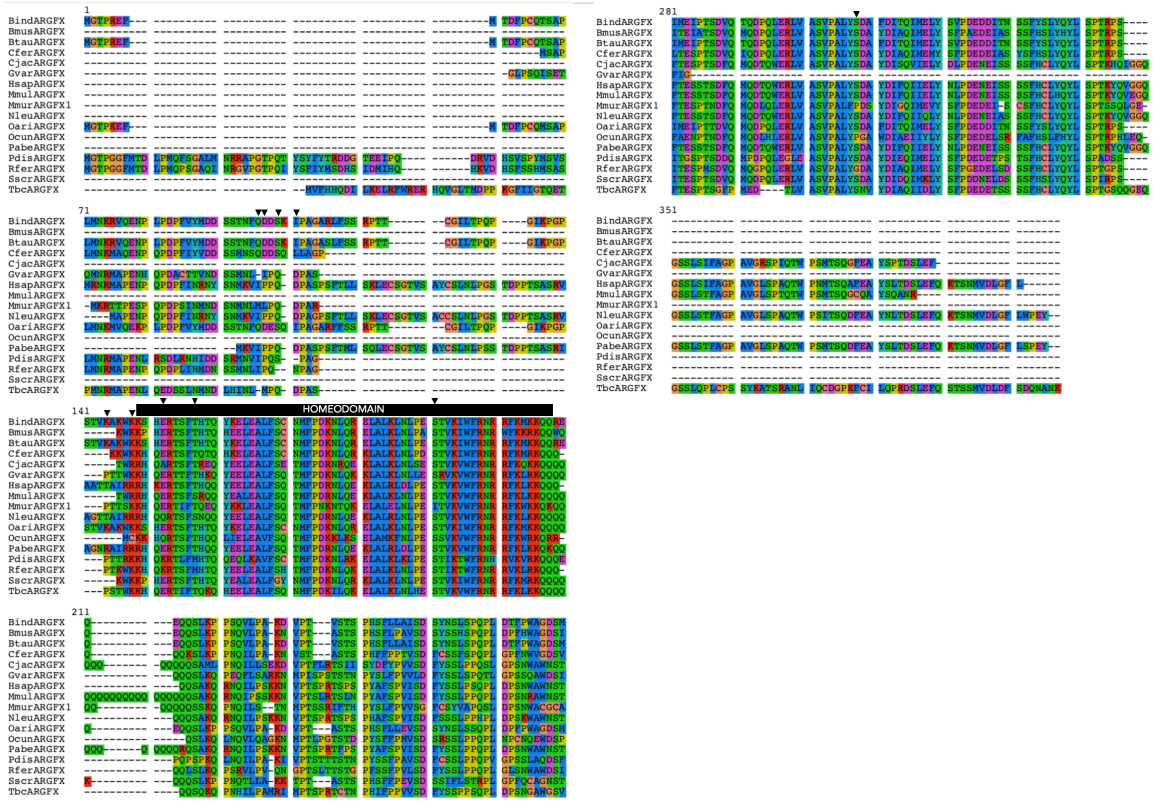


Figure S2.7A Untrimmed multiple sequence alignment of ARGFX protein sequences. Sites under positive selection, as identified by MEME[493], are marked with a black arrow. The homeodomain is marked with a black box. Species abbreviations as in Figure 2 of main text.

Figure S2.7B

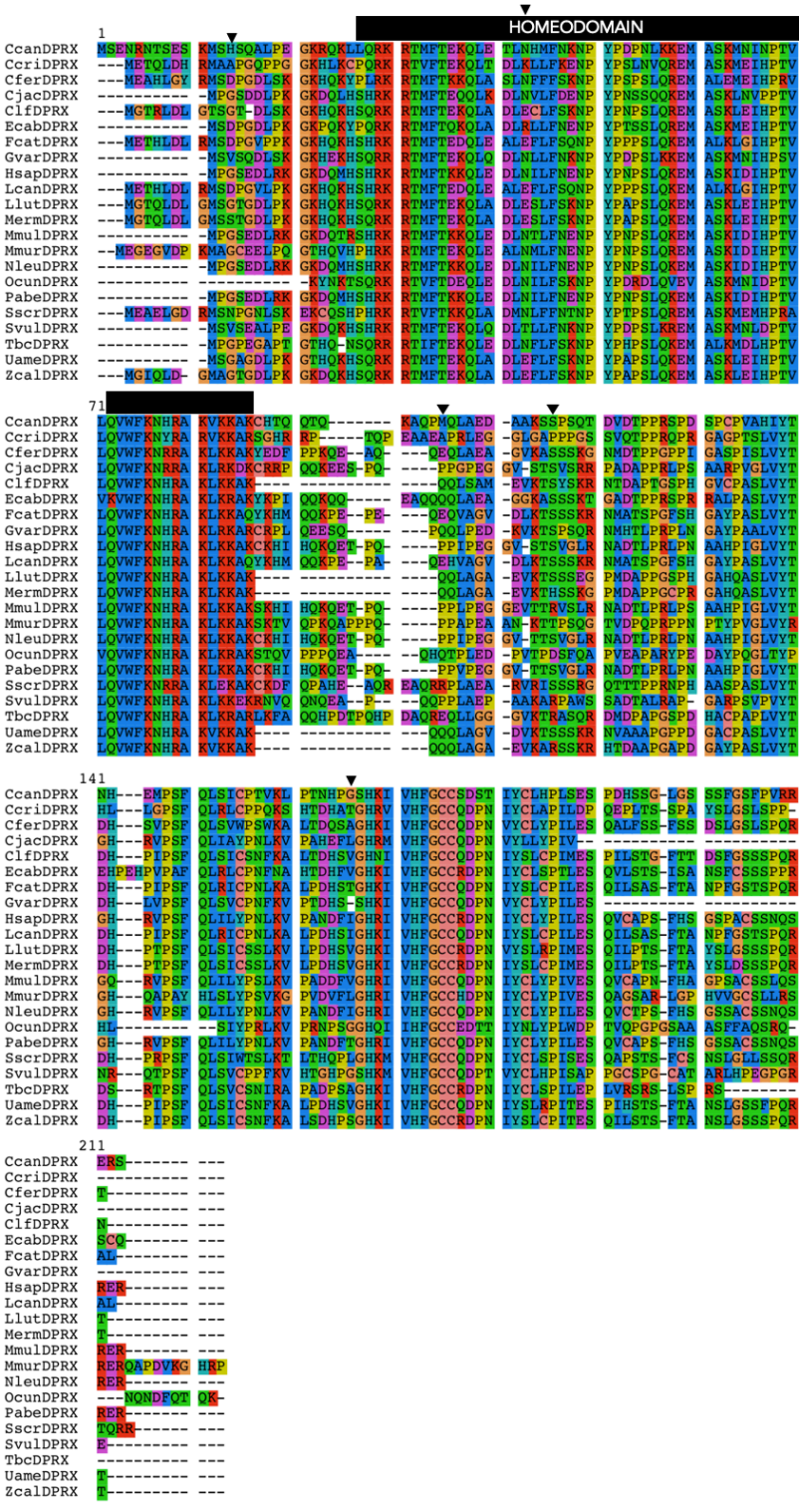


Figure S2.7B Untrimmed multiple sequence alignment of DPRX protein sequences. Sites under positive selection, as identified by MEME[493], are marked with a black arrow. The homeodomain is marked with a black box. Species abbreviations as in Figure 2 of main text.

Figure S2.7C

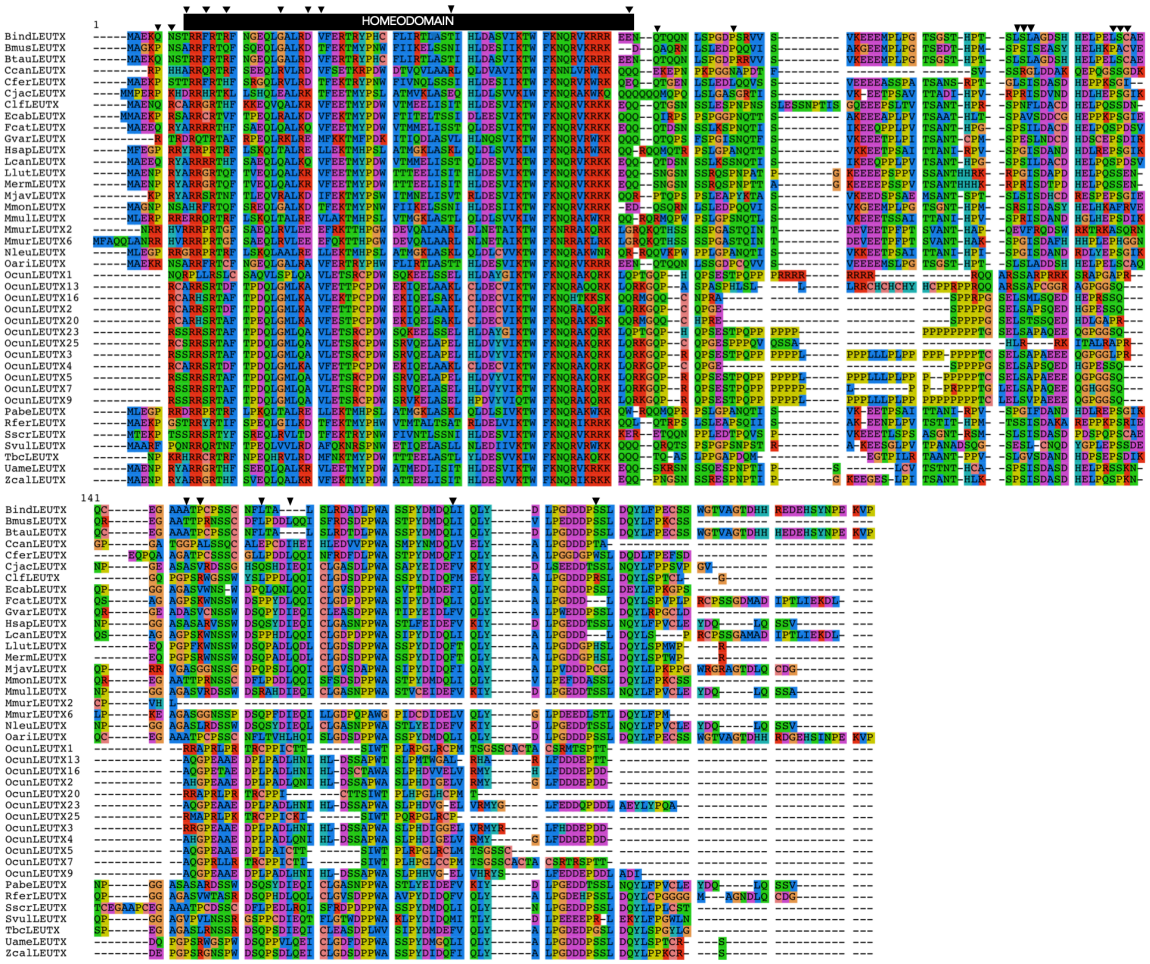


Figure S2.7C Untrimmed multiple sequence alignment of LEUTX protein sequences. Sites under positive selection, as identified by MEME[493], are marked with a black arrow. The homeodomain is marked with a black box. Species abbreviations as in Figure 2 of main text.

Figure S2.7D

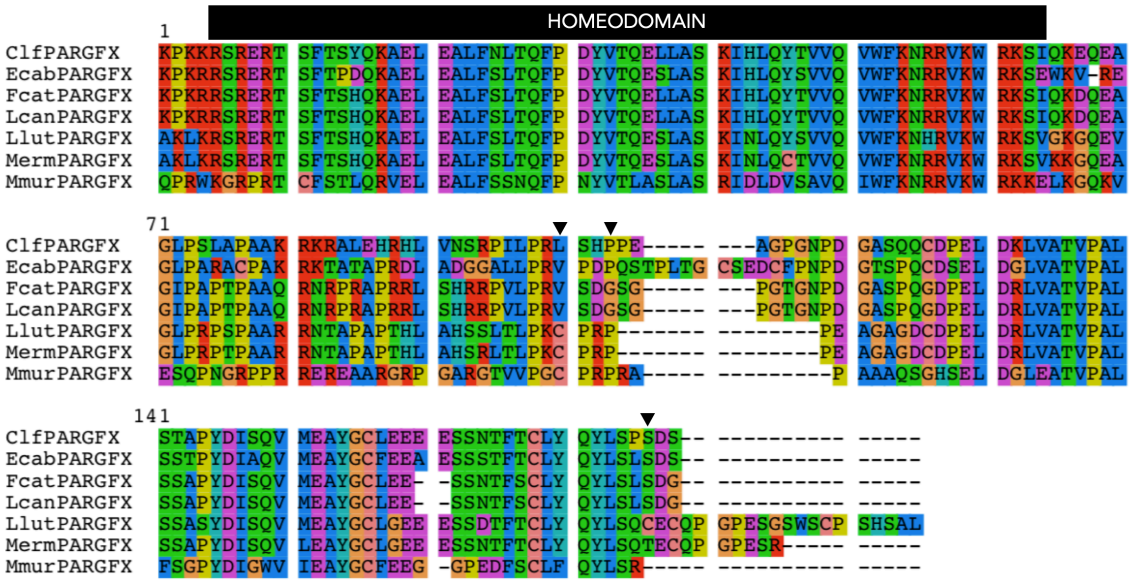


Figure S2.7D Untrimmed multiple sequence alignment of PARGFX protein sequences. Sites under positive selection, as identified by MEME[493], are marked with a black arrow. The homeodomain is marked with a black box. Species abbreviations as in Figure 2 of main text.

Figure S2.7E

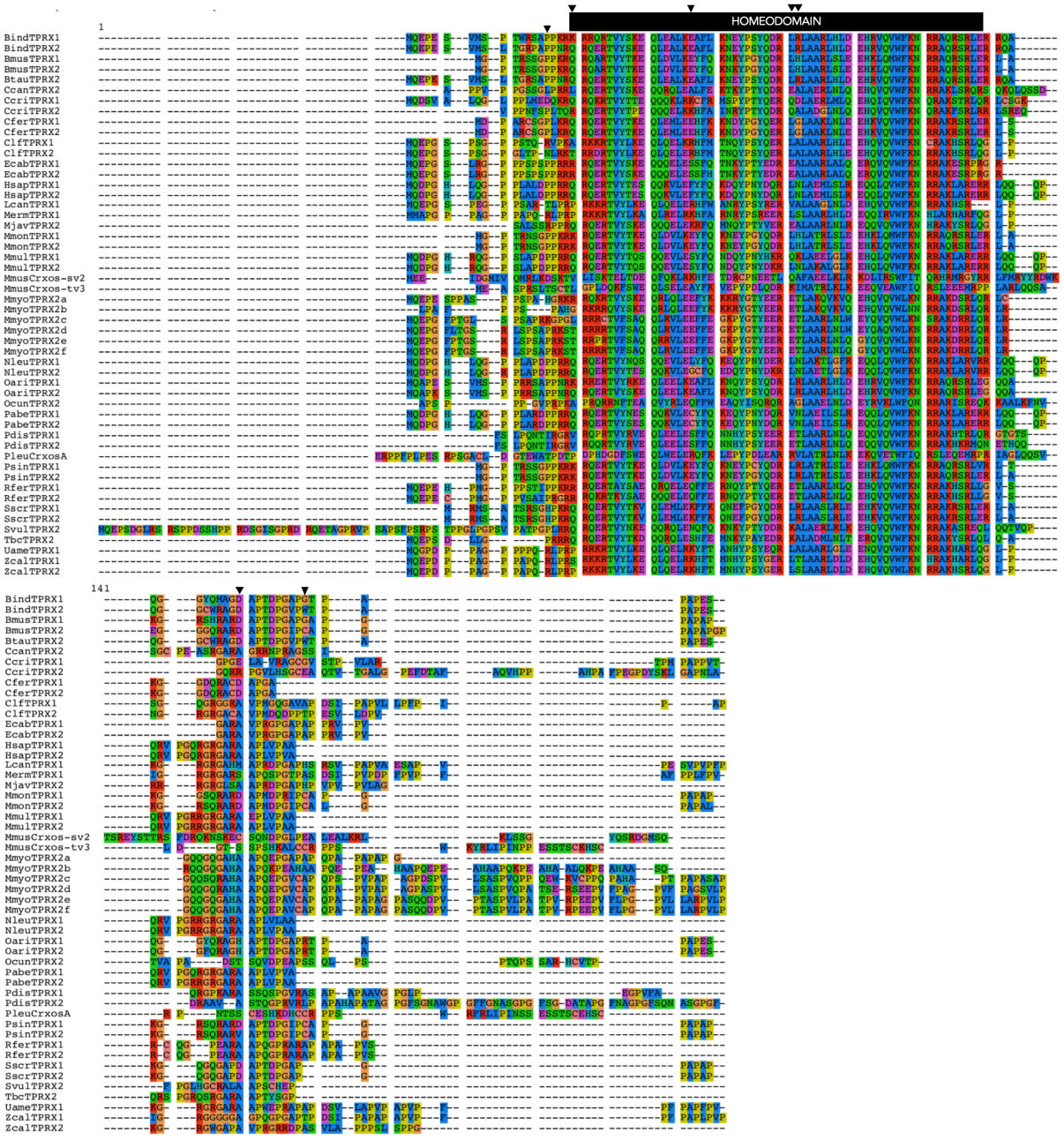
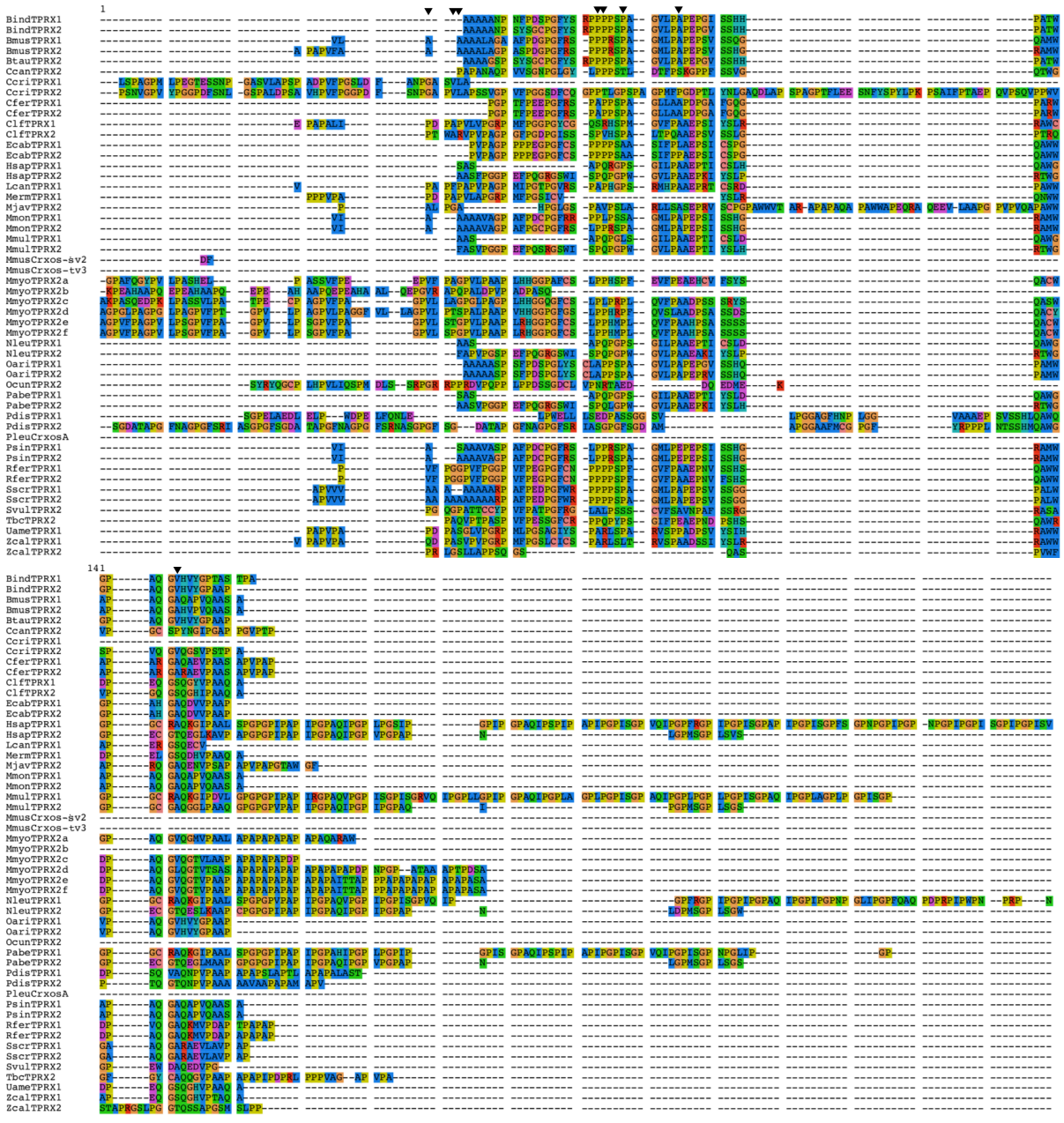


Figure S2.7E Untrimmed multiple sequence alignment of TPRX partition 1 protein sequences. Sites under positive selection, as identified by MEME[493], are marked with a black arrow. The homeodomain is marked with a black box. Species abbreviations as in Figure 2 of main text.

Figure S2.7F



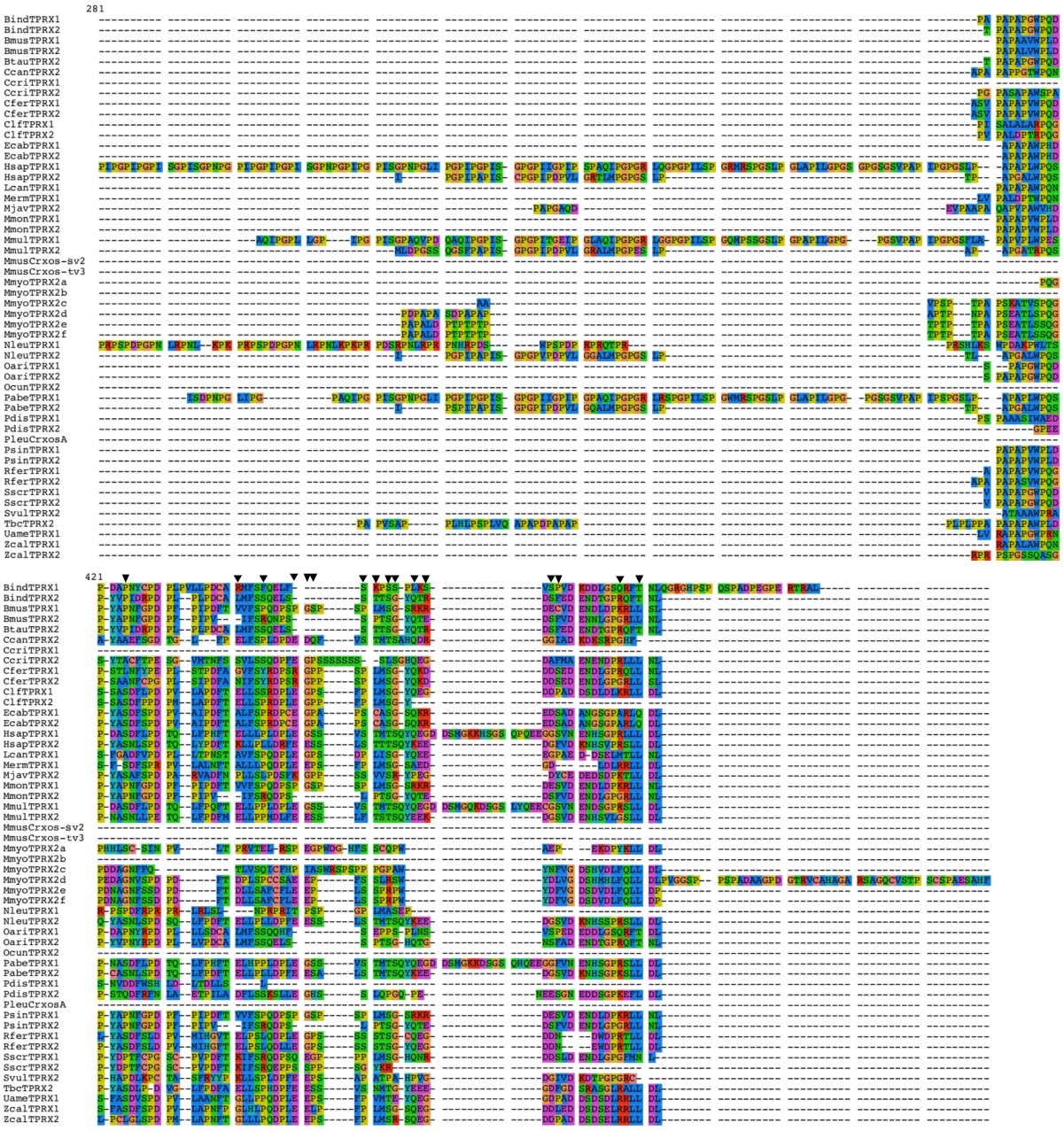


Figure S2.7F Untrimmed multiple sequence alignment of TPRX partition 2 protein sequences. Sites under positive selection, as identified by MEME[493], are marked with a black arrow. Species abbreviations as in Figure 2 of main text.

Figure S2.8

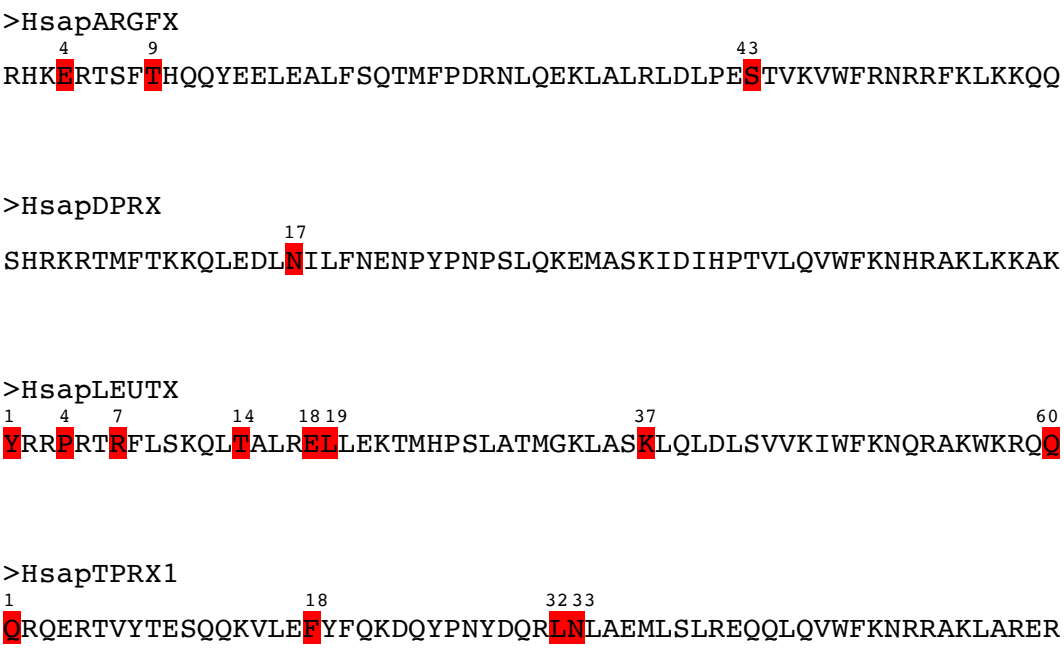


Figure S2.8 Positive selection in ETCHbox proteins. Figure shows homeodomains of *Homo sapiens* ETCHbox proteins with residues deduced to have been under positive selection highlighted in red. Numbers show the location of positively selected residues within the homeodomain.

Electronic supplementary material

ESM2.1 Supplementary Tables S2.1 – S2.13.

Chapter 3: PRD-class homeobox genes in bovine early embryos: function, evolution and overlapping roles

Thomas D. Lewin, Ali A. Fouladi-Nashta and Peter W. H. Holland

Introduction to manuscript

This chapter has been published as a first author original research article in *Molecular Biology and Evolution* (doi: https://doi.org/10.1093/molbev/msac098). It was first published online on 5th May 2022, and is reproduced here with small formatting changes as permitted as part of an integrated thesis. The authors credited on this manuscript, in order, are Thomas D. Lewin, Ali A. Fouladi-Nashta and Peter W. H. Holland (corresponding author).

Author contributions

Thomas D. Lewin and Peter W. H. Holland conceived and designed the study. Ali A. Fouladi-Nashta and TDL generated foetal fibroblast cell line. Wet lab experiments and data analysis were performed by TDL with input from PWHH. TDL and PWHH wrote the manuscript with input from AAF-N. All authors read and approved the final manuscript.

3.1 Abstract

Eutherian Totipotent Cell Homeobox (ETCHbox) genes are mammalian-specific PRD-class homeobox genes with conserved expression in the preimplantation embryo but fast-evolving and highly divergent sequences. Here we exploit an ectopic expression approach to examine the role of bovine ETCHbox genes and show that ARGFX and LEUTX homeodomain proteins upregulate genes normally expressed in the blastocyst; the identities of the regulated genes suggest that, in vivo, the ETCHbox genes play a role in coordinating the physical formation of the blastocyst structure. Both genes also downregulate genes expressed earlier during development and genes associated with an undifferentiated cell state, possibly via the JAK/STAT pathway. We find evidence that bovine *ARGFX* and *LEUTX* have overlapping functions, in contrast to their antagonistic roles in humans. Finally, we characterise a mutant bovine *ARGFX* allele which eliminates the homeodomain and show that homozygous mutants are viable. These data support the hypothesis of functional overlap between ETCHbox genes within a species, roles for ETCHbox genes in blastocyst formation and the change of their functions over evolutionary time.

3.2 Introduction

Eutherian Totipotent Cell Homeobox (ETCHbox) genes are a group of highly variable PRD-class homeobox genes specific to mammals. In humans, ETCHbox genes are expressed almost exclusively in the preimplantation embryo between the 4-cell stage and blastocyst, and this temporal expression pattern is broadly conserved in cattle[75,90]. The genes are thought to be involved in key transitions during embryonic development, acting as a vital component of regulatory networks active at the 8-cell stage and in the morula[75,90,96,101,102,104,524].

The six ancestral ETCHbox genes (*ARGFX*, *DPRX*, *LEUTX*, *PARGFX*, *TPRX1* and *TPRX2*) duplicated from the Otx-family gene *CRX* on the stem lineage leading to eutherians, and the ancestral eutherian ETCHbox repertoire likely consisted of one copy of each gene[90]. A curious property of the genes is their high propensity for duplication and loss, and the subsequent variation in the gene repertoires possessed by different eutherian species[90,96,525] (Figure 3.1). While humans (*Homo sapiens*) have a complement similar to that of the eutherian ancestor, having lost only *PARGFX* and undergone no duplications, the ETCHbox cluster of mice (*Mus musculus*) is very different, with *ARGFX*, *DPRX* and *PARGFX* completely lost, only remnants of *LEUTX* detectable and *TPRX1* (called *Crxos*) and *TPRX2* (known as *Obox*) both duplicated[96]. Indeed, *M. musculus* possesses a huge array of 66 *Obox* loci[96]. The ETCHbox repertoire of European cattle (*Bos taurus*) is more similar to that of human, except that *DPRX* is a pseudogene and there is an extra *TPRX* duplicate[90,525].

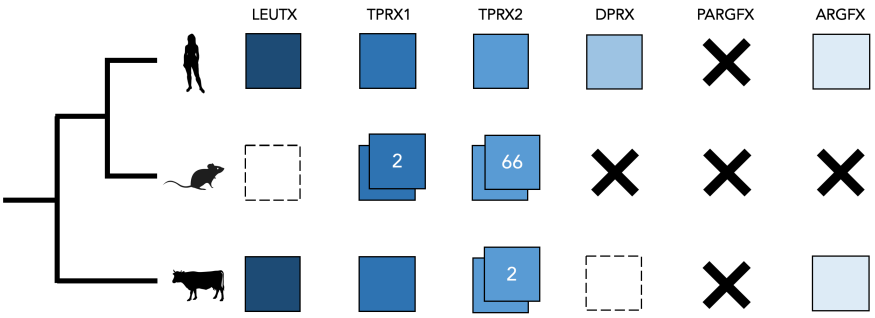


Figure 3.1 ETCHbox gene repertoires of human, mouse and cattle. Coloured squares indicate the presence of a functional gene. White squares with dashed borders represent pseudogenes, and a black X represents complete gene loss. Double squares indicate duplication; numbers show the quantity of duplicates of a given gene present in the genome.

In addition to extensive gain and loss, ETCHbox genes exhibit a much faster rate of sequence change across the eutherians than do other homeobox genes[90,525]. They are an archetypal example of asymmetric evolution, as their 'parent' *CRX* has been conserved while ETCHbox sequences have changed drastically[90,95,96,525]. These features are unusual for homeobox genes, which are generally highly conserved because they perform key roles in embryonic development and tissue patterning[106]. ETCHbox genes therefore pose a conundrum: they show extensive evolutionary change in gene number and amino acid sequence, yet they seem to have conserved temporal expression patterns and are hypothesised to possess key roles in the developing embryo. To resolve this paradox we need to know if similar expression is associated with similar and overlapping gene properties, and/or whether conserved expression between species is masking changes in gene functions driven by extensive amino acid sequence change.

To address these questions, we focus on *Bos taurus* (European cattle) because its repertoire of ETCHbox genes is similar to that of the ancestral eutherian and to that of humans. This allows comparisons to be made between orthologues of *Bos* and human. Furthermore, since humans are within the clade Euarchontoglires and cattle are within Laurasiatheria, this comparison may permit insight into ancient roles of ETCHbox genes in Eutheria.

There is also intrinsic interest in studying genes that may play regulatory roles in cattle embryonic development. Cattle are enormously important agriculturally: there are an estimated 1 to 1.5 billion head of cattle worldwide, almost 72 million tonnes of beef and buffalo meat and 840 million tonnes of milk is produced annually, and demand may increase further if the human population rises as expected[526–529]. It is therefore worrying that selective breeding in cattle has sometimes been associated with decreased fertility[530–533], and such infertility is a major cause of cattle culling worldwide[534–537]. Reduced fertility is caused in large part by a high failure rate of preimplantation embryos, the point at which ETCHbox genes are expressed, with up to 56% of cattle embryos lost before the blastocyst stage[538]. The cost of early embryonic mortality in cattle was estimated at an annual $1.28 trillion worldwide[539] and hence there is a clear imperative to better understand preimplantation development in cattle and identify genes and cellular processes that may be compromised in certain cattle breeds.

In this study, we aimed to examine the function of proteins encoded by ETCHbox genes of cattle and compare these to their orthologues in humans. The timing and location of ETCHbox gene expression makes them particularly challenging to study in their native setting, and previous work has therefore utilised ectopic expression to gain functional insights[75,90,96]. This approach, in which the homeobox genes are expressed in cultured cells, asks whether the encoded proteins can elicit changes to cells in which they are not normally expressed. Since homeobox genes encode transcription factors, this can be assayed by analysing the transcriptome after ectopic expression, a method which has proved highly informative. In this work, we utilise transcriptomic analysis after ectopic expression to understand the function of ETCHbox genes in *B. taurus*, characterise a mutant allele of *B. taurus ARGFX,* and test whether functional roles have changed during mammalian evolution.

3.3 Results

3.3.1 Transfection of homeobox genes into bovine foetal fibroblasts

In previous work, human ETCHbox gene function was studied by analysing transcriptomic changes induced by ectopic expression in cultured fibroblasts, in which the genes are not natively expressed[90]. To facilitate comparison between orthologous bovine and human ETCHbox genes, here we conducted similar experiments using bovine genes in bovine foetal fibroblast (BFF) cells. First, we established a primary culture of fibroblasts from a biopsy of *B. taurus* foetal skin (Supplementary Figure S3.1). Second, we constructed expression plasmids carrying bovine homeobox genes *ARGFX* or *LEUTX* driven by a constitutive promoter, each connected by a peptide linker to a C-terminal 3xFLAG tag to facilitate downstream analysis. We also constructed an expression plasmid carrying a mutant allele of *ARGFX* identified in the ARS-UCD1.2 reference genome[525]. This allele has a 13 base pair (bp) deletion before the homeodomain resulting in a frameshift and truncated predicted protein. Immunocytochemistry showed that bovine ARGFX and LEUTX proteins both localise to the nucleus when transfected into cells, as expected for transcription factors (Figure 3.2). The mutant ARGFX protein was detected in the nucleus but was also abundant in cytoplasm, suggesting presence of a weaker nuclear localisation signal. Control transfections were negative.

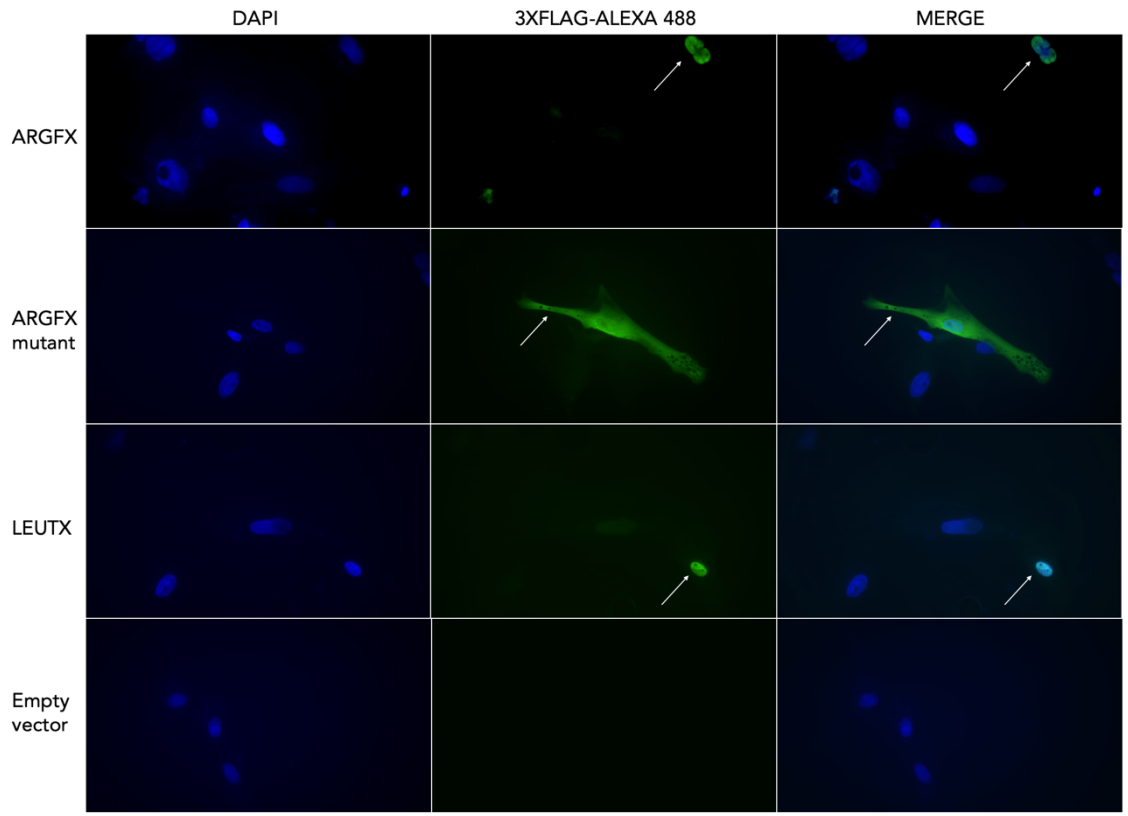


Figure 3.2 Immunocytochemistry analysis of BFF cells transfected with bovine ETCHbox genes. White arrows indicate cells with high expression of ETCHbox proteins. DNA is stained with DAPI (blue) to show cell nuclei. ETCHbox proteins are detected using Alexa Fluor 488-labelled antibodies to anti-FLAG antibodies (green). WT ARGFX and LEUTX proteins are localised to the nucleus; the mutant ARGFX is detected in the nucleus and cytoplasm. No green fluorescence is detected in samples transfected with the control plasmid.

3.3.2 Transcriptomic changes induced by bovine homeobox genes

We used transcriptomic analysis to determine whether ectopic expression of bovine *ARGFX* and *LEUTX* elicited gene expression changes in cells (Supplementary Table S3.1). After transfection, RNA-sequencing (RNA-seq) followed by differential gene expression (DE) analysis identified which loci in the genome increased or decreased in expression level (Figure 3.3A, B; Supplementary Tables S3.2a-f). We found 518 and 725 genes up- and downregulated respectively following *ARGFX* ectopic expression, which likely includes direct and indirect targets. *LEUTX* expression caused changes to a larger number of genes: 1286 increased in expression level and 1238 decreased. *ARGFX* expression downregulates more genes than it upregulates, and the genes that

decreased in expression after *ARGFX* transfection had a greater average fold change and lower *p*-value than those that increased in expression (Figure 3.3A); this suggests that ARGFX primarily has a repressive role in this cellular context. We performed gene ontology (GO) analysis on the DE gene sets (Supplementary Tables S3.3a-d).

We find that 47% of genes upregulated following *ARGFX* ectopic expression are also upregulated after *LEUTX* expression; 61% of genes downregulated by *ARGFX* expression are also downregulated by *LEUTX*. These overlaps between genes upregulated after *ARGFX* and *LEUTX* transfection (Fisher's exact test $p = 1.492 \times 10^{-90}$) or downregulated in response to both genes ($p = 5.176 \times 10^{-251}$) are much greater than overlaps for genes affected antagonistically by *ARGFX* and *LEUTX*, for which there are no significant overlaps (6% genes upregulated in response to *ARGFX* are downregulated in response to *LEUTX*, $p = 1.000$; 3% genes downregulated in response to *ARGFX* are upregulated in response to *LEUTX*; $p = 1.000$) (Figure 3.3C). These analyses suggest a large degree of overlap in the cellular roles of bovine ARGFX and bovine LEUTX: the proteins perform similar functions rather than acting antagonistically. This is in contrast to data from humans, in which ARGFX and LEUTX act in opposite directions on the same group of genes[90].

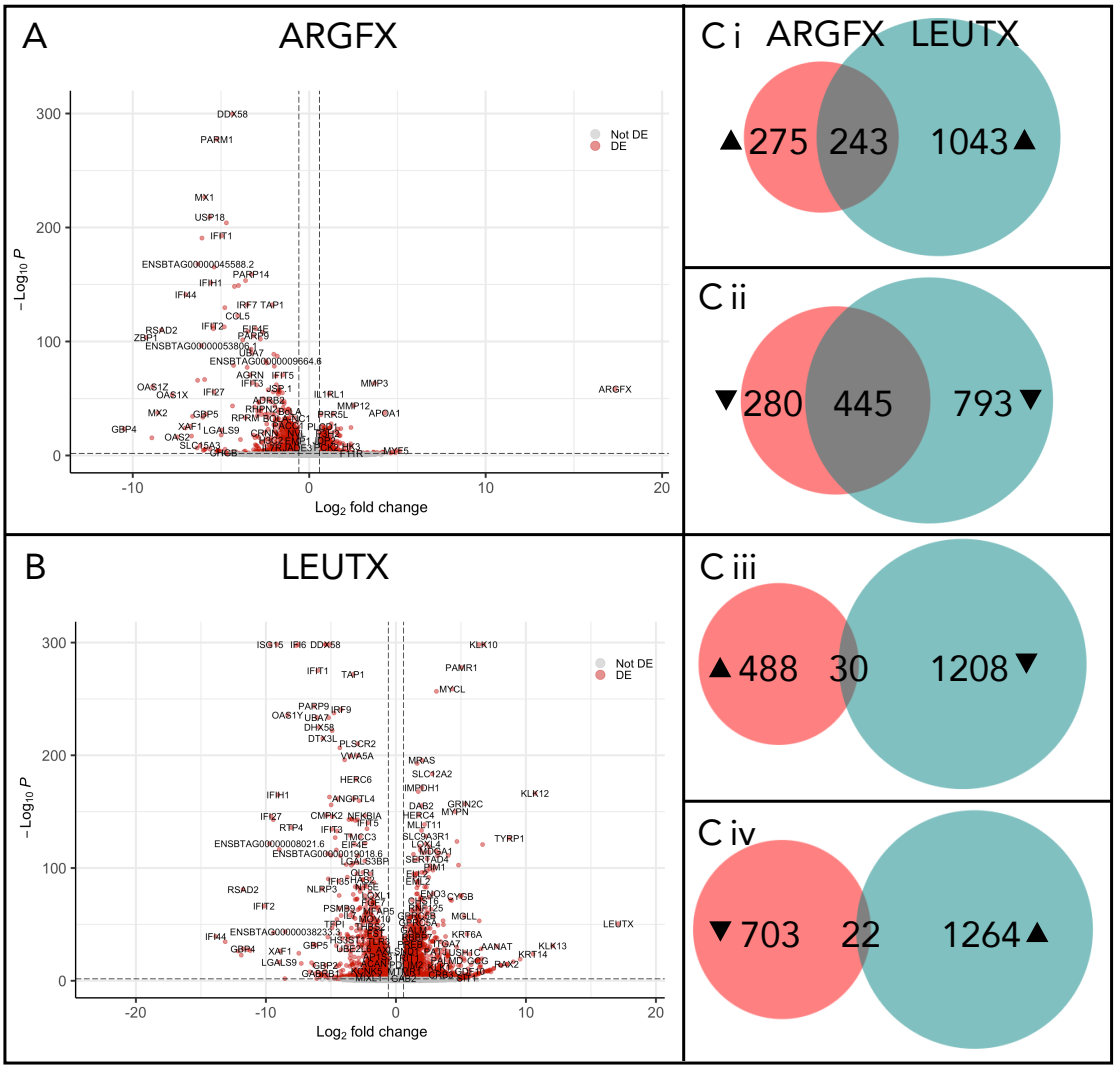


Figure 3.3 Transcriptional response to ETCHbox gene ectopic expression. Volcano plots show genes differentially expressed (DE) in response to wild type (WT) **(A)** *ARGFX* and **(B)** *LEUTX* ectopic expression. Each point represents a gene. Points in red are considered DE (adjusted $p < 0.05$, fold change [FC] > 1.5). **(C)** Overlaps between sets of genes DE in response to *ARGFX* and *LEUTX* ectopic expression: **(i)** *ARGFX* (red) up, *LEUTX* (blue) up; **(ii)** *ARGFX* down, *LEUTX* down; **(iii)** *ARGFX* up, *LEUTX* down; **(iv)** *ARGFX* down, *LEUTX* up.

3.3.3 Properties of bovine, human and mouse homeobox genes

The comparisons above suggest that bovine and human ETCHbox genes may not be performing identical roles. To investigate further, we compiled lists of genes differentially expressed in response to ETCHbox expression in three systems: bovine genes in bovine foetal dermal fibroblasts (this work), human genes in human dermal fibroblasts[90] and mouse genes in mouse embryonic fibroblasts[96]. We restricted lists of downstream genes

to 1:1 orthologues; our dataset includes 21,666 bovine genes, of which 71.7% (15,534) had identifiable 1:1 orthologues in human and 71.6% (15,503) in mouse. Fisher's exact test was then used to examine overlaps: this analysis asks whether overlap in the function of two genes is significantly greater than expected by chance. Indeed, many of the tests give apparently significant overlaps, hence we compared negative logs of adjusted *p*-values, which we call 'overlap scores', to extract biologically meaningful information: the higher the overlap score, the more the overlap exceeds the expected similarity. To provide context, previous work suggests that in humans ARGFX acts antagonistically to LEUTX and TPRX1, which have overlapping roles[90].

Considering genes differentially expressed in response to bovine *LEUTX* expression, the highest overlap scores are for overlaps between: bovine *LEUTX* down ('down' referring to downstream genes that decrease in expression) genes and human *LEUTX* down genes (overlap score = 62.032), bovine *LEUTX* down genes and human *TPRX1* down genes (overlap score = 56.611), and then bovine *LEUTX* up ('up' referring to downstream genes that increase in expression) genes and human *LEUTX* up genes (overlap score = 26.833) (Figure 3.4A, points marked '1'). For example, 170 of the 1019 bovine *LEUTX*-downregulated genes remaining after restriction to 1:1 orthologues are also downregulated by human *LEUTX*. Our test seeks to ask if this number is greater than expected by chance, and we conclude that it is (Fisher's exact test *p* value = 9.28×10^{-63}); from this we deduce some conservation of function. Overall, these data suggest that **at least some aspects of the function of LEUTX are similar between humans and cattle.** Bovine *LEUTX*-downregulated genes also have a large overlap (overlap score = 29.979) with human *ARGFX* upregulated genes (Figure 3.4A, point marked '2').

Considering genes downstream of bovine *ARGFX*, we find minimal overlap with gene sets affected by human *ARGFX*, suggesting **bovine and human ARGFX have different**

properties (Figure 3.4A, points marked '3'). Larger overlaps, indeed the most significant overlaps in the whole study, are found when bovine *ARGFX* is compared to human *LEUTX* or *TPRX1*: bovine *ARGFX* down genes and human *LEUTX* down genes (overlap score = 70.383); bovine *ARGFX* down genes and human *TPRX1* down genes (overlap score = 68.166) (Figure 3.4A, points marked '4'). This suggests that **ARGFX in cattle has similar properties to LEUTX and TPRX1 in humans.** There are also large overlaps between bovine and human *ARGFX* in opposite directions: bovine *ARGFX* down genes and human *ARGFX* up genes (overlap score = 21.460); bovine *ARGFX* up genes and human *ARGFX* down genes (overlap score = 10.320) (Figure 3.4A, points marked '5'). While this final comparison has a smaller overlap score, it is still the highest overlap for bovine *ARGFX* upregulated genes with any human gene. This suggests that **bovine ARGFX acts in the opposite direction to human ARGFX.**

Comparing bovine and murine genes, the highest similarity of lists of differentially expressed genes is between bovine *LEUTX* down and mouse *Crxos, OboxA1, Oboxa4* and *Oboxa7* down genes, but these similarities are smaller than bovine/human overlaps (overlap scores = 22.113, 23.008, 27.870 and 26.094, respectively) (Figure 3.4B, points marked '1'). This suggests that bovine ETCHbox genes are more similar in function to human ETCHbox genes than to those of mice. The highest overlap score for bovine *ARGFX* up genes is murine *Oboxa4* up genes (overlap score = 18.860) and for bovine *ARGFX* down genes is murine *Oboxa4* (overlap score = 16.250) and *Oboxa7* (overlap score = 17.984) down genes (Figure 3.4B, points marked '2'). These data suggest that both bovine ARGFX and LEUTX act in broadly the same direction as the tested mouse ETCHbox genes.

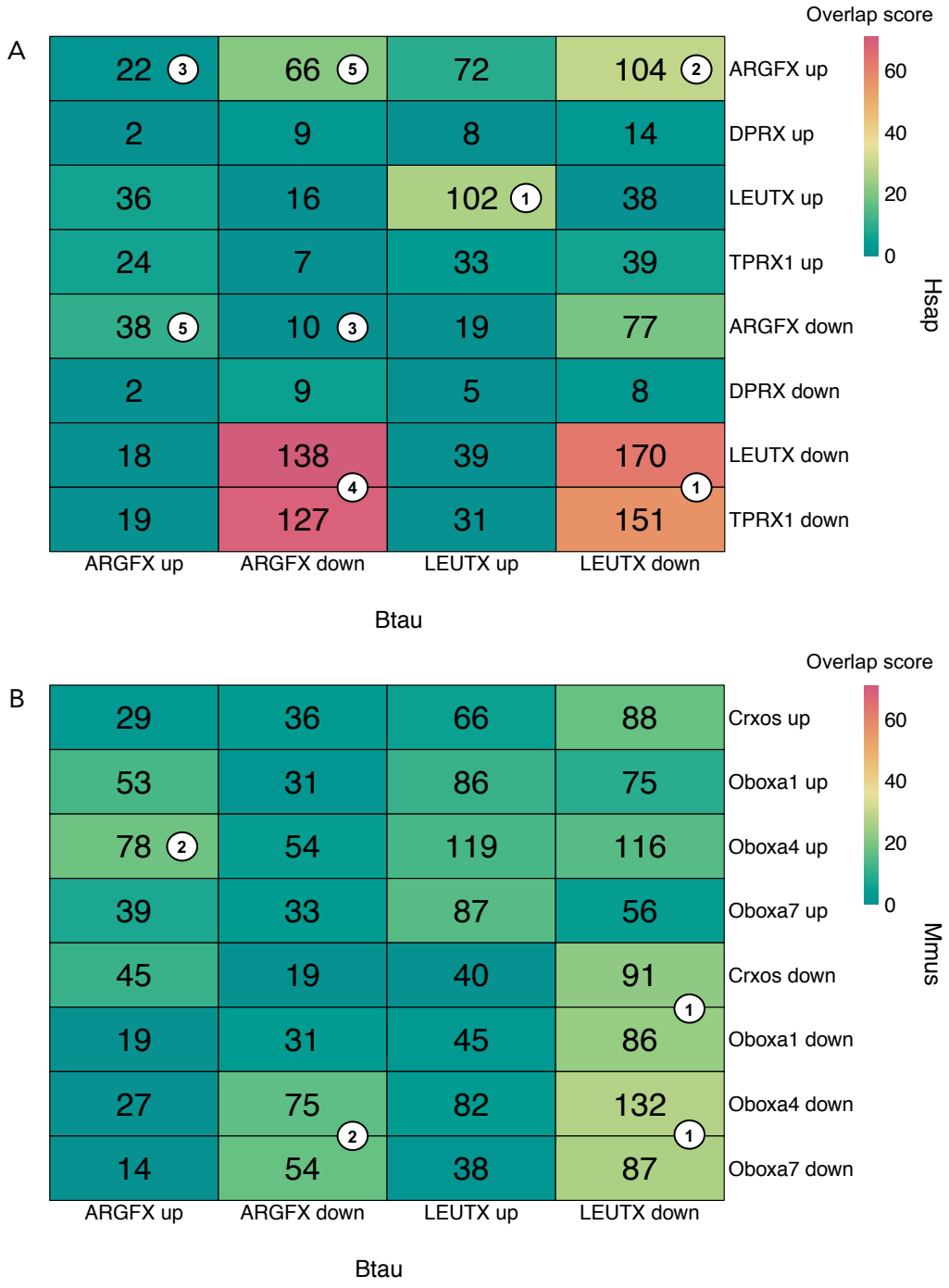


Figure 3.4 Overlap between genes differentially expressed in response to ectopic expression of (A) bovine and human and (B) bovine and mouse ETCHbox genes. For each comparison, numbers in the heatmap show the number of shared DE genes. Heatmap colour scale reflects overlap scores. The greater the overlap score, the more the overlap of regulated genes exceeds that which is expected by chance. Overlap scores consist of -$\log_{10}$ multiple comparison-adjusted *p*-values from Fisher's exact test for overlaps between gene sets. White dots with numbers mark points referred to in the main text. Abbreviations: Btau = *Bos taurus*, Hsap = *Homo sapiens*, Mmus = *Mus musculus*.

3.3.4 Relevance of target genes to bovine development

The bovine *ARGFX* and *LEUTX* genes are expressed in a tight temporal window during preimplantation embryonic development[90]. To investigate if the gene expression changes induced by ectopic expression in fibroblasts have functional relevance, we employed temporal expression clustering. This approach tests whether the set of downstream target genes in fibroblast cells are enriched for genes normally expressed in specific temporal patterns in the embryo. Using published transcriptomic data, we determined temporal expression profiles for all bovine genes expressed from oocyte to blastocyst and clustered these into 100 groups sharing similar temporal expression profiles. We then tested whether the gene sets differentially expressed in response to *ARGFX* or *LEUTX* expression were enriched for genes in any expression profiles (Supplementary Table S3.4).

Genes that increased in expression after ectopic expression of both *ARGFX* and *LEUTX* are enriched for profile 48 (Fisher's exact test adjusted p = 0.005 and p = 0.041, respectively), which comprises genes that show minimal expression before the blastocyst but are then strongly expressed at the blastocyst stage (Figure 3.5A; Supplementary Figure S3.2). This supports the above conclusion that ARGFX and LEUTX functions are overlapping. The gene set downregulated by ectopic expression of *LEUTX* is enriched for genes belonging to expression profile 45 (Fisher's exact test adjusted p = 0.005). This profile comprises genes expressed specifically at the two-cell stage (Figure 3.5B). Finally, profile 36 is enriched in the gene set downregulated following *ARGFX* ectopic expression (Fisher's exact test adjusted p = 0.011); these genes are activated specifically at the 8-cell stage (Figure 3.5C).

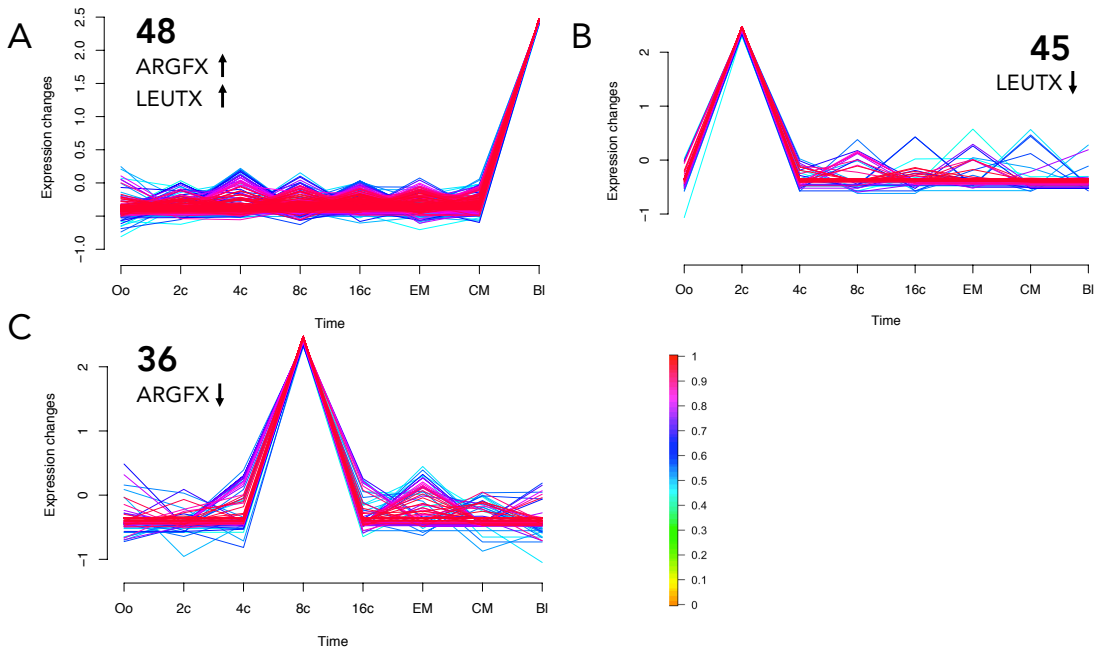


Figure 3.5 Temporal expression profiles for bovine genes enriched in the downstream target gene sets of ARGFX and LEUTX. Plots show normalised, standardised expression level as a function of developmental stage. Each line is the expression profile of one gene. Colours of lines show membership score from 0 to 1; membership score is a measure of how well the gene fits the profile. Genes with a score < 0.5 were excluded from profiles. **(A)** Profile 48, *ARGFX* up and *LEUTX* up. **(B)** Profile 45, *LEUTX* down. **(C)** Profile 36, *ARGFX* down. Abbreviations: Oo = oocyte, 2c = 2-cell, 4c = 4-cell, 8c = 8-cell, 16c = 16-cell, EM = early morula, CM = compact morula, Bl = early blastocyst.

We note that profile 48, upregulated by *ARGFX* and *LEUTX*, includes several genes with known roles in cell junctions and cell adhesion (*DSC2*, *EPDR1, FLRT2*), the cytoskeleton (*ACTG2, EPS8L2, HCK, MIB2, MYADM, PALLD, PLD2, SLC9A3R1, SMTN*) or the extracellular matrix (*COL15A1*, *COL5A2*, *COL5A3*, *COL6A3*, *CTHRC1*, *PXDN*) (Supplementary Figure S3.2), functions which are central to embryo compaction and the physical formation of the blastocyst. This suggests that bovine *ARGFX* and *LEUTX* regulate aspects of structural formation of the blastocyst. Moreover, GO cellular component analysis of the shared gene set upregulated by *LEUTX* in both cattle and in humans (Supplementary Table S3.5) shows extracellular matrix (GO:0031012) as an enriched function (p = 7.17 x 10^{-4}), suggesting that this functional role may be conserved between the two species.

3.3.5 ETCHbox genes and cellular potency

Induced pluripotent stem cells (iPSCs) can be generated from mouse and human somatic cells by the addition of just four transcription factors: MYC, KLF4, POU5F1 and SOX2[148,149]. In this work, we find that bovine ARGFX and LEUTX both downregulate *MYC*, and ARGFX also downregulates *KLF4*. Inspecting DE gene lists from published works, we find that human DPRX, LEUTX, and TPRX1 all downregulate *KLF4* and that four tested mouse ETCHbox proteins (Crxos, Oboxa1, Oboxa4, Oboxa7) also downregulated *Sox2* when expressed in fibroblasts[90,96]. Furthermore, bovine ARGFX and LEUTX both downregulate the pluripotency-related gene *GBP4* and, indeed, the total bovine LEUTX-downregulated gene set is enriched for genes defining an 'undifferentiation network signature', which is composed of genes that are considered markers of an undifferentiated cell state[540] (Fisher's exact test, $p = 0.002$; Supplementary Table S3.6). Overall, both bovine ARGFX and LEUTX downregulate genes linked with an undifferentiated cell state.

3.3.6 Frequency and properties of a mutant bovine *ARGFX* allele

We previously identified a 13 bp deletion in the coding region of the *ARGFX* gene of the *B. taurus* reference genome ARS-UCD1.2 and showed that the indel is polymorphic, as several RNA-seq datasets exhibit the wild type (WT) allele[525]. Here we investigate the frequency of the deletion allele and its effect on protein function. The mutation, a 13 bp deletion in exon 2, causes a frameshift that eliminates the homeodomain and truncates the protein (Figure 3.6A). We used the Bovine Genome Variation Database[541], which includes data from 432 *B. taurus* and *B. indicus* samples from 54 breeds, to assess distribution of the two alleles. We find 25 of 54 (46%) breeds have at least one sampled individual with the mutant allele, and the limited sampling of some breeds suggests that the incidence may be higher. The reported mutant allele frequency is 0.122 across all samples and is highest in Jersey (0.917, 12 samples) and Hereford cattle (0.667, 21 samples) (Figure 3.6B; Supplementary Table S3.7). Consistent with this, the reference

genome ARS-UCD1.2 is from a Hereford individual. The deletion is present at least in low frequency in breeds originating from Africa, Asia and Europe, and is found in both *B. taurus* and *B. indicus*. No other indels in the coding sequence of *ARGFX* are identified in the 432 samples.

To test whether individuals homozygous for the mutant *ARGFX* allele can be viable, we used PCR to screen samples from the Jersey breed. From 11 individuals, we found three that are homozygous for the mutant allele, demonstrating that the homozygous mutant is not lethal (Figure 3.6C). Five individuals were heterozygous, two were homozygous WT and one sample was inconclusive. We sequenced the *ARGFX* gene of all three homozygous mutant samples to confirm that the deletion found by PCR in Jersey cattle is the same 13 bp deletion identified in the reference genome. We also used PCR to genotype the BFFs used for ectopic expression; these were homozygous for the WT *ARGFX* allele (Figure 3.6C).

Analysis of RNA-seq data[442] from embryo pools containing both alleles shows that *ARGFX* continues to be expressed from a mutant locus. To understand the impact on protein function, we generated an expression construct containing the mutant *ARGFX* allele, transfected this into bovine fibroblasts, conducted transcriptome sequencing, and compared the sets of genes that changed in expression in response to transfection of WT or mutant *ARGFX* (Figure 3.6D). We found that the mutant allele has a much smaller set of downstream effects than does WT *ARGFX*, but that the DE genes overlapped strongly with those regulated by WT *ARGFX*. Just 18% and 41% of genes up- and downregulated by WT *ARGFX*, respectively, were also regulated by the mutant *ARGFX* but 89% of the genes that increased in expression and 92% of those that decreased in expression after transfection of mutant *ARGFX* had the same response to transfection of WT *ARGFX* (Figure 3.6E). This suggests that the truncated mutant protein is able to perform a subset of the functions of WT ARGFX, but not additional functions. This

retention of a limited function despite protein truncation may be facilitated by cofactors that bind to the intact part of the protein.

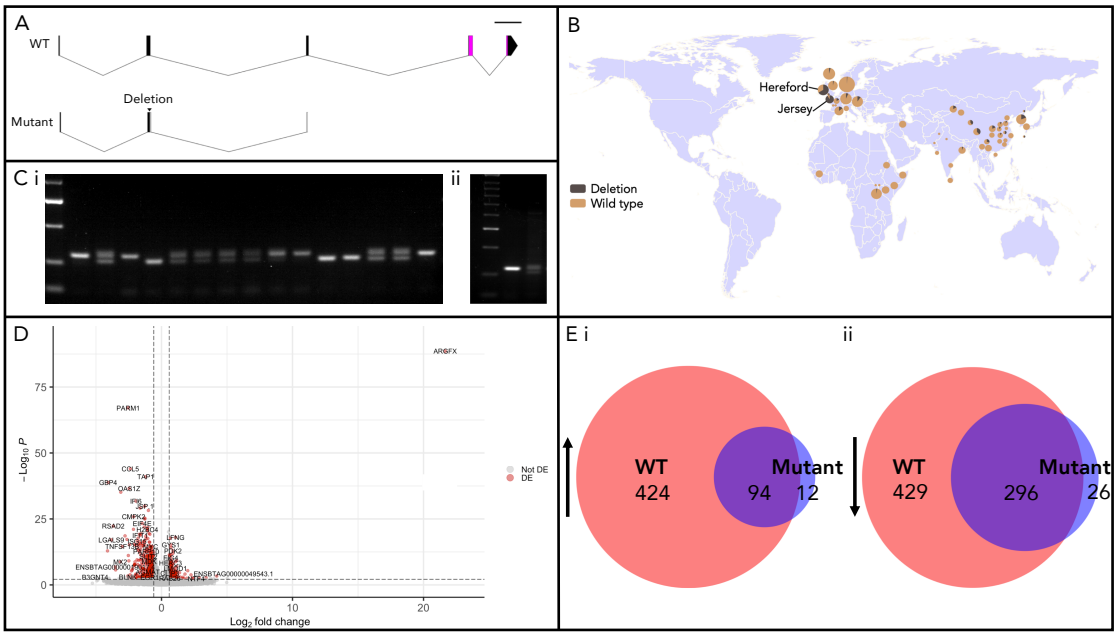


Figure 3.6 Characterisation of a mutant bovine *ARGFX* allele. (A) Structure of wild type (WT) and mutant genes and deduced proteins. The WT gene comprises five coding exons, with the homeobox (pink) split across exons 4 and 5. The mutant allele has a 13 bp deletion in exon 2 causing a frameshift (grey) and an early stop codon truncating the protein in exon 3 before the homeodomain. Scale bar 1 kilobase (kb). **(B)** Map showing the distribution of the mutant allele. Dark brown/black = deletion, pale brown = WT. Pie charts show allele frequency for each breed; pie size proportional to sample size. Pie charts are located at the approximate location of the breed's origin. Map generated using Bovine Genome Variation Database[541]. **(Ci)** Agarose gel showing that individuals can be homozygous for the mutant *ARGFX* allele. PCR Primers were used that anneal either side of the deletion site, producing a shorter amplicon in the deletion allele. Lane 1 = Bioline HyperLadder 50 bp. Lanes 2 and 16 = homozygous WT control. Lanes 3 and 15 = heterozygous control. Lanes 4 and 11 are homozygous WT Jersey individuals, lanes 5, 12 and 13 are homozygous mutant Jersey individuals and lanes 6, 7, 8, 9 and 14 are heterozygous Jersey individuals. Lane 10 is inconclusive, with a strong WT band and very weak mutant band. **(Cii)** Agarose gel showing that BFF cells have a WT *ARGFX* gene**.** Lane 1 = Bioline HyperLadder 50 bp. Lane 2 = BFF band. Lane 3 = heterozygote bands. **(D)** Volcano plot showing genes differentially expressed (DE) in response to *ARGFX* mutant ectopic expression. Each point represents a gene. Points in red are considered DE (adjusted $p < 0.05$, fold change [FC] > 1.5). **(E)** Overlaps between sets of genes differentially expressed in response to WT or mutant *ARGFX* ectopic expression: **(i)** upregulated by both and **(ii)** downregulated by both.

3.4 Discussion

The ETCHbox genes exhibit several puzzling characteristics. They encode homeodomain proteins with a highly specific, conserved temporal expression pattern in the preimplantation embryo, and which are inferred to play roles at key developmental milestones such as embryonic genome activation[75,90,104]. Given this, one might expect them to be highly conserved but, in reality, the genes display rapid sequence evolution, and are also duplicated and lost at an unusually high rate[90,96,525].

We utilised ectopic expression to examine the roles of ETCHbox proteins in cattle, and find evidence for large functional overlap between bovine ARGFX and LEUTX. This is consistent with the hypothesis that redundancy is one of the causes of the genes' unusual evolutionary patterns[525]. Such redundancy may also explain why a frameshift mutation of *ARGFX* is not lethal when homozygous and has risen to a population allele frequency of over 10%; however, we stress that we do not know if homozygous mutant embryos have the same rate of survival as those with a fully functional copy of *ARGFX*. Regardless, there would be reduced selection against a sub-optimal allele if gene function in the preimplantation embryo is redundant with *LEUTX*. Indeed, of the 424 ARGFX up genes and 429 ARGFX down genes which the mutant ARGFX protein fails to properly regulate, 45% and 48% respectively are also regulated in the same direction by LEUTX. This provides further evidence that LEUTX is able to directly compensate for the presence of a truncated ARGFX protein. Such redundancy may underpin why ETCHbox genes are lost in evolution at an elevated rate and, in this example, we may be observing this process in action: redundancy between two ETCHbox genes allowing one to move toward pseudogenisation.

To test whether ETCHbox gene sequence divergence is accompanied by functional change, we compared the transcriptional response to ectopic expression between

humans and cattle. Human and bovine ETCHbox orthologues share 48 to 77% amino acid sequence identity in their homeodomains and 36 to and 54% identity across the full protein. We were surprised to find that *B. taurus* and *H. sapiens* ARGFX proteins show minimal functional overlap, and that *B. taurus* ARGFX likely performs a function similar to that of *H. sapiens* LEUTX, and opposite to that of *H. sapiens* ARGFX. The change in broad functional roles is not seen, however, for LEUTX proteins, which have highly similar effects in human and bovine cells. We also found that bovine ETCHbox genes regulate gene sets that are more similar to those regulated by human ETCHbox genes than murine ETCHbox genes; this is consistent with the greater similarity of bovine ETCHbox sequences to those of humans than those of mice[96]. In interpreting these results, we note the caveat that the human and cattle experiments were similar but not identical: analysis of human ETCHbox genes used ectopic expression in adult cells and analysed transcriptome change 48 hours (h) post-transfection[90], the bovine experiment used foetal cells, with RNA extracted after 72 h. These differences might cause variation in the gene sets regulated. Overall, however, these data suggest that significant changes in functional roles are a feature of ETCHbox evolution.

We utilised temporal expression profiling of bovine embryonically-expressed genes to gain insights into the function of ETCHbox proteins in the preimplantation embryo. The logic to this analysis is that if, during normal embryonic development, bovine ETCHbox proteins upregulate or downregulate suites of genes at a particular stage, this may be replicated when the genes are ectopically expressed in cell culture. We find one temporal profile enriched in both the *ARGFX* and *LEUTX* upregulated gene sets: this profile consists of genes that have minimal expression up until the blastocyst but are then strongly induced. We suggest, therefore, that bovine *ARGFX* and *LEUTX* are involved in blastocyst formation. Interestingly, many blastocyst genes that are upregulated by *ARGFX* and/or *LEUTX* possess known roles in cell junctions and cell adhesion, the cytoskeleton or the extracellular matrix (Supplementary Figure S3.2). These functions

are critical to embryo compaction and the structural changes that occur during morula and blastocyst formation[542–555]; this suggests that ETCHbox genes are involved in coordinating the physical process of structural formation of the blastocyst. We note, however, that, despite GO analyses suggesting that this role may be conserved in humans, a similar response of 'blastocyst genes' was not detected in experiments using human cells[90]. One possible explanation for the difference is that blastocyst gene expression could be a downstream effect, not observed in the human experiments when transfected cells were harvested sooner[90]. We do not find induction of the proinflammatory genes *S100A9* or *TNFA*, implicated in inducing an inflammation response in uterine cells at implantation[556].

The timing of ETCHbox genes' expression in cattle and human, peaking just before blastocyst formation, and for *ARGFX* continuing into the blastocyst[90], is consistent with the hypothesis that bovine and human ETCHbox genes are involved in coordinating the physical process of blastocyst formation. A function for ETCHbox genes in embryo compaction and the structural formation of the blastocyst is also consistent with their absence in the sister lineage to placentals, marsupials, which do not undergo compaction, possess a morula stage or form an inner cell mass as part of their blastocyst[92,127,557,558]. The upregulation of the same temporal profile by both ARGFX and LEUTX supports the contention that bovine ARGFX and LEUTX have overlapping functions.

The bovine LEUTX-downregulated gene set is enriched for genes expressed at the 2-cell stage, suggesting that LEUTX may downregulate the 2-cell programme as part of the exit from the 2-cell stage. Intriguingly, ARGFX downregulates a temporal profile comprising genes expressed at the 8-cell stage. This profile is similar to that identified by Maeso et al. (2016)[90] as downregulated by ectopic expression of human LEUTX and TPRX1, and upregulated by ARGFX (Figure 3.7); this supports our earlier conclusion

that *B. taurus* ARGFX performs the opposite role to human ARGFX and is more similar functionally to human LEUTX. Overall, the temporal expression profiles enriched in the gene sets regulated by ectopic expression of ETCHbox genes are consistent with the timing of ETCHbox expression in cattle development, indicate that the ectopic expression experiments lead to the activation of realistic embryonic targets, and suggest that ETCHbox genes are involved in blastocyst formation.

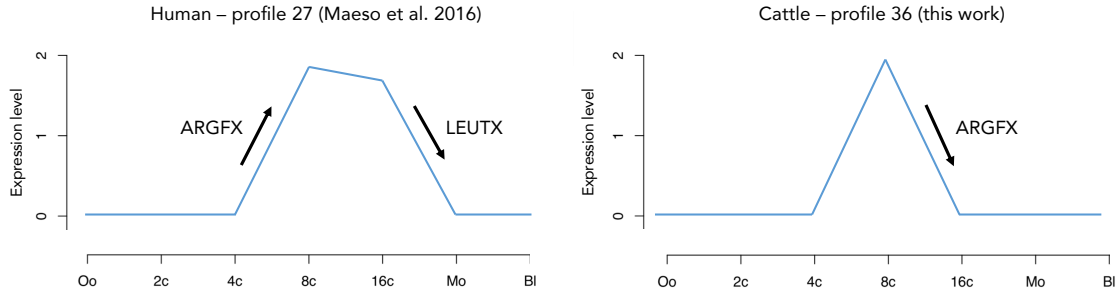


Figure 3.7 ETCHbox target gene expression profiles in human and bovine. Temporal expression profiles for genes upregulated in response to *ARGFX* expression in human[90] and downregulated in response to *ARGFX* expression in bovine (this work). Abbreviations: Oo = oocyte, 2c = 2-cell, 4c = 4-cell, 8c = 8-cell, 16c = 16-cell, Mo = morula, Bl = early blastocyst.

We find that human and bovine ETCHbox genes downregulate the expression of Yamanaka factors, and that the total bovine *LEUTX*-downregulated gene set is enriched for genes of the 'undifferentiation network signature' of the early preimplantation embryo[540]. Cells of the preimplantation embryo undergo a reduction in their potency as the totipotent blastomeres of early development undergo lineage specification and differentiation to form the blastocyst [282,559]; the downregulation of genes associated with an undifferentiated state is therefore consistent with the timing of ETCHbox expression just prior to blastocyst formation.

Finally, we also examined which gene targets had the strongest response to ETCHbox expression. Eight of the 10 downregulated genes with the highest fold change were common between *ARGFX* and *LEUTX*. Seven of these eight genes (*GBP4*, *IFI44*, *MX2*, *OAS1X*, *OAS1Z*, *RSAD2* [*VIPERIN*], *ZBP1*) are known to be interferon-stimulated genes

(ISGs), which are regulated via the JAK/STAT signalling pathway; this pathway is active during preimplantation development and is involved in the maintenance of pluripotency and the prevention of differentiation[560–571]. Two more ISGs, *XAF1* and *IFIT2*, are in the top 10 downregulated genes for *ARGFX* and *LEUTX*, respectively[572,573]. Moreover, the overall gene sets downregulated in response to *ARGFX* and *LEUTX* expression are each enriched for six GO Biological Process terms related to the interferon response (Table 3.1). Overall, these data suggest that the expression of ISGs is downregulated by bovine *ARGFX* and *LEUTX*.

Interferon-related GO term	Bovine *ARGFX* FDR-adjusted *p*-value	Bovine *LEUTX* FDR-adjusted *p*-value	Bovine *LEUTX* and human *LEUTX* overlap FDR-adjusted *p*-value
cellular response to interferon-beta (GO:0035458)	NS	NS	2.89E-03
cellular response to interferon-gamma (GO:0071346)	NS	NS	1.24E-03
cellular response to type I interferon (GO:0071357)	1.32E-02	1.66E-02	5.63E-04
interferon-gamma-mediated signaling pathway (GO:0060333)	NS	NS	1.31E-05
positive regulation of interferon-alpha production (GO:0032727)	NS	NS	1.28E-02
positive regulation of interferon-beta production (GO:0032728)	NS	NS	3.99E-06
positive regulation of type I interferon production (GO:0032481)	NS	NS	1.48E-06
regulation of interferon-alpha production (GO:0032647)	NS	NS	1.30E-03
regulation of interferon-beta production (GO:0032648)	NS	NS	1.34E-07
regulation of type I interferon production (GO:0032479)	NS	NS	3.63E-08
response to interferon-alpha (GO:0035455)	1.74E-03	2.71E-03	8.50E-03
response to interferon-beta (GO:0035456)	2.54E-04	3.91E-03	7.75E-05
response to interferon-gamma (GO:0034341)	4.69E-03	1.18E-03	5.27E-05
response to type I interferon (GO:0034340)	8.14E-04	3.50E-03	6.51E-07
response to type III interferon (GO:0034342)	NS	NS	3.40E-02
type I interferon signaling pathway (GO:0060337)	4.57E-02	3.82E-02	6.83E-03

Table 3.1 Interferon-related GO terms enriched in *ARGFX* and *LEUTX*-downregulated gene sets. *P*-values are from binomial tests with FDR corrections for

multiple testing. Abbreviations: FDR = false discovery rate; GO = gene ontology; NS = not significant.

Through what mechanism are ETCHbox proteins regulating ISGs? We suggest that ETCHbox proteins are capable of downregulating the JAK/STAT pathway: we find that bovine LEUTX and/or ARGFX stimulate downregulation of expression of *JAK2*, *STAT1*, *STAT2* and upregulation of the JAK/STAT inhibitors *PIAS3* and *PIAS4* (Supplementary Table S3.2). Since cell surface receptors compete for JAKs, meaning that JAK concentration is often a limiting factor for JAK/STAT signalling[568,574], this downregulation of *JAK2* may result in a direct downregulation of JAK/STAT signalling potential. We note that there is also downregulation of the JAK/STAT inhibitors SOCS1-5; while this may initially seem counterintuitive, these are also ISGs and are upregulated by JAK/STAT, so a decrease in their expression is consistent with reduced JAK/STAT signalling[575–577]. The JAK/STAT signalling pathway plays important roles in preimplantation development at the time of normal ETCHbox expression. For example, in mice the pluripotency factor POU5F1 acts via JAK/STAT3 signalling to induce pluripotency[578], and JAK/STAT3 signalling alone can induce naïve pluripotency during reprogramming[579,580]. In cattle, the JAK/STAT pathway is vital for inner cell mass gene expression and development[581]. Indeed, it has been found that JAK/STAT signalling is essential for pluripotency maintenance, self-renewal and the prevention of differentiation[578,582–584]; the downregulation of JAK/STAT signalling is therefore consistent with the above result that genes involved in the maintenance of an undifferentiated cell state are also downregulated by ETCHbox genes. Together with the temporal profile analyses, these data suggest that ETCHbox genes are involved in the promotion of blastocyst structural formation, and downregulate genes involved in the undifferentiation network in cells, possibly via the downregulation of JAK/STAT signalling. GO terms related to interferon signalling are strongly enriched in the set of genes commonly downregulated by human

and bovine LEUTX (Table 3.1 and Supplementary Table S3.8); this suggests that this function is conserved between humans and cattle.

3.5 Conclusions

Fast-evolving homeobox genes pose a conundrum in developmental biology since they contrast so markedly from the evolutionary conservation typical of such genes. In this work, we utilised an ectopic expression approach to compare the functions of bovine ETCHbox proteins to those of humans and mice and found evidence suggesting that, remarkably, the function of bovine ARGFX is highly different to that of its human orthologue and is more similar to that of human LEUTX. This suggests that the gene's function has undergone a transition since the divergence of humans and cattle. In addition, we found that bovine ARGFX and LEUTX upregulate blastocyst-expressed genes, including genes involved in cell adhesion, the extracellular matrix and the cytoskeleton; we suggest that these genes have overlapping roles in orchestrating structural changes during blastocyst formation. Our results also suggest that ETCHbox genes downregulate JAK/STAT signalling, and that this may result in the downregulation of genes associated with an undifferentiated cell state. Finally, the existence of a frameshift allele that truncates the bovine ARGFX protein and eliminates many of its functions, but is non-lethal when homozygous, supports the hypothesis of redundancy between ETCHbox proteins.

3.6 Materials and methods

3.6.1 Cell isolation and culture

Primary bovine foetal fibroblasts (BFFs) were isolated from the torso skin of a 6-month old *Bos taurus* foetus from a pregnant female sacrificed at an abattoir using a method similar to that used to isolate human dermal fibroblasts[585]. Briefly, a 1 cm^2 section of skin was excised, transported on ice and washed 3 times with sterile phosphate buffered

saline (PBS) (Gibco #10010015). Subcutaneous fat was scraped off and the tissue minced with a scalpel. Five 2 mm^2 pieces were placed in each well of a 0.001% poly-L-lysine-coated (Sigma-Aldrich, #P4832-50ML) 6-well plate, covered with a sterile glass coverslip, and 2 mL BFF media added. BFF media is Dulbecco's Modified Eagle Medium (Gibco #41965039) with 10% heat-inactivated foetal bovine serum (Gibco #10500064), 1% penicillin-streptomycin (Gibco #15140122) and 1% amphotericin B (Gibco #15290026). Outgrowth was allowed for 18 days at 37°C with 5% CO_2, with the culture media changed every 3 days before cells were washed three times in 2 mL PBS and detached with 2 mL TrypLE™ Express enzyme (Gibco #12604013) at 37°C with 5% CO_2 for 20 minutes. TrypLE™ Express was neutralised with 5 mL BFF medium, the coverslip removed, cells and medium transferred to a 50 mL Falcon tube and centrifuged for 5 minutes at 1200 rpm. Cells were washed twice with 10 mL BFF medium and centrifuged, then seeded into 0.001% poly-L-lysine-coated T75 flasks. To maintain growth, fibroblasts were incubated at 37°C with 5% CO_2 and passaged one flask to four every 3-4 days at 70-80% confluency. Genomic DNA was extracted using a DNeasy Blood and Tissue Kit (Qiagen #69504), and the *Cytochrome C Oxidase Subunit I* (*COI*) gene amplified by PCR (primer pair 1; Supplementary Table S3.9) and sequenced to verify species identity of the cultured cells (Supplementary Figure S3.3). Cells were tested and negative for mycoplasma contamination.

3.6.2 Ectopic expression

Bos taurus ARGFX and *LEUTX* intron/exon structures were determined previously[525]. Where multiple isoforms were present, the isoform with the highest FPKM value and possessing an exon structure matching that of the human gene was used. *ARGFX* and *LEUTX* complete coding sequences with a C-terminal GGGGSGGGGS linker and 3xFLAG were synthesised by ThermoFisher GeneArt and Twist Bioscience respectively (Supplementary Figure S3.4). Sequences were cloned into plasmid pSF-CMV-Puro-COOH (OXGENE #OG3422) under the control of a CMV promoter using XhoI (Thermo

Scientific #FD0695) and NotI (Thermo Scientific #FD0593) (primer pair 2). An *ARGFX* mutant containing the 13 bp deletion found in the ARS-UCD1.2 reference genome was produced using a Q5® Site-Directed Mutagenesis Kit (New England BioLabs #E0554S). The deletion was induced with primer pair 3, then primer pair 4 used to replace the premature stop codon caused by the frameshift with the GGGGSGGGGS linker and in-frame 3xFLAG. Constructs were validated by Sanger sequencing (primer pair 5).

Cells were transfected with either the wild type bovine *ARGFX* gene, *LEUTX* gene, mutant *ARGFX* gene or empty pSF-CMV-Puro-COOH vector as a control. For transfection, 1 million cells (passage <6) were resuspended in 98 µL Opti-MEM (Gibco #11058021), combined with 2 µL of 5 µg/µL endotoxin-free plasmid, electroporated with a 2 mm gap using a NEPA21 Super Electroporator (NEPAGENE) and seeded in 2 mL BFF medium. Medium was changed after 24 h, and puromycin (Sigma Aldrich #P8833) added at 4 µg/mL to select for transfected cells. After 72 h, RNA was extracted using an RNeasy Plus Micro kit (Qiagen #74034). RNA purity was measured using a Nanodrop ND-1000 spectrophotometer and RNA integrity checked using an Agilent 2100 Bioanalyzer.

For one sample of each construct, immunocytochemistry was undertaken to check for protein expression using an antibody to 3xFLAG, as described by Maeso et al. (2016)[90] with the following differences: primary antibody = monoclonal mouse anti-FLAG M2 antibody (Sigma Aldrich #F1804-50UG) 1:100, 4 h incubation time; secondary antibody = goat anti-mouse IgG (H+L) superclonal recombinant secondary antibody with Alexa Fluor 488 (Invitrogen #A28175) 1:1000, 1 h incubation time. Cells were mounted using SlowFade™ Gold Antifade Mountant with DAPI (Invitrogen #S36938) to stain DNA and visualised with a Zeiss Axioskop 2 Plus fluorescence microscope and Zeiss Axiocam 202 mono camera. Images were processed using ImageJ.

3.6.3 Analysis of RNA-seq data

RNA Sequencing was performed on four biological replicates for each plasmid using the Illumina NovaSeq 6000 Sequencing System (Novogene), giving between 77.7 million and 120.9 million paired-end 150 bp reads per sample. Reads were filtered to remove those containing adapters, those with N > 10% and those of low quality (Qscore < 5). Quality control was performed using FastQC version 0.11.8[404] and MultiQC version 1.8[586] and Trimmomatic version 0.39 used to trim reads[587]. The pseudoaligner kallisto version 0.46.1[407] was used to align reads to the *B. taurus* ARS-UCD1.2 transcriptome and quantify abundances; *B. taurus* ETCHbox sequences are not annotated and were therefore added to the transcriptome before pseudoalignment. STAR version 2.7.9a[445], StringTie version 2.1.7[446] and the Integrative Genomics Viewer[449] were used to align, assemble and visualise reads of *ARGFX* and *ARGFX* mutant samples to verify that the correct version was expressed. R version 4.1.0 was used for data analysis (R Core Team, 2020). Tximport version 1.20.0 was used to create count tables[588], and DESeq2 version 1.32.0 used for differential expression analysis[411]. The false discovery rate method (FDR = 0.05) was used to correct for multiple testing. Genes with an adjusted *p*-value < 0.05, fold change > 1.5 and mean TPM (transcripts per million) > 2 were considered differentially expressed. One of the four *ARGFX* mutant samples was excluded as it was an outlier on PCA plots, clustering with *ARGFX* WT samples (Supplementary Figure S3.5). Volcano plots were created using R package EnhancedVolcano version 1.10.0[589] and Venn diagrams with BioVenn[590]. Gene ontology (GO) analysis[591] was completed using PANTHER version 16.0[592–594] using binomial tests and an FDR correction for multiple testing (FDR = 0.05).

We tested whether there were significant overlaps between the genes differentially expressed in response to the expression of *B. taurus*, *M. musculus* and *H. sapiens* ETCHbox genes. To achieve this, we combined the bovine data from this work with lists of genes reported to be DE in response to *H. sapiens ARGFX*, *DPRX*, *LEUTX* and

TPRX1 ectopic expression and mouse *Crxos*, *Oboxa1*, *Oboxa4* and *Oboxa7* ectopic expression[90,96]. Gene lists were first restricted to 1:1 orthologues, identified using OMA[595] (Supplementary Tables S3.10a, b). Fisher's exact test was then used to test for overlaps in the gene sets up- or downregulated by bovine *ARGFX* or *LEUTX* and each of either the human or mouse ETCHbox genes. Many of these tests produce significant overlaps so, to identify the 'most significant' overlaps and facilitate clearer comparisons, we calculated an overlap score consisting of the $-\log_{10}$ of the multiple comparison-corrected *p*-value. The greater the overlap score, the more the overlap in the genes' activity exceeds that expected by chance. For these tests, the Benjamini-Yekutieli[488] false discovery rate method (false discovery rate = 0.05) was used to correct for multiple testing because it does not assume independence of tests.

To gain insights into the role of bovine ETCHbox genes in the preimplantation embryo, temporal expression profile clustering was employed. To obtain a dataset for clustering, raw reads from the bovine embryonic RNA-seq dataset produced by Jiang et al. (2014)[441], covering oocyte, 2-cell, 4-cell, 8-cell, 16-cell, early morula, compact morula and blastocyst stages, were downloaded from NCBI SRA (PRJNA254699) and mapped with STAR[445], then transcripts were quantified with StringTie[446] as above. Mfuzz version 2.52.0[596] was then used to cluster bovine embryonically-expressed genes into temporal expression profiles. The fuzzy c-means algorithm of Mfuzz was selected as it is more robust to the noise present in RNA-seq data than is k-means clustering. Clusters were restricted to genes with a membership value > 0.5, and Fisher's exact test was used to test for enrichment of genes of each cluster in the sets of genes differentially expressed in response to ETCHbox expression. The Benjamini-Hochberg[597] method was used to correct for multiple testing.

Fisher's exact test with a Benjamini-Hochberg[597] adjustment for multiple testing was used to test for enrichment of 'undifferentiation' marker genes in the lists of genes

differentially expressed in response to bovine ETCHbox ectopic expression. The undifferentiation network signature consists of 266 genes[540], of which 234 had identifiable 1:1 orthologues in bovine.

3.6.4 Frequency of an *ARGFX* mutant allele

The Bovine Genome Variation Database[541] was used to assess the allele frequency and distribution of a mutant *ARGFX* allele identified in the *B. taurus* ARS-UCD1.2 reference genome. To test whether individuals can be homozygous for the mutant form of *ARGFX* or whether it is lethal, the polymerase chain reaction (PCR) (GoTaq G2 DNA polymerase, Promega #M7841) was used to characterise the *ARGFX* gene of 11 individuals of the Jersey breed, which possess a high frequency of the mutant allele. Nasal fluid was collected from *B. taurus* individuals using GenoTube Livestock swabs (ThermoFisher Scientific #9062010). To extract genomic DNA, a 5 mm^2 fragment was cut from the swab with sterile scalpels and minced, and then treated as a tissue sample for extraction using the DNeasy Blood and Tissue Kit (QIAGEN #69504). A primer set (pair 6) was designed with primers either side of the deletion, giving a 13 bp shorter amplicon for a mutant allele versus WT. To distinguish bands with a 13 bp difference in length, a 3% TBE (Tris-Borate-EDTA) gel was used. To test which form of *ARGFX* is present in the BFF cells, genomic DNA was extracted with a DNA Isolation Kit (Roche #11814770001) using the standard protocol, and the above primer sets used.

3.7 Acknowledgements and declarations

Data availability

Raw and processed sequencing datasets have been deposited to the NCBI Gene Expression Omnibus (GEO) (www.ncbi.nlm.nih.gov/geo) under accession GSE192356.

Acknowledgements

We thank David Marden (Elm Farm) and Paul Redmore (Neston Park Home Farm) for access to Jersey cows for DNA extraction. We also thank Dunbia Cardington for access to tissues to facilitate fibroblast isolation, and Aziz Aboobaker, Yichen Dai, Tim Davies, Yi-Jyun Luo, Peter Mulhair, Rodrigo Pracana, Amy Royall, Sebastian Shimeld and Sonia Trigueros for helpful discussions and advice. We appreciate the efforts of two anonymous reviewers whose comments improved the manuscript. This work was supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) (Grant Number BB/M011224/1) and an Oxford-Wolfson Marriott BBSRC Graduate Scholarship.

3.8 Supplementary figures

Figure S3.1

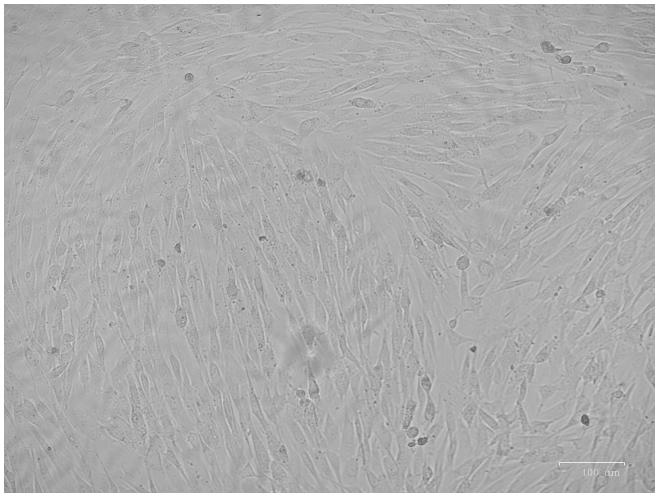


Figure S3.1 Bovine foetal fibroblasts (BFFs). Image of primary BFF cells at passage 3 post-isolation. Image captured with ZOE Fluorescent Cell Imager (Bio-Rad).

Figure S3.2

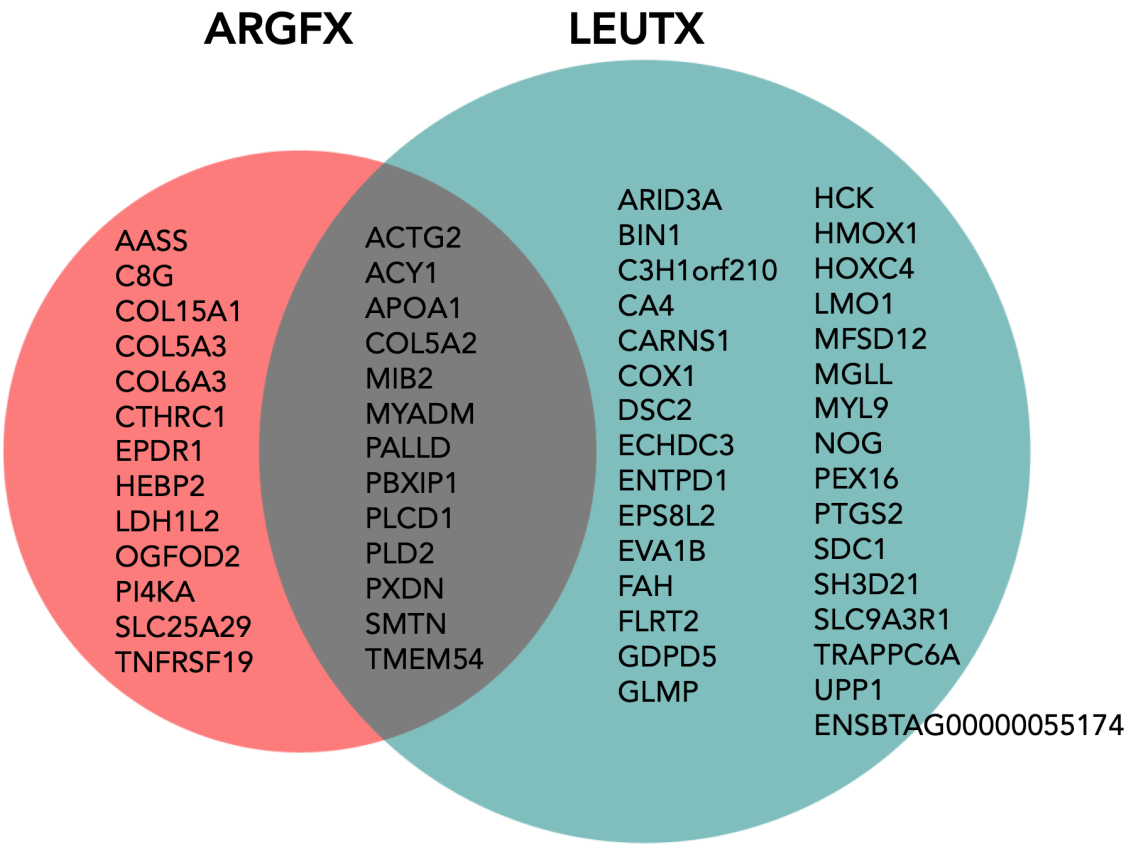


Figure S3.2 Blastocyst genes (profile 48) that are upregulated by ARGFX and LEUTX.

Figure S3.3

```
>BFF COI gene
TATGGCGATAGGCACCGGACTCTGCTCGGAGACGACCAAATCTACAACGTAGTNGTTNC
GCACACGCATTTGTAATAATCTTCTTCATAGTAATACCAATCATAATTGGAGGATTCGG
TAACTGACTTGTTCCCCTAATAATTGGTGCTCCCGATATAGCATTTCCCCGAATAAATA
ATATAAGCTTCTGACTCCTCCCTCCCTCATTCCTACTACTCCTCGCATCCTCTATAGTT
GAAGCTGGGGCAGGAACAGGCTGAACCGTGTACCCTCCCTTAGCAGGCAACCTAGCCCA
TGCAGGAGCTTCAGTAGATCTAACCATTTTCTCTTTACACTTAGCAGGAGTTTCCTCAA
TTTTAGGAGCCATCAACTTCATTACAACAATTATCAACATAAAGCCCCCCGCAATGTCA
CAATACCAAACCCCTCTGTTCGTATGATCCGTAATAATTACCGCCGTACTACTACTACT
CTCGCTCCCTGTATTAGCAGCCGGCATCACAATGCTATTAACAGACCGGAACCTAAATA
CAACCTCNTTCCGAACCGGAA
```

Figure S3.3 DNA barcode confirmation of BFF species identity. *Cytochrome c oxidase subunit I* (*COI*) gene of the cultured bovine foetal fibroblasts, sequenced by Sanger sequencing.

Figure S3.4

>LEUTX
ATGGCAGAAAAACAAAATTCAACTCGGCGATTCCGCACACGCTTTAATGGAGAACAGCT
AGGAGCACTAAGAGATGTATTTGAAAGGACCAGGTACCCACATTGTTTCCTCATAAGAA
CACTTGCTTCAACTATTCATCTTGATGCGTCAGTTATAAAGACTTGGTTTAAAAACCAA
CGTGTCAAAAGGAGGAGGGAGGAGAATCAGACTCAGCAAAATCTGTCACCAGGAGACCC
GCGCCGGGTTGTCTCAGTGAAGGAGGAAGAGATGCCCTTACCGGGCACTTCCGGAAGCA
CTCATCCCACGAGTCTCAGCCTTGCAGGTGATTCTCATCACGAGCTACCTGAGCTTTCT
TGTGCTGAGCAGTGTGAAGGGGCTGCTGCCACTCCATGCCCTTCATCCTGCAATTTCCT
GACTGCTCTAAGTCTCAGAGACACTGATCTTCCTTGGGCCTCCAGTCCTTATGACATGG
ATCAACTTATACAGTTATACGACTTACCTGGGGATGATGACCCCAGCAGTCTGGATCAG
TACCTCTTCCCAGAGTGCTCCAGCTGGGGGACTGTGGCCGGCACTGATCACCATCATGA
GGACGAACACAGCTACAATCCTGAGAAAGTTCCGGGCGGGGGGGGCTCCGGAGGAGGTG
GCAGCGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGAT
GACGATGACAAGTGA

>ARGFX
ATGGGCACTCCAAGGGAATTCATGACAGATTTCCCTTGCCAGACATCAGCACCCCTGAT
GAACAAAAGAGTCCAAGAGAATCCCCTGCCAGACCCTTTTGTCTACATGGATGATTCTA
GCACGAACTTCCAGGATGATTCAAAGATACCAGCTGGTGCCAGTCTTTTTTCATCACGG
CCTACGACGTGTGGAATCTTAACTCCCCAACCAGGGATCAAACCCGGGCCTTCGACAGT
GAAAGCAAAATGGAAGAAGAGTCATGAACGCACCTCATTCACCCACACACAGTACAAGG
AGTTGGAGGCTCTGTTTAGCTGCAACATGTTTCCAGATAAAAACCTCCAGAGAGAACTT
GCTTTAAAACTCAATCTACCAGAGTCAACAGTAAAGATTTGGTTCAGGAACCGGCGGTT
CAAAATGAAGAAGCAGCAGCGGGAGCAAGAGCAGCAATCACTAAAGCCACCAAGCCAGG
TCCTTCCAGCCAAGGATGTGCCCACAGTATCAACCAGCCCTCATTCTTTTCTCCTTGCA
ATTTCAGATTCCTATAACTCCCTCTCACCTCAGCCCTTAGACACTTTCCCCTGGGCAGG
GGACTCTATGATCATGGAGATTCCTACAAGTGATGTCCAAACGCAAGATCCTCAACTGG
AGAGGCTAGTGGCCTCAGTTCCTGCTTTGTACTCTGATGCATTTGACATCACCCAAATC
ATGGAACTGTACAGTGTTCCTGATGAGGATGACATCACCAACTCTTCCTTCTATTCTCT
ATATCAGTATCTCTCACCGACAAGGCCCAGTGGCGGGGGGGGGCTCCGGAGGAGGTGGCA
GCGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGAC
GATGACAAGTGA

>ARGFX mutant
ATGGGCACTCCAAGGGAATTCATGACAGATTTCCCTTGCCAGACATCAGCACCCCTGAT
GAACAAAAGAGTCCAAGAGAATCCCCTGCCAGACCCTATGATTCTAGCACGAACTTCCA
GGATGATTCAAAGATACCAGCTGGTGCCAGTCTTTTTTCATCACGGCCTACGACGTGTG
GAATCTGGCGGGGGGGGGCTCCGGAGGAGGTGGCAGCGACTACAAAGACCATGACGGTGA
TTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAGTGA

Figure S3.4 Ectopic expression constructs. Sequences of ectopically expressed ETCHbox genes, including GGGGSGGGGS linkers (blue) and 3xFLAG (red).

Figure S3.5

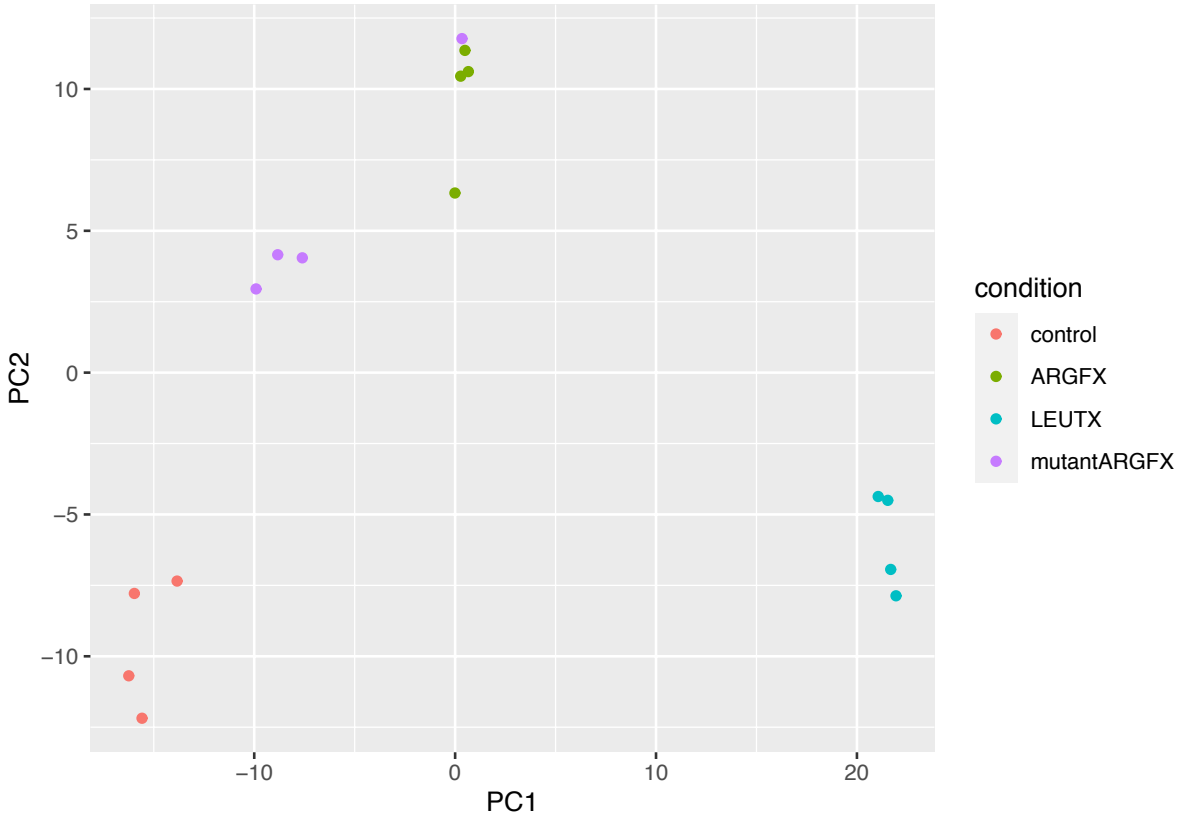


Figure S3.5 Principle Component Analysis (PCA) of RNA-seq replicates. Figure shows PCA biplot including control, *LEUTX*, WT *ARGFX* and mutant *ARGFX* samples. *LEUTX*, WT *ARGFX* and mutant ARGFX samples are well separated by PCA, except for one *ARGFX* mutant sample, which is an outlier, clustering with WT *ARGFX*. This sample was removed before differential expression analysis.

Electronic supplementary material

ESM3.1 Supplementary Tables S3.1 – S3.10.

Chapter 4: Rapid evolution of the embryonically-expressed homeobox gene *LEUTX* within primates

Thomas D. Lewin, Josephine R. Blagrove and Peter W. H. Holland

Introduction to manuscript

This chapter has been published on the preprint server *bioRxiv* (doi: 10.1101/2023.02.02.526829) and is under review at *Genome Biology and Evolution* with myself as first and corresponding author. It is reproduced here with small formatting changes as permitted as part of an integrated thesis. The authors credited on this manuscript, in order, are Thomas D. Lewin (corresponding author), Josephine R. Blagrove and Peter W. H. Holland.

Author contributions

The experiments underlying this chapter were initially conceived through discussions between Peter W. H. Holland, Thomas D. Lewin and 2021-22 MBiol student Josephine R. Blagrove, to whom TDL and PWHH were co-supervisors. Preliminary comparative genomics and preliminary experiments were performed by JRB under the co-supervision of TDL and PWHH. TDL performed vector modifications to create the empty vector control. The final comparative genomic work, ectopic expression experiment, subsequent bioinformatic analyses and writing were performed by TDL. I am grateful to the lab of Prof. Thomas Richards (Department of Biology, University of Oxford) for providing access to an Olympus CKX53 inverted fluorescence microscope for cell imaging.

4.1 Abstract

LEUTX is a homeodomain transcription factor expressed in the very early embryo with a function around embryonic genome activation. The *LEUTX* gene is found only in eutherian mammals, including humans, but unlike the majority of homeobox genes, the encoded amino acid sequence is very different between divergent mammalian species. However, whether dynamic evolution has also occurred between closely related mammalian species remains unclear. In this work, we perform a comparative genomics study of *LEUTX* within the primates, revealing dramatic evolutionary sequence change between closely related species. Positive selection has acted on sites in the LEUTX protein, including six sites within the homeodomain; this suggests that selection has driven changes in the set of downstream targets. Transfection into cell culture followed by transcriptomic analysis reveals small functional differences between human and marmoset LEUTX, suggesting rapid sequence evolution has fine-tuned the role of this homeodomain protein within the primates.

4.2 Significance

Homeodomain proteins are key regulators of animal development and are therefore often highly conserved between divergent species. However, recent work has uncovered several apparently fast-evolving homeobox genes expressed during mammalian development, one of which is *LEUTX*. Here we show that the *LEUTX* genes of primates are highly variable and that these sequence changes have resulted in small modifications to its protein function. This continued rapid divergence of an embryonic gene between closely related species is significant to the areas of evolutionary, developmental and genome biology.

4.3 Introduction

Homeobox genes are renowned for their conservation across large evolutionary timescales. Many homeodomain (HD) transcription factors (TFs) play essential roles in fundamental animal developmental processes, such as axial patterning, cellular differentiation and cell proliferation[105,106,424], and have been at the centre of the idea of the conserved genetic toolkit due to their striking similarity across widely divergent animal species[29,30]. It is therefore intriguing that in recent years an increasing number of homeobox genes have been found to be lineage-specific and rapidly-evolving, contrary to the evolutionary conservation typical of this group. This is particularly the case for the PRD-like class of homeobox genes, to which the fast-evolving and mammal-specific Cphx, Dux, Rhox, and Eutherian Totipotent Cell Homeobox (ETCHbox) genes belong. All of these gene groups have been recruited to roles in early mammalian development[56,72,75,76,90].

The ETCHbox genes duplicated from the *CRX* homeobox gene in the ancestor of eutherians and the last eutherian common ancestor possessed six group members: *ARGFX*, *DPRX*, *LEUTX*, *PARGFX*, *TPRX1* and *TPRX2*[90]. ETCHbox genes are expressed exclusively during early preimplantation development[90]; recent work has shown that they function around or immediately after embryonic genome activation (EGA) in humans and mice, with early transcriptional programs defective when they are knocked down[75,94,100,104,119]. Mouse ETCHbox genes are necessary for proper blastocyst development and hatching[115], and we recently showed that bovine ETCHbox genes have probable roles in blastocyst formation[598]. Moreover, the ETCHbox gene *TPRX1* is necessary for transforming pluripotent human embryonic stem cell cultures into totipotent 8-cell-like cells, suggesting a role in totipotency[101].

This body of work implies that ETCHbox proteins are critical regulators of developmental processes in the mammalian preimplantation embryo. The paradox is that, despite these roles, ETCHbox genes seem to be evolving rapidly. ETCHbox repertoires have undergone dramatic evolutionary changes across the eutherians, with high rates of both gene duplications and losses leading to dramatic copy number variation between species[90,95,525]. An illuminating comparison is between human, which has lost just *PARGFX* and has a single copy of the other five genes, and mouse, in which *ARGFX*, *DPRX*, *PARGFX* and *LEUTX* are all lost or pseudogenised, *TPRX1* is present in two copies and *TPRX2* in 66 copies[90,96]. Other large tandem arrays of ETCHbox genes have been found in *Oryctolagus cuniculus* (rabbit – 27 *LEUTX* copies) and *Myotis myotis* (greater mouse-eared bat – six *TPRX2* copies)[525]. However, previous work employed a broad sampling strategy, leaving open the question of whether closely related species possess different ETCHbox repertoires.

The differences between mammalian taxa are not restricted to gene duplication and loss. ETCHbox genes and their 'ancestor' *CRX* exhibit asymmetric sequence evolution: *CRX* has been conserved while ETCHbox sequences have diverged extensively between taxa, and this divergence has been driven at least in part by positive selection[90,525]. In previous work, we compared the transcriptional activity of ETCHbox genes between humans, mice and cattle, and found evidence for changes in gene function[598]; we define 'function' here as the gene sets up- and downregulated by a putative transcription factor.

Overall, previous work has shown that extensive sequence divergence and changes in ETCHbox protein function are seen between deeply diverged evolutionary lineages of eutherian mammals. We asked to what extent there have been changes between more closely related species. This will help answer whether ETCHbox homeobox genes are truly 'fast-evolving' or whether they underwent change during mammalian diversification followed by relative stasis. To address this question, we characterised the ETCHbox

gene *LEUTX* across the order Primates, with species spanning from a few million years to circa 75 million years of divergence[599–601]. We find that *LEUTX* sequences have continued to diverge at a rapid rate across primates and that positive selection has driven substitutions at key HD residues, suggesting selection for divergence of protein function. Experimental characterisation using transfection followed by RNA-seq suggests small but significant differences exist in the TF function of LEUTX between primate species.

4.4 Results

4.4.1 Duplication and divergence of *LEUTX* within primates

We identified the *LEUTX* genes in publicly-available genome sequences of 52 primate species representing all major evolutionary lineages (Figure 4.1A, B; Supplementary Figure S4.1; Supplementary Table S4.1). Although *LEUTX* has been lost in several other mammals[525], the gene is present in all of the sampled primate genomes. Of 52 species analysed, 48 have a single putatively functional *LEUTX* locus in the expected location in the genome. We find four species with duplications: (a) ten *LEUTX* loci in *Microcebus murinus* as reported previously[525]; (b) two tandem *LEUTX* loci in *Lemur catta*; (c) a divergent, intron-containing copy on a separate scaffold in *Nycticebus bengalensis*; (d) a partial gene duplication affecting exons 1 and 3 in *Hylobates pilateus.* The first three examples are all members of the Strepsirrhini, which includes the lemurs and lorises (Figure 4.2A).

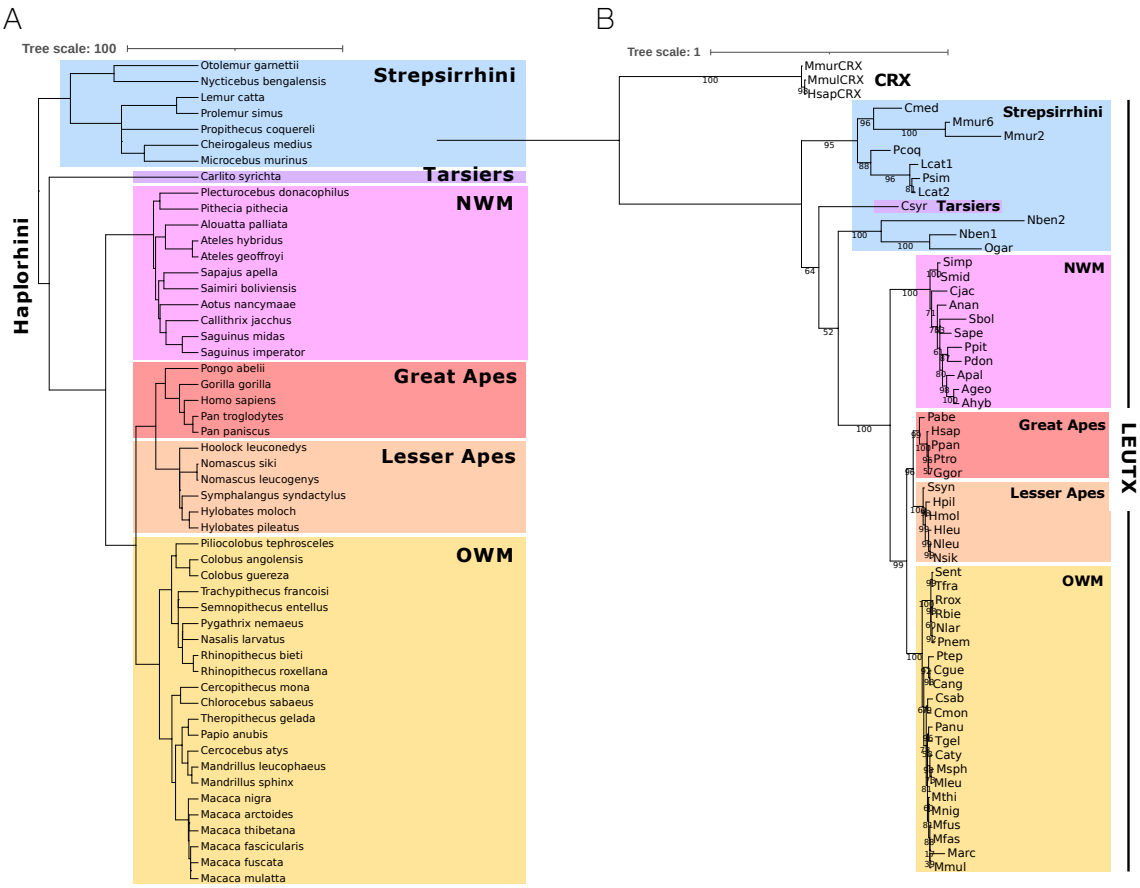


Figure 4.1 Phylogenetic analysis of primate LEUTX sequences. (A) Species tree of the 52 primates used in this analysis. Branch lengths are proportional to divergence times (million years). **(B)** Tree of primate LEUTX made with full protein-coding sequences. Great ape, lesser ape, Old World monkey and New World monkey clades are recapitulated with the expected topology. Nben and Ogar are expected to group with the lemurs (Cmed, Mmur, Pcoq, Lcat, Psim). Abbreviations: Ageo = *Ateles geoffroyi*; Ahyb = *Ateles hybridus*; Anan = *Aotus nancymaae*; Apal = *Alouatta palliata*; Cang = *Colobus angolensis*; Caty = *Cercocebus atys*; Cgue = *Colobus guereza*; Cimi = *Cebus imitator*; Cjac = *Callithrix jacchus*; Cmed = *Cheirogaleus medius*; Cmon = *Cercopithecus mona*; Csab = *Chlorocebus sabaeus*; Csyr = *Carlito syrichta*; Ggor = *Gorilla gorilla*, Hleu = *Hoolock leuconedys*; Hmol = *Hylobates moloch*; Hpil = *Hylobates pileatus*; Hsap = *Homo sapiens*; Lcat = *Lemur catta*; Marc = *Macaca arctoides*; Mfasc = *Macaca fascicularis*; Mfus = *Macaca fuscata*; Mleu = *Mandrillus leucophaeus*; Mmul = *Macaca mulatta*; Mmur = *Microcebus murinus*; Mnig = *Macaca nigra*; Msph = *Mandrillus sphinx*; Mthi = *Macaca thibetana*; Nben = *Nycticebus bengalensis*; Nlar = *Nasalis larvatus*; Nleu = *Nomascus leucogenys*; Nsik = *Nomascus siki*; NWM = New World Monkeys; Ogar = *Otolemur garnettii*; OWM = Old World Monkeys; Pabe = *Pongo abelii*; Panu = *Papio anubis*; Pcoq = *Propithecus coquereli*; Pdon = *Plecturocebus donacophilus*; Pnem = *Pygathrix nemaeus*; Ppan = *Pan paniscus*; Ppit = *Pithecia pithecia*; Psin = *Prolemur simus*; Ptep = *Piliocolobus tephrosceles*; Ptro = *Pan troglodytes*; Rbie = *Rhinopithecus bieti*, Rrox = *Rhinopithecus roxellana*; Sape = *Sapajus apella*; Sbol = *Saimiri boliviensis*; Sent = *Semnopithecus entellus*; Simp = *Saguinus imperator*; Smid = *Saguinus midas*; Ssyn = *Symphalangus syndactylus*; Tfra = *Trachypithecus francoisi*; Tgel = *Theropithecus gelada.*

We asked whether LEUTX protein sequences are fast-evolving within the primates. Using all vs. all pairwise comparisons of sequence identity, we find that primate LEUTX HDs show extensive divergence, increasing gradually with phylogenetic distance (Figure 4.2B; Supplementary Table S4.2). The two most different LEUTX HDs (*Callithrix jacchus and M. murinus*) share just 35% sequence identity. Indeed, across the full coding sequence, only 12% (23/198) of amino acid sites are invariable between all sampled primates. Coding sequences are most variable within the Strepsirrhini (Figure 4.1B; Supplementary Table S4.2). This contrasts markedly with *CRX*, from which *LEUTX* originated by gene duplication, which is highly conserved across species as is typical for homeobox genes (Figure 4.2B). Across a sample of 20 species representing all major evolutionary lineages (Supplementary Figure S4.2), 19 of the CRX HDs are identical, while *Pan troglodytes* has one substitution (A18T). Overall, we show that *LEUTX* protein-coding sequences have undergone rapid evolution within the primate lineage, including at sites within the HD.

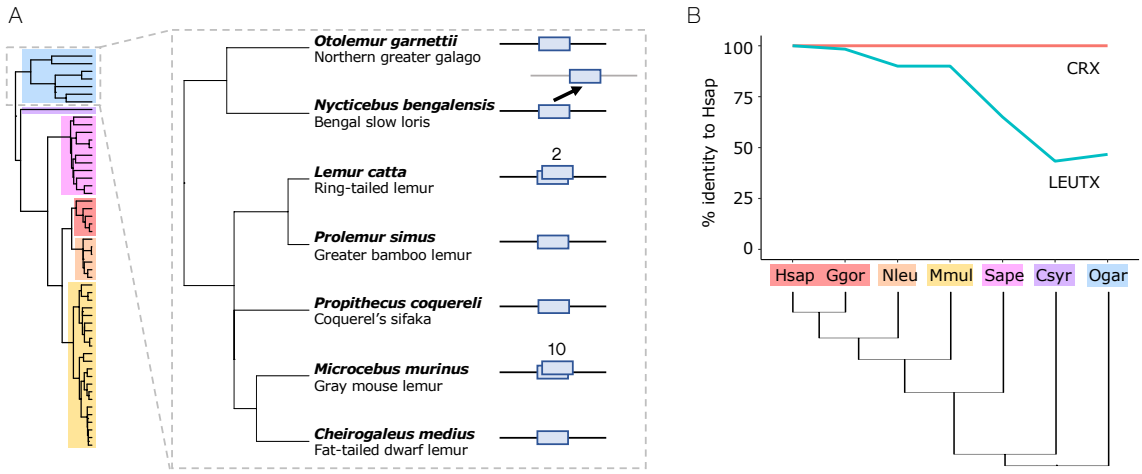


Figure 4.2 Evolution of LEUTX within primates. (A) *LEUTX* copy number within strepsirrhine primates. *Nycticebus bengalensis* has a divergent second *LEUTX* copy on a separate scaffold. *Lemur catta* has two *LEUTX* tandem duplicates, and *Microcebus murinus* has 10 *LEUTX* loci. **(B)** Divergence in primate LEUTX HDs. Plot shows percent identity of representative species' LEUTX HDs to that of human. CRX is shown for reference. Abbreviations as in Figure 4.1.

The primate CRX sequences show no variation in gene structure: the start and stop codons and intron/exon boundaries are conserved, and there are no indels. In contrast, of the 52 LEUTX sequences analysed, there are six different predicted start codon positions and seven different stop codon positions. For example, there are different predicted start codons in Old World monkeys/apes (x2); New World monkeys (x2); tarsiers; and lemurs. Additionally, we uncover indels at four separate locations (Supplementary Figure S4.3). Overall, within primates, we observe *LEUTX* duplication, rapid sequence evolution and significant changes to gene structure, but no cases of *LEUTX* gene loss.

Within the genus *Macaca*, we were able to test the extent of variation between very closely related species. Among six species, we find two have identical deduced LEUTX proteins (*M. fascicularis* and *M. fuscata*), two differ from this reference by one substitution (*M. thibetana* and *M. nigra*; both H101R); and one species has a different substitution (*M. mulatta*; P92S) (Supplementary Figure S4.4). However, *M. arctoides* has 11 amino acid differences: six of these due to a frameshift-causing indel 16 residues from the end of exon 3, changing the frame of the last seven amino acids of the protein and creating a premature stop codon. We find there is more difference between LEUTX protein sequences within the *Macaca* genus than there is between CRX protein sequences across the entire primate order.

4.4.2 Evolution of functional motifs

We tested whether positive selection has been a driver of LEUTX sequence divergence. We detected evidence for episodic diversifying selection within the primate lineage using the branch-site model of BUSTED[489] (likelihood ratio test [LRT] p-value = 9.322×10^{-7}). Analysis using MEME[493] indicated that 27 residues within the protein have been under positive selection at some point in the primate phylogeny, six of which lie within the HD

(Figure 4.3A, Supplementary Figure S4.3). This suggests that positive selection has played a role in the divergence of LEUTX proteins.

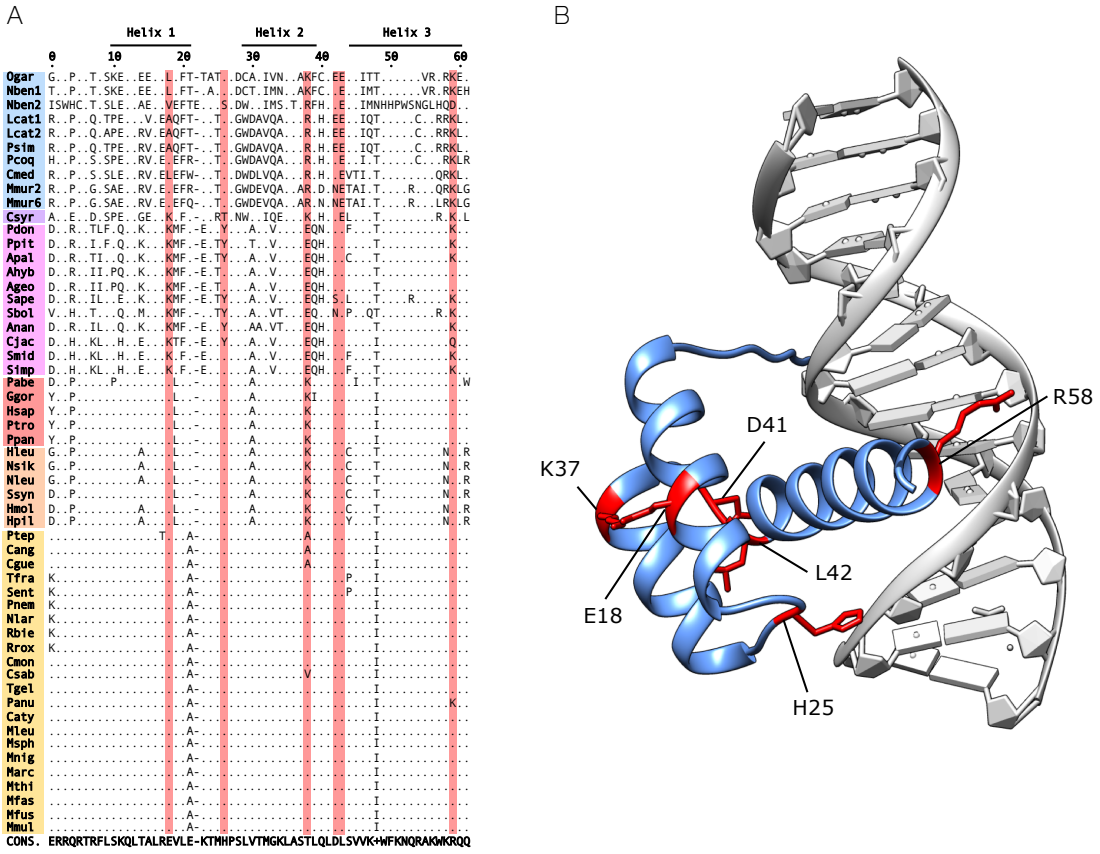


Figure 4.3 Positive selection in LEUTX homeodomains. (A) Primate LEUTX homeodomains. Only residues divergent from the consensus are shown; consensus sequence is below alignment. Residues under positive selection are highlighted in red. Abbreviations as in Figure 4.1. **(B)** Structure of human LEUTX homeodomain (blue) in complex with DNA (grey). Residues at which positive selection was detected in primates are shown in red; side chains shown only for these residues.

One of the residues inferred to have changed due to positive selection is HD residue 58, found within the critical 'recognition helix' (helix 3), which the structural modelling approach of Katayama et al. (2018)[95] identified as a specificity-determining residue in LEUTX. In almost all Old World monkeys and apes (including human) this residue is R58, while the majority of New World monkeys and prosimians (strepsirrhines plus tarsiers) have K58. *C. jacchus* is notable for its unusual Q58 residue. Comparative structural modelling suggests that residue 58 contacts the major groove of the DNA double helix (Figure 4.3B). This implies that within primates there has been selection for

divergent specificity of LEUTX DNA-binding properties. Modelling also suggests that the side chains of residues under positive selection in HD helices 1 and 2, at positions 18 and 37 respectively, are in close proximity. Their opposite charges and HD position[602] suggest the formation of salt bridges between these residues, implying selection for possible changes in the structure or stabilisation of the HD. Positively selected residue H25 contacts the DNA sugar-phosphate backbone.

Changes to other specificity-determining residues defined by Katayama et al. (2018)[95] have also occurred, but are not confirmed as under positive selection with the current dataset. First, A54 to V54 in *O. garnettii* and *N. bengalensis* (galago and loris). Second, position 47 has 'flipped' between I47 and T47 several times: T47 is seen in prosimians and New World monkeys, changing to I47 in *C. jacchus*; I47 is also seen in Old World monkeys, but changed to T47 in the ancestor of apes, again reverting to I47 in *Homo*, *Pan* and *Gorilla*. This complex evolutionary history suggests lability in this part of the LEUTX protein, consistent with previous work, which found that this site has minimal functional influence alone but may undergo compensatory substitutions in response to changes at other positions[95]. Other known specificity-determining residues (R2, R3, R5, K50, and N51) are invariant across primates (with the exception of the divergent *N. bengalensis* duplicate), and we identify pervasive purifying selection at R2, R5 and N51, along with 11 other residues within the HD (Supplementary Figure S4.3).

Katayama et. al (2018)[95] annotated a 'Leutx domain', a conserved peptide motif downstream of the HD with conservation across mammals. Within this region, the authors propose two 9-amino acid transactivation domains (9aaTADs) in every mammalian sequence analysed; 9aaTADs mediate the activation of transcription and are therefore key to TF function[603]. We find both 9aaTADs are highly conserved across 44 anthropoids analysed (New World monkeys, Old World monkeys and apes); we detect evidence for purifying selection at four residues in the first 9aaTAD and two in the

second (Supplementary Figure S4.3). As above, increased change is observed within the prosimians.

We also asked whether ubiquitination motifs in LEUTX showed evolutionary conservation across primates. Using an evolutionary screening algorithm[604], we identified three high-likelihood putative ubiquitination motifs in human LEUTX (Supplementary Figure S4.3). Each is conserved across all anthropoids, suggesting evolutionary constraint, but prosimians show notable divergence. For instance, *O. garnettii* is missing the target lysine at two out of three motifs, but putatively compensatory lysine substitutions are present within both of these motifs. Several other species are missing the target lysine in the ubiquitination motifs but evolved new lysine residues elsewhere. The conservation of ubiquitination motifs across anthropoids and the evolution of putative compensatory changes in prosimians points to functional importance, consistent with the genes' fleeting temporal expression and subsequent requirement for rapid degradation.

4.4.3 Evolution of LEUTX expression profiles

We asked whether the expression profiles of *LEUTX* in the preimplantation embryo are conserved across primates. Human *LEUTX* is expressed in a distinct temporal pattern, with expression peaking sharply at the 8-cell stage[90]. We quantified *LEUTX* expression across preimplantation development in publicly-available human*, M. mulatta* (Old World monkey) and *C. jacchus* (New World monkey) RNA-seq datasets and found strong conservation of 8-cell stage-specific expression between human and *M. mulatta* (Figure 4.4). In *C. jacchus*, *LEUTX* is expressed in a more protracted pulse comprising both the 4-cell and 8-cell stages. Thus, *LEUTX* expression profiles can vary but remain constrained within the limits of embryonic genome activation and morula formation.

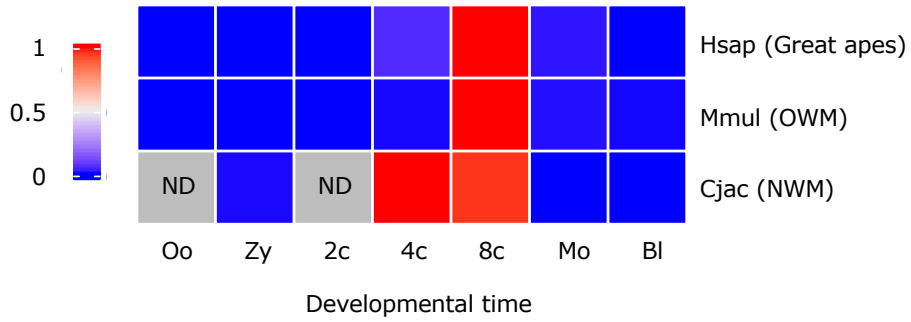


Figure 4.4 LEUTX expression in primate preimplantation embryos. Heatmap of scaled *LEUTX* expression in human (Hsap), rhesus macaque (Mmul) and common marmoset (Cjac) preimplantation embryos. Grey squares indicate no data available. Abbreviations: 2c = two-cell; 4c = four-cell; 8c = 8-cell; Bl = blastocyst; Cjac = *Callithrix jacchus*; Hsap = *Homo sapiens*; Mmul = *Macaca mulatta*; Mo = morula; ND = no data; NWM = New World monkeys; Oo = oocyte; OWM = Old World monkeys; Zy = zygote.

4.4.4 Evolutionary divergence of LEUTX downstream targets

We hypothesised that the selection-driven sequence divergence observed between primates has caused divergence of LEUTX protein functions. We used transcriptome analysis after transfection into cultured cells to test this, targeting *Homo sapiens* (representing great apes) and the common marmoset *C. jacchus* (New World monkeys) for experimental comparison. The *C. jacchus* LEUTX HD has 73% sequence identity to human, including substitutions at four sites within the HD at which we identified positive selection, one of which is the specificity-determining residue 58 (Figure 4.5A). *LEUTX* gene sequences of *H. sapiens* and *C. jacchus*, each with a C-terminal V5 tag, were cloned into a constitutive mammalian expression vector and transfected into human dermal fibroblasts (HDFs). Previous work has shown that expression of ETCHbox genes in a cell culture setting, including in fibroblasts, elicits changes to the expression of embryonic genes[75,90,96,104,525]. Immunocytochemistry confirmed protein expression and nuclear localisation of the homeodomain TF in both human and marmoset-transfected samples (Figure 4.5B).

RNA-sequencing was performed on three replicates for human and marmoset *LEUTX*, and gene expression was then quantified with Kallisto[407] (Supplementary Table S4.3).

Human (mean transcripts per million [TPM] = 3797) and marmoset (mean TPM = 4514) *LEUTX* genes were successfully expressed in the expected samples (Figure 4.5C). Differential expression analysis was performed to identify differences in the downstream genes responding to human versus marmoset *LEUTX*. We found that expression of human and marmoset *LEUTX* elicited small but notable differences in the transcriptomic response within the transfected cells: 68 genes were more highly expressed in the marmoset-transfected samples, and 44 more highly expressed in the human-transfected samples (Figure 4.5D; Supplementary Tables S4.4; S4.5). Previous work found expression of human LEUTX to downregulate 754 and upregulate 481 genes[90]; this suggests that approximately 9% of the transcriptomic response to human LEUTX is different when marmoset LEUTX is expressed.

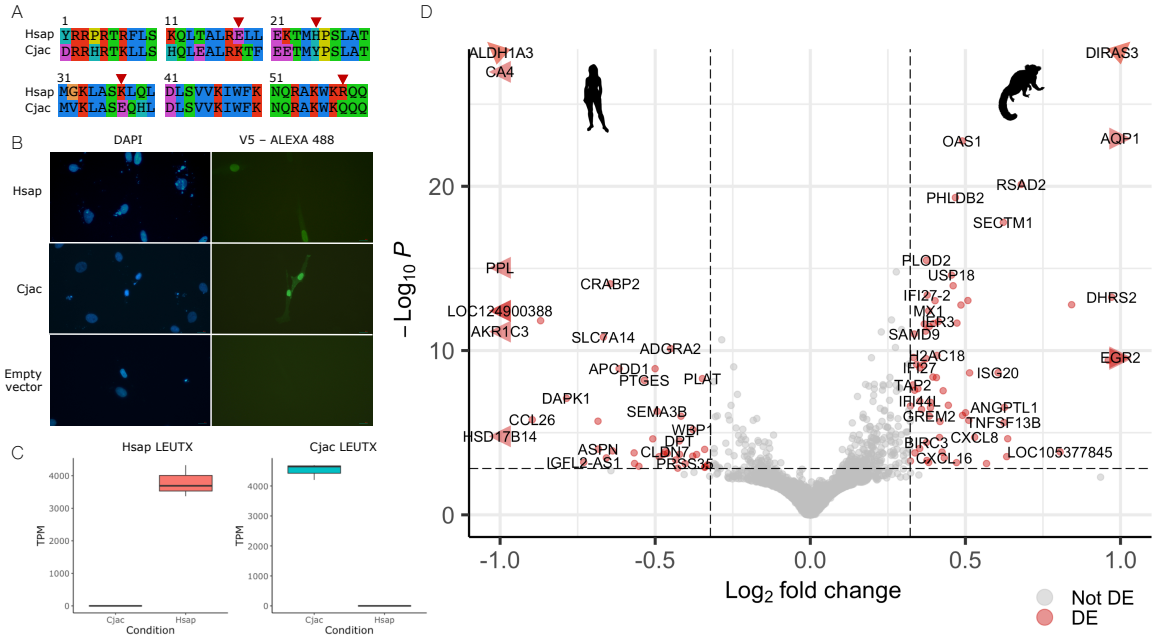


Figure 4.5 Ectopic expression of human and marmoset *LEUTX* genes. (A) Alignment of LEUTX HD sequences of the species used in ectopic expression experiments. Red arrowheads mark residues under positive selection with a substitution between human and marmoset. **(B)** Immunocytochemistry stain of fibroblasts transfected with human and marmoset *LEUTX* genes. DAPI stains DNA blue in cell nuclei. Expression of V5-tagged LEUTX proteins is detected with anti-V5 primary antibodies and Alexa Fluor 488 (green fluorescence)-labelled secondary antibodies. Empty vector transfections showed no green fluorescence. **(C)** Expression of transfected *LEUTX* genes in cell culture samples. **(D)** Transcriptional response to expression of

human and marmoset *LEUTX* genes. Points to the left of centre are genes more highly expressed in response to human *LEUTX* expression than marmoset; points to the right are genes more highly expressed in response to marmoset *LEUTX* expression. DE genes (adjusted $P < 0.05$ and fold change > 1.25) are labelled with gene IDs and shown in red. Abbreviations: Cjac = *Callithrix jacchus*; DE = differentially expressed; Hsap = *Homo sapiens*; TPM = transcripts per million.

We sought to understand the significance of these transcriptional differences. We find that of the 68 genes more highly expressed in the marmoset treatment compared to human treatment, 33 were previously shown to be downregulated by human LEUTX[90] (Figure 4.6A). This suggests that some genes downregulated by human LEUTX are not downregulated (or significantly less so) by marmoset LEUTX, revealing a change in transcription factor function. We performed biological process gene ontology (GO) analysis on these 68 DE genes: all of the top 20 GO terms without exception relate to the response to external biotic stimuli (Figure 4.6B; Supplementary Table S4.6). These terms do not appear in the gene set more highly expressed in response to human *LEUTX* than marmoset (Supplementary Table S4.7).

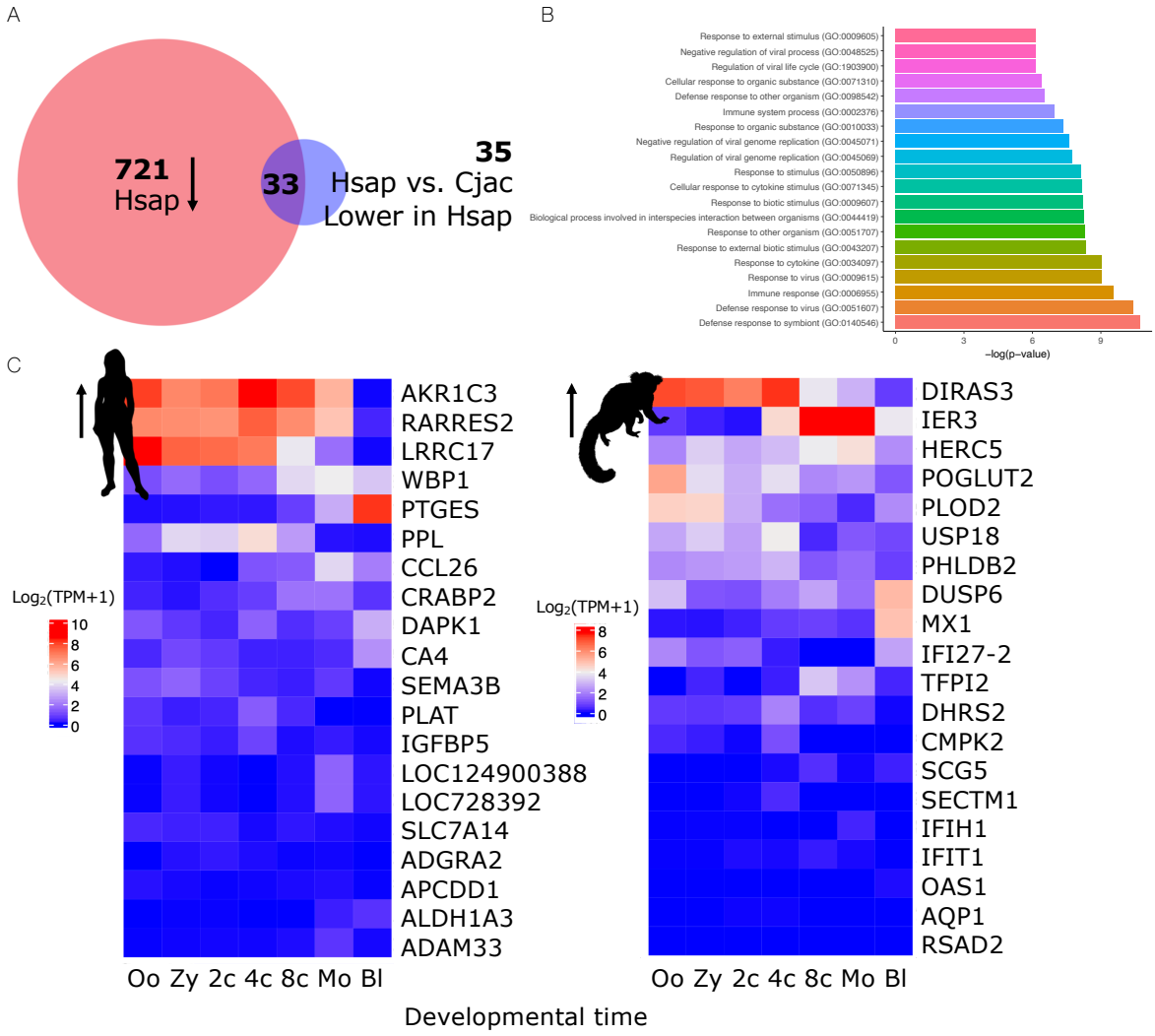


Figure 4.6 Differing transcriptional responses to human and marmoset LEUTX expression. (A) Of the 68 genes more highly expressed in the marmoset LEUTX treatment compared to human LEUTX, 33 are known to be downregulated by human LEUTX. This suggests that they are not (or significantly less) downregulated by marmoset LEUTX. **(B)** Top 20 GO terms enriched in the gene set upregulated in response to marmoset LEUTX compared to human LEUTX. **(C)** Expression of DE genes in the human preimplantation embryo. Left heatmap shows top 20 genes upregulated in response to human LEUTX compared to marmoset, right heatmap shows top 20 genes upregulated in response to marmoset LEUTX compared to human. Abbreviations: 2c = two-cell; 4c = four-cell; 8c = 8-cell; Bl = blastocyst; Cjac = *Callithrix jacchus*; Hsap = *Homo sapiens*; Mo = morula; Oo = oocyte; TPM = transcripts per million; Zy = zygote.

We also explored the genes with the strongest difference in response to ectopic expression of human and marmoset *LEUTX*. This allowed us to test whether the genes identified by the ectopic expression approach are realistic embryonic targets of *LEUTX*. We profiled the temporal expression of the top 20 DE genes using published transcriptomic data[605] spanning human preimplantation development (Figure 4.6C;

Supplementary Table S4.8). Ten of the top 20 genes upregulated in the human treatment compared to marmoset are expressed (TPM > 2 in at least one embryonic stage) during preimplantation development. Similarly, 12 of the 20 genes most differentially upregulated in response to marmoset LEUTX are embryonic genes. In addition, several of the most highly expressed DE genes (*AKR1C3*, *RARRES2*, *LRRC17*, *DIRAS3*) are strongly downregulated around the 8-cell to early morula stage, consistent with the timing of *LEUTX* expression. This suggests that the downstream targets differentially affected by marmoset and human LEUTX are realistic in vivo targets of this homeodomain protein.

Overall, there are clear, significant but relatively minor differences in the downstream targets of human and marmoset LEUTX, suggesting that the evolution of *LEUTX* sequences within the primate lineage has served to subtly modify the proteins' transcription factor function rather than elicit dramatic shifts in target gene sets.

4.5 Discussion

Fast-evolving homeobox genes may have received less attention than their highly conserved counterparts, but it is becoming increasingly clear that they play important roles in early embryonic development in mammals[37,69,72,75,90,606]. One group with known roles in preimplantation development is ETCHbox, a set of genes in which copy number, protein-coding sequence, and protein functions have been shown to vary greatly between mammalian taxa[90,96,525]. In this work, we characterised one of the ETCHbox genes within one taxonomic order, the primates, examining copy number, amino acid sequence evolution and divergence of protein function.

Comparative genomic analysis revealed that rapid evolution of the LEUTX protein-coding sequence has occurred to a remarkable extent within the primate lineage. While

the CRX HD has remained almost completely unchanged, LEUTX has undergone divergence between primate clades, resulting in an amino acid sequence identity of only 35% between the two most divergent LEUTX HDs in our study, and an average of 70% across all sampled HDs. Positive selection acting on *LEUTX* sequences played an important role in this divergence, acting on key residues across the protein, including six within the HD. The most notable target of positive selection is residue 58, known to be a factor in determining the specificity of LEUTX proteins[95], suggesting there has been selection for changes in protein targets.

Structural modelling revealed that targets of positive selection at HD positions 18 and 37 are positioned on the exterior of their respective helices. A network of salt bridges is known to form between the surfaces of helices 1 and 2 to stabilise HD structure[602]. Salt bridges are bonds between oppositely charged glutamic acid or aspartic acid (negatively charged) and arginine or lysine (positively charged) residues which contribute to protein structure, stability and specificity[607,608]. In the human LEUTX protein, the residues at positions 18 and 37 are glutamic acid and lysine, respectively; this suggests selection for modifications to salt bridge formation has occurred within primates. Previous work sampling mammals more widely also found positive selection at these residues[525], suggesting that they have been consistent targets for selection across the Eutheria.

LEUTX is not lost or pseudogenised in any of the sampled primate species, implying selection for its retention. Although a small number of duplications are observed, these are almost entirely limited to the Strepsirrhini. The relative stability of *LEUTX* copy number within primates is a notable contrast to the situation across mammals more widely, in which this gene has been lost on at least four independent occasions and has duplicated in multiple species[525].

It is enlightening to compare the scenario of *LEUTX* with that of the *Reproductive Homeobox* (*RHOX*) and *Double Homeobox* (*DUX*) gene families. Both *RHOX* and *DUX* families are PRD-like genes that are mammal-specific, expressed during early development and have rapidly-evolving sequences[56,62,72,609]. Like *LEUTX*, the *RHOXF2* protein-coding sequence has diverged rapidly between primates, and copy number variation facilitated by nearby endogenous retroviral sequences also exists between closely related species, such as the presence of two copies in human and six in chimpanzee[69]. From *DUX* genes, we learn that the presence of rapid evolutionary change does not indicate a lack of functional importance, as mouse Dux and its human orthologue DUX4 are both central to EGA despite minimal sequence conservation[57,58,60–64]. The parallels between these three gene families support the idea that selection pressures are acting to drive the evolutionary divergence of groups of homeobox genes with key roles in preimplantation development.

Bioinformatic analyses can reveal evolutionary constraint and the action of positive selection, but do not alone reveal the functional consequences of these changes. Using ectopic expression in primary cells, we compared the downstream actions of human *LEUTX* to the orthologous gene in the common marmoset *Callithrix jacchus*. Differential expression analysis revealed that expression of human and marmoset LEUTX proteins elicits small but notable differences in transcriptomic response. While this stands in stark contrast to the striking differences observed in the function of ARGFX when it was compared across a larger phylogenetic distance between human and cattle[598], it suggests that positive selection has driven minor but detectable changes in LEUTX target specificity between primate species.

What explains this divergence of protein function? LEUTX is a transcription factor activated at EGA with expression at a critical point of mammalian embryonic development[104]. At a molecular level, the gene regulatory networks (GRNs) underlying

early preimplantation development at the time of and immediately following *LEUTX* expression are largely similar across primates but do exhibit small differences[610–612]. For instance, expression of factors forming the core pluripotency network of the epiblast (NANOG, POU5F1 and SOX2) is conserved between human and marmoset, but further epiblast-specific factors such as CREB3L1, HEY2, INSR and VENTX are species-specific[88]. The relatively minor differences in LEUTX function between human and marmoset are consistent with the small-scale divergence of the GRNs coordinating preimplantation development; this suggests that positive selection on LEUTX proteins is fine-tuning their roles, changing targets at the periphery of largely conserved GRNs rather than initiating whole-scale changes to the core factors. The observed small differences in timing of LEUTX expression, which is highly specific to the 8-cell stage in humans but expressed in both 4-cell and 8-cell blastomeres in the marmoset, also support the conclusion that rapid sequence evolution has driven small functional adjustments within the primate order. However, such adjustments should not be disregarded as superficial; early development in human and marmoset do indeed entail notable differences, including the duration of preimplantation development and the manner of implantation[88,613,614].

4.6 Conclusions

LEUTX is a fast-evolving homeobox gene recruited to a role in EGA in the early mammalian embryo. We characterised the *LEUTX* loci of all available chromosome-level primate genome assemblies, revealing dramatic divergence of protein-coding sequences but limited copy number variation. This divergence has been driven at least in part by positive selection, and six residues in the LEUTX HD were identified as targets of selection within the primate lineage. Ectopic expression experiments suggest that evolutionary sequence change has led to a small divergence in LEUTX function between primate species.

4.7 Materials and methods

4.7.1 Comparative genomics

All reference assemblies of primates with a scaffold N50 of at least 1 Mb were downloaded from NCBI Genome (www.ncbi.nlm.nih.gov/genome/), with selected other species added to improve taxon representation (Supplementary Table S4.1). *LEUTX* genes were identified using blastn and tblastn searches and synteny; gene trees and reciprocal blast searches were used to confirm gene identities. The full human *LEUTX* sequence determined from transcriptome data[90] was used as the basis for inferring gene structures. Genes with a complete HD are considered putatively functional. Intronless genes (putative retrocopies) are likely to be non-functional due to the absence of regulatory elements[615] and are therefore excluded. In two species, *O. garnettii* and *N. bengalensis*, we were unable to identify the first exon of *LEUTX*.

For phylogenetics, the maximum likelihood (ML) algorithm of IQ-TREE[616] was run with 1000 bootstraps made using UFBoot2[617] and automated model selection by ModelFinder[618]. Sequence alignments were made using Clustal Omega[454] implemented in Seaview version 4.7[455]. A species tree was made using TimeTree 5, which uses a global time-calibrated tree of life synthesised from 4,075 studies[461,619]. Homeodomain sequences of PRD-class proteins were obtained from HomeoDB[620,621].

BUSTED[489] was used to test whether positive selection has acted on *LEUTX* within the primates. The mixed effects branch-site model of MEME[493] was then used to infer sites at which positive selection has acted, and the Fixed Effects Likelihood (FEL) model used to identify pervasive purifying selection (residues where purifying selection is detectable across the whole tree)[622]. Tests for selection were run with default parameters using Datamonkey[475]. Where species have a *LEUTX* duplication, only one gene was used in the tests. *Cercopithecus mona* and *Chlorocebus sabaeus* sequences were included up

to the ancestral start codon even though this has been lost; their complete HD suggests them to be functional.

The protein structure of the LEUTX homeodomain was modelled by comparative structural modelling using UCSF Chimera 1.16[498] to implement Modeller[497]. The *Drosophila melanogaster* Aristaless (Al) homeodomain (PRD-class) in complex with DNA (RCSB Protein Data Bank entry 3LNQ)[495,496] was taken as a reference. Putative ubiquitination sites were detected with ESA-UbiSite[604]. Homeodomain residues were excluded as potential sites of ubiquitination.

For expression analysis, raw RNA-seq reads from human (*Homo sapiens*; PRJNA153427)[605], rhesus macaque (*Macaca mulatta*; PRJNA401876)[623] and common marmoset (*Callithrix jacchus*; PRJEB29285) preimplantation embryos were obtained from NCBI BioProject (www.ncbi.nlm.nih.gov/bioproject/). *LEUTX* expression was quantified at each developmental stage using Kallisto version 0.48.0[407]. Heatmaps were made using ComplexHeatmap version 2.8.0[624] in R version 4.1.0[466].

4.7.2 Ectopic expression

Primary human dermal fibroblasts (HDFs) (Stemnovate #SV-HF21-17-500) were maintained in HDF medium at 37°C with 5% CO_2 and passaged at approximately 70% confluency every 3-4 days. HDF medium consists of Dulbecco's modified Eagle medium (Gibco #41965039) with 10% heat-inactivated fetal bovine serum (Gibco #10500064) and 1% penicillin-streptomycin (Gibco #15140122). Testing for mycoplasma (Sigma Aldrich #MP0035) revealed no contamination.

Codon-optimised sequences of *Homo sapiens* and *Callithrix jacchus LEUTX* with a GGGGSGGGGS linker and C-terminal V5 tag (Supplementary Figure S4.5) were synthesised by ThermoFisher GeneArt and cloned into a pcDNA3.1 mammalian

expression vector. For transfection, 65000 cells per well were seeded into 6-well plates. After 16 hours (h), medium was replaced with 2 mL antibiotic-free HDF medium. For each biological replicate, 108 µL Opti-MEM (Gibco #31985-062) was combined with 9.6 µL FUGENE 6 (Promega #E2691) and incubated for 5 minutes, then 2.4 µL of 1 µg/µL appropriate expression construct added before another 15-minute incubation. To each well, 120 µL of this mixture was added, and cells kept at 37°C with 5% CO_2. After 24 h, transfection medium was removed and replaced with 2 mL HDF medium with 800 µg/mL G418 selective antibiotic (Gibco #10131-035). At 48 h post-transfection, RNA was extracted using an RNeasy Plus Micro kit (Qiagen #74034), and integrity tested using an Agilent 2100 Bioanalyzer.

To confirm expression of full-length proteins, the immunocytochemistry protocol of Maeso et al. (2016)[90] was used with minor modifications: primary antibody (V5 tag monoclonal antibody; Invitrogen #37-7500) 1:500, 4 h incubation; secondary antibody (goat anti-mouse IgG H + L superclonal recombinant secondary antibody with Alexa Fluor 488; Invitrogen #A28175) 1:1000, 1 h incubation. Cells were incubated with DAPI (Invitrogen #S36938) to label nuclei. Results were visualised with an Olympus CKX53 inverted fluorescence microscope.

4.7.3 Analysis of RNA-seq data

Three replicates for each treatment were sequenced on the Illumina NovaSeq 6000 platform (Novogene). FastQC version 0.11.8[404] and MultiQC version 1.8[586] were used for quality control, and reads (150 bp paired-end) were subjected to filtering to remove adapter-containing reads, low-quality reads (Qscore < 5), and reads with > 10% Ns (undetermined bases), resulting in an average of 45.8 million reads per sample. Pseudoalignment to the human transcriptome from genome build GRCh38.p14 (RefSeq annotation) was performed with Kallisto version 0.48.0[407]; pseudoalignments were found for an average of 93.8% of reads. Gene-level abundance estimates were created using

tximport version 1.20.0[588], and then differential expression analysis completed in DESeq2 version 1.32.0[411] using apeglm[625] for log fold change (LFC) shrinkage. EnhancedVolcano[589] version 1.16.0 was used to create volcano plots. Genes with an adjusted *p*-value < 0.05, fold change > 1.25 and mean TPM over 2 were considered differentially expressed. To check whether differentially expressed genes represented realistic embryonic targets, raw reads from human preimplantation development (PRJNA153427)[605] were quantified with Kallisto as above[407]. Gene ontology (GO) analysis was performed using PANTHER version 17.0[626] with Fisher's exact test and a false discovery rate (FDR) correction of 0.05.

4.8 Acknowledgements and declarations

We would like to thank Peter Mulhair for productive discussions, advice and critical reading of the manuscript. This work was supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) (grant number BB/M011224/1); an Oxford-Wolfson Marriott BBSRC Graduate Scholarship; and Merton College, Oxford.

Data availability

Raw and processed sequencing datasets are available from the NCBI Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) under accession GSE224384.

Conflict of interest

The authors declare no conflicts of interest.

4.9 Supplementary figures

Figure S4.1

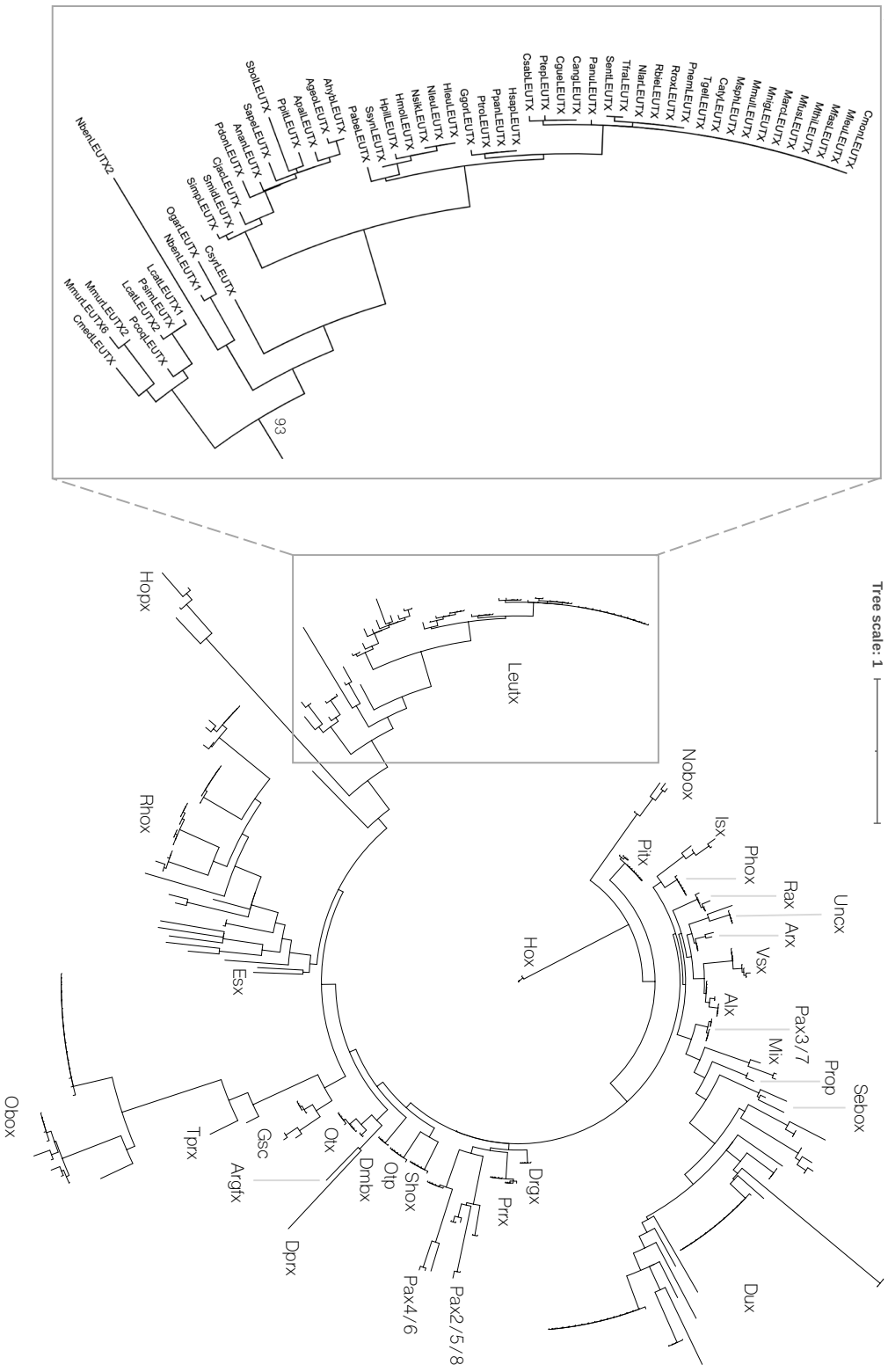


Figure S4.1 Position of LEUTX sequences within a tree of all PRD-class homeodomains. Homeodomain sequences of PRD-class proteins were obtained from HomeoDB [620,621]. Species name abbreviations as in Figure 4.1 main text.

Figure S4.2

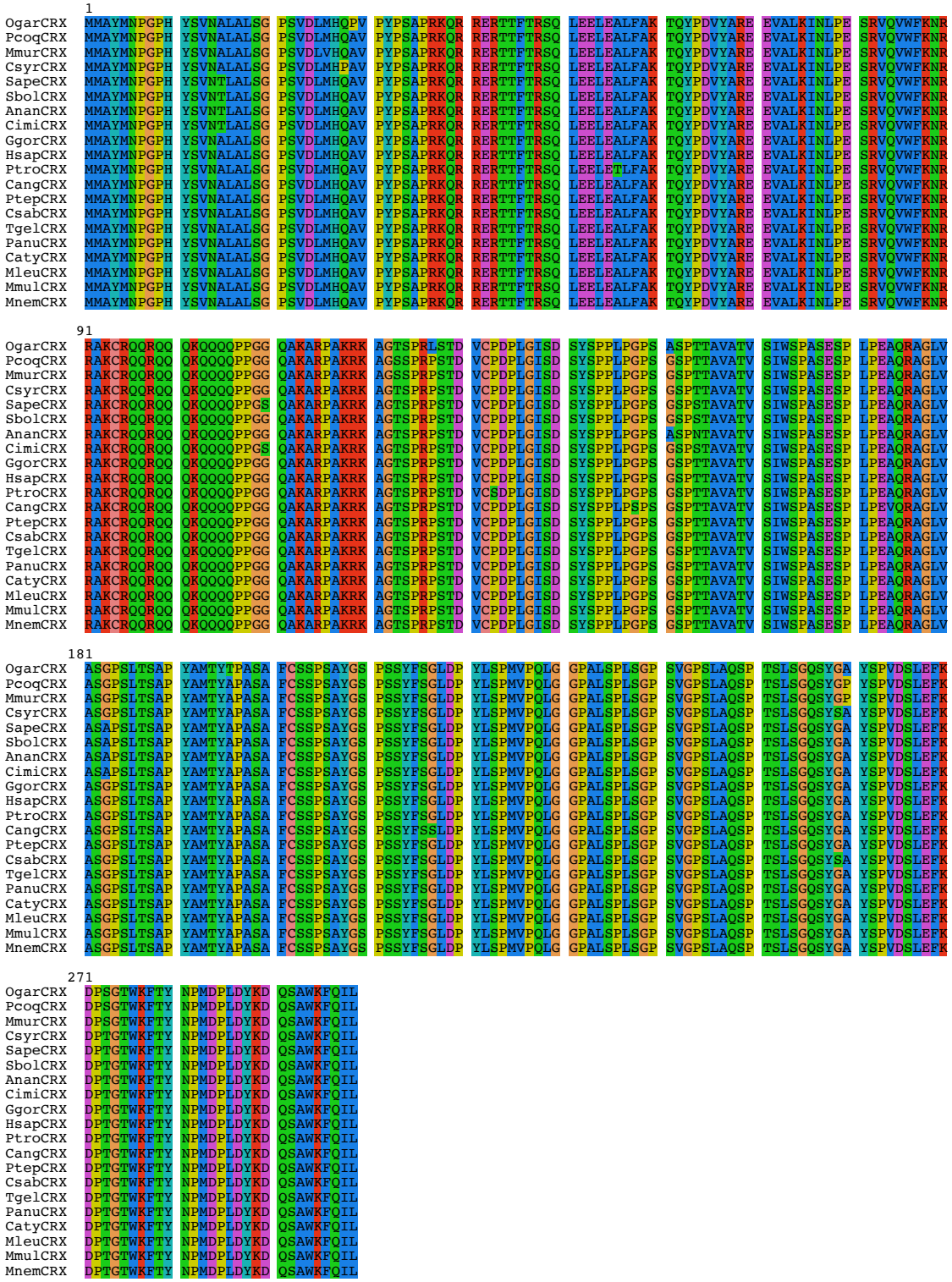


Figure S4.2 Alignment of primate CRX proteins. Species name abbreviations as in Figure 4.1 main text.

Figure S4.3

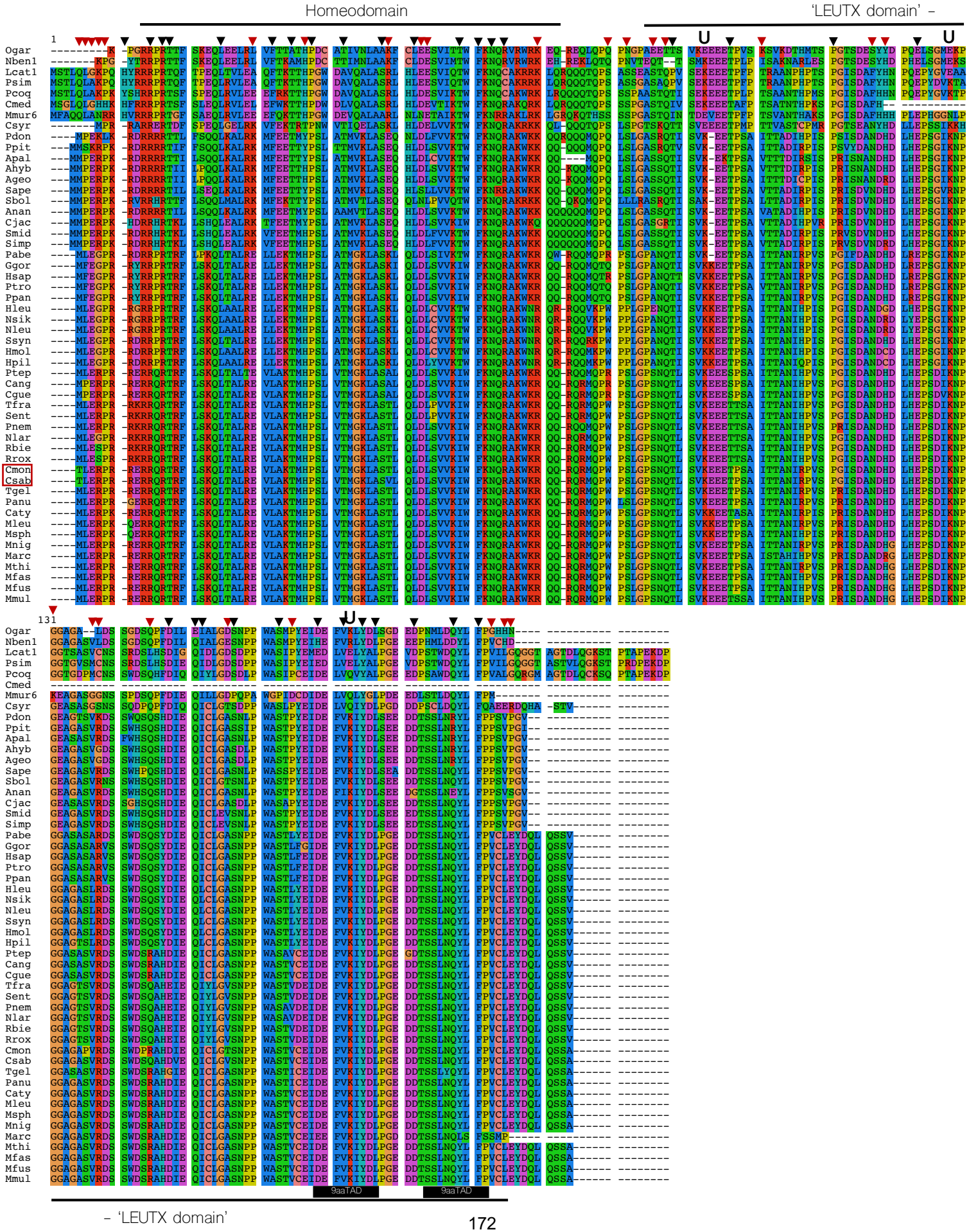


Figure S4.3 LEUTX protein sequence alignment used for tests for selection. Red arrowheads indicate residues identified as under positive selection, residues marked with a black arrowhead are under pervasive purifying selection. Human ubiquitination sites are marked with a U. The position of human 9aaTADs is marked with a labelled black bar. The Cmed sequence is shortened due to an incomplete assembly, not a premature stop codon. The red square marks Cmon and Csab, in which the New World monkey start codon has mutated to a threonine residue. Ogar and Nben lack a start codon because the first LEUTX exon could not be identified. Species name abbreviations as in Figure 4.1 main text.

Figure S4.4

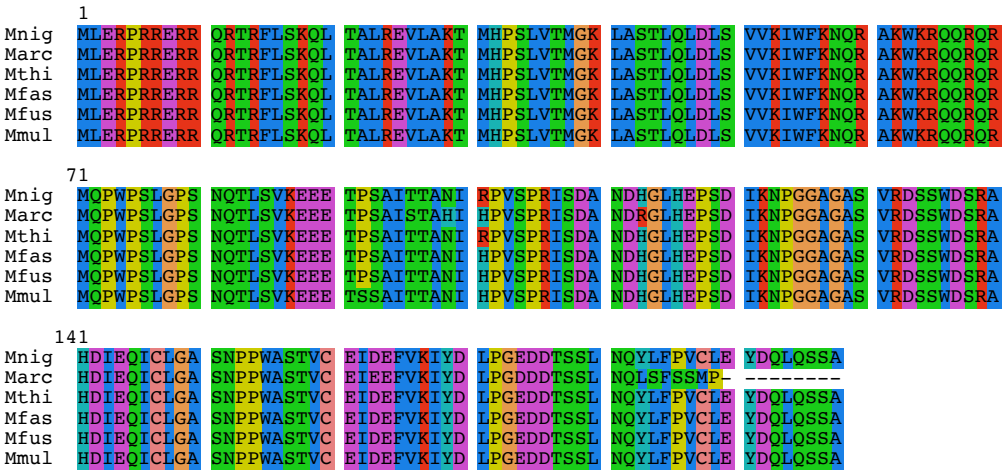


Figure S4.4 LEUTX proteins of species in the genus *Macaca*. Species abbreviations as in Figure 4.1 main text.

Figure S4.5

>Hsap_LEUTX_V5
ATGTTTGAGGGCCCCAGAAGATACAGACGGCCCAGAACCAGATTCCTGAGCAAGCAGCT
GACAGCCCTGAGAGAGCTGCTGGAAAAGACAATGCACCCCAGCCTGGCCACCATGGGAA
AGCTGGCTTCTAAACTGCAGCTGGACCTGAGCGTGGTCAAGATCTGGTTCAAGAACCAG
CGGGCCAAGTGGAAGCGGCAGCAGAGACAGCAGATGCAGACCAGACCTTCTCTGGGCCC
TGCCAATCAGACCACCAGCGTGAAGAAAGAGGAAACCCCTAGCGCCATCACCACCGCCA
ACATCAGACCTGTGTCTCCCGGCATCAGCGACGCCAACGATCACGATCTGAGAGAACCC
AGCGGCATCAAGAATCCTGGCGGAGCCTCTGCCTCTGCCAGAGTGTCATCTTGGGACAG
CCAGAGCTACGACATCGAGCAGATCTGTCTGGGCGCCAGCAATCCTCCTTGGGCCAGCA
CACTGTTCGAGATCGACGAGTTCGTGAAGATCTACGACCTGCCTGGCGAGGACGATACC
AGCAGCCTGAACCAGTATCTGTTCCCCGTGTGCCTGGAATACGATCAGCTGCAGAGTTC
TGTTGGCGGCGGAGGATCTGGCGGAGGCGGTTCTGGAAAGCCCATTCCTAATCCTCTGC
TGGGCCTCGACAGCACCTGATGATAA

>Cjac_LEUTX_V5
ATGATGCCCGAAAGACCCAAGCACGACAGACGGCACAGAACAAAGCTGCTGAGCCACCA
GCTGGAAGCCCTGAGAAAGACCTTCGAGGAAACAATGTACCCCAGCCTGGCCACCATGG
TCAAGCTGGCCTCTGAACAGCACCTGGACCTGAGCGTGGTCAAGATCTGGTTCAAGAAC
CAGCGGGCCAAGTGGAAGCAGCAGCAACAGCAACAACAGATGCAGCCCCAGCTGTCTCT
GGGCGCCTCTGGCAGAACAATCAGCGTGAAAGAAGAGACACCCAGCGCCGTGACCACCG
CCGATATTCACCCTGTGCGGCCTAGAATCAGCGACGTGAACGACCACGATCTGCACGAG
CCTAGCGGCATCAAGAATCCTGGCGAAGCCTCTGCCAGCGTGCGGGATTCTTCTGGCCA
CAGCCAGAGCCACGACATCGAGCAGATTTGTCTGGGAGCCAGCGACCTGCCTTGGGCCT
CTGCTCCTTATGAGATCGACGAGTTCGTGAAGATCTACGACCTGTCCGAAGAGGACGAC
ACCAGCAGCCTGAACCAGTACCTGTTTCCTCCAAGCGTGCCAGGCGTTGGAGGCGGAGG
ATCTGGCGGAGGCGGATCTGGAAAGCCCATTCCTAATCCTCTGCTGGGCCTCGACTCCA
CCTGATGATAA

Figure S4.5 *LEUTX* expression constructs. Sequences of ectopically expressed *LEUTX* genes, including GGGGSGGGGS linkers (blue) and V5 tag (red). Abbreviations: Cjac = *Callithrix jacchus*; Hsap = *Homo sapiens*.

Electronic supplementary material

ESM4.1 Supplementary Tables S4.1 – S4.8.

Chapter 5: ETCHbox proteins and transposable elements

Statement of author contributions

This project was conceived by discussions between Thomas D. Lewin and Peter W. H. Holland. All wet-lab experimental work, data analysis, data interpretation and writing were carried out by TDL. PWHH supervised the work. The retrotransposition assay used was developed by Prof. John V. Moran in the laboratory of Prof. Haig H. Kazazian Jr at Johns Hopkins University School of Medicine (Baltimore, Maryland, USA). The protocol used for the assay followed a published methods paper[627]. HeLa-HA cells and all plasmids for the assay (except the ETCHbox expression plasmids) were a generous gift from Dr. José L. García Pérez (Institute of Genetics and Molecular Medicine, University of Edinburgh, UK). Alexandra Bisia (Sir William Dunn School of Pathology, Oxford, UK) provided assistance with microscopy.

5.1 Abstract

A striking feature of mammalian preimplantation development is the brief surge in the expression levels of many transposable elements (TEs). The peak of this wave, occurring between the 4-cell and blastocyst stages, is conspicuous in its overlap with the time at which ETCHbox proteins are active. This chapter presents several exploratory analyses which examine whether this is simply coincidence or whether there is a link between these two occurrences. Using RNA-sequencing data, I first show that several human and bovine ETCHbox proteins downregulate a large suite of TE-repressing genes. Based on this, I hypothesise that expression of ETCHbox proteins increases the rate of TE retrotransposition, and use a cell culture retrotransposition assay to test this for one such protein, TPRX1. Preliminary data suggest that the presence of TPRX1 does indeed increase the retrotransposition rate, but additional work is required to confirm this. I also quantify TE transcripts in published RNA-sequencing data and show that human ETCHbox proteins have a negligible impact on TE expression levels; this suggests that they are not transcriptional regulators of TEs. Overall, this work suggests that ETCHbox proteins act as upstream regulators of the machinery governing TE activity but do not affect TE transcription; it opens up several avenues for further investigation, particularly with regard to the evolutionary fitness benefit of downregulating TE repressive genes.

5.2 Introduction

An important but often underestimated driver of genome evolution is the action of transposable elements (TEs). TEs are mobile DNA sequences with the ability to move from one genomic position to another (for a more complete discussion see Chapter 1 section 1.6). Archetypes of the idea of the selfish gene, in general (though with notable exceptions) it is their ability for self-replication that has led to their accumulation in genomes rather than selection on host genomes for their presence. Indeed, an astonishing 45% of both human and bovine genomes consists of TEs alone[122,628]. TEs are insertional mutagens, and their insertion at protein-coding loci and subsequent disruption of gene function can have highly damaging effects on fitness. Numerous human diseases are caused by de novo TE insertions[372]; Duchenne muscular dystrophy caused by LINE-1 insertion into the *DMD* gene is a prominent example[629].

TEs' mutagenic potential has placed strong selection on genomes to exert control over their TE complements to limit their activity and mutagenicity, and mammalian genomes like those of humans, mice and cattle possess an arsenal of mechanisms with which TEs are marshalled. One such mechanism is genomic DNA methylation of TEs, which in mice is mediated by Dnmt3a, Dnmt3b and Dnmt3L[630,631]. Small interfering RNAs (siRNAs) or piwi-interacting RNAs (piRNAs) and their interaction with P-element induced wimpy testes (PIWI) proteins also contribute to this process[357,632–634]. Furthermore, a suite of highly duplicated transcription factors (TFs) known as the Krüppel-associated box (KRAB) zinc-finger (KZNF) family mediate the transcriptional repression of a wide variety of transposons by associating with KRAB-associated protein-1 (KAP1) and recruiting a silencing complex that results in histone methylation and the formation of heterochromatin[635–639]. TEs are also regulated at a post-translational level by myriad pathways each with distinct mechanisms; these include the binding of RNA-binding proteins like hnRNPL, cytosine deamination by APOBEC3 proteins, and the action of

further cellular proteins like IFITM1, TRIM5A, SAMHD1, TREX1 and MOV10 and ZC3HAV1 (ZAP) among others, whose mechanisms include splicing and premature polyadenylation[357,375,640–653]. Overall, genomes exert authoritarian control over their TEs with sophisticated and redundant mechanisms of repression.

The heritability of germline mutations means that the protection of the germline from TE insertions is of paramount importance. It is therefore surprising that the fine-tuned process of preimplantation development in humans, mice and other mammals is accompanied by the brief activation of vast swathes of TEs, especially between the 4-cell stage and blastocyst formation[131]. This process consists of stage-specific transcription of TEs from various groups, in humans including but not limited to ERVs and SVA elements[129–131], and is caused at least in part by the remodelling of the heterochromatin landscape and the removal of repressive marks that is an essential component of preimplantation development, but also the selective pressure acting on TEs to replicate at this time[296,388,654,655].

Based on several lines of evidence, we hypothesised that ETCHbox genes may have evolved to function in germline defence by governing TE activity in the early embryo. First, it is conspicuous that the rapid upregulation of ETCHbox expression coincides with this pulse of TE activity. Second, ETCHbox proteins' TF capability makes transcriptional regulation of TEs a plausible function. Third, the emergence of ETCHbox genes at the base of the placentals coincides with a large evolutionary expansion of TEs; around one-third to one-half of each mammalian genome is comprised of TEs[656]. Fourth, their dynamic duplication and loss is consistent with the fast-evolving nature of species' TE complements and, if an arms race developed between TEs evolving to escape repression and ETCHbox genes evolving to repress them, this would provide an explanation for the genes' rapid sequence evolution. Finally, this would not be without precedent. Members of the mammalian-specific fast-evolving Rhox homeobox gene

family act to inhibit LINE-1 expression and retrotransposition during embryonic development[132].

The properties of TEs present numerous challenges for bioinformatic analyses, and especially with respect to their quantification within sequencing data. The most obvious issue is the repetitive nature of their transcripts, but further challenges derive from population polymorphism and transcript diversity[657–659]. These characteristics make the task of accurate quantification a bioinformatician's nightmare, with just two of many difficulties being extensive multi-mapping (especially an issue with short-read RNA-seq data) and an inability to distinguish autonomous TE transcription from that which is passively associated with neighbouring genes[657]. Several tools have been specifically designed to tackle these challenges; a recent comparison of tools for TE quantification in RNA-seq data[660] found SalmonTE[661] to be the most accurate, and it is that which is used here.

In this work, I explore the possibility of a link between ETCHbox proteins and TEs. I focus on human and cattle, as their ETCHbox repertoires are highly similar to each other and that of the eutherian ancestor.

5.3 Aims

1. Do ETCHbox proteins regulate genes that repress TEs?
2. Does the presence of ETCHbox proteins affect the rate of TE retrotransposition?
3. Does the presence of ETCHbox proteins affect the level of TE expression?
4. Is there an excess of TEs around ETCHbox loci?

5.4 Materials and methods

5.4.1 Identification of ETCHbox-regulated TE-repressive genes

To explore whether ETCHbox proteins regulate genes involved in TE-repressive pathways, differentially expressed (DE) gene sets from two previously published studies were utilised: 1) human ETCHbox expression in human dermal fibroblasts[90], 2) bovine ETCHbox expression in bovine foetal fibroblasts (Chapter 3)[598] (Supplementary Table S5.1). Genes with known roles in TE regulation were identified in DE gene sets. Genes that are known to affect TE expression by piRNA/epigenetic mechanisms were excluded, as there are widescale changes in these processes at the time of ETCHbox expression in the embryo so their appearance in DE gene sets cannot be directly linked to TE regulation. Lists of KZNFs were taken from Corsinotti et al. (2013)[662] (human) and Imbeault et al. (2017)[637] (cattle). Of 374 bovine KZNF genes, 219 have unique Ensembl IDs and were therefore present in our dataset. Two-sided Fisher's exact tests were used to test for over- and underrepresentations of KZNFs in the DE gene sets, and for overlaps in the KZNFs regulated by different ETCHbox genes. Gene ontology (GO) analysis was performed using the protocol described in Chapter 4 section 4.7.3.

5.4.2 Single-cell RNA-seq analysis

A single-cell RNA-sequencing (scRNA-seq) dataset covering the 8-cell to blastocyst stage of human preimplantation development[663] was used to explore whether the ETCHbox-regulated TE-repressive genes identified in fibroblast cultures represent realistic embryonic targets. Data were processed using Seurat[664,665]; a full explanation of the data analysis pipeline is presented in Chapter 6 section 6.4.1. Multiple comparison-corrected Wilcoxon rank sum tests were used to test for differential expression between clusters.

5.4.3 LINE-1 retrotransposition assay

The cultured-cell retrotransposition assay was first developed by Moran et al. in 1996 and is used to test whether exogenous proteins affect the rate of retrotransposition events occurring in a cell population[666]. The work in this chapter is based on a protocol for the assay described by Kopera et al. (2016)[627]; it uses the LINE-1 mammalian expression plasmid pJM101/L1.3 (Figure 5.1), which encodes a full-length, retrotransposition-competent LINE-1 element under the control of a CMV promoter. When this construct is transfected into cell culture, the LINE-1 element is expressed and retrotransposition events occur in a very small proportion of cells. Antibiotic resistance is used to track the occurrence of retrotransposition events: within the 3' untranslated region (UTR) of the LINE-1 element is, in reverse orientation, an *mneoI* reporter cassette containing an independent promoter controlling a neomycin resistance gene (neomycin phosphotransferase). This cassette is split by an intron in the same direction as the LINE-1 element. When expressed from the plasmid, the neomycin resistance gene is non-functional. However, should the LINE-1 element be spliced and then reverse transcribed and inserted into genomic DNA (i.e. retrotransposed), the neomycin resistance gene will be expressed in its functional form, leading to the selectable marker of neomycin resistance. In short, if a retrotransposition event occurs in a cell, that cell and its daughters will possess neomycin resistance. To test whether a gene of interest affects the rate of retrotransposition, the pJM101/L1.3 plasmid is transfected into cells with and without the gene of interest, and the number of neomycin-resistant colonies (each representing one retrotransposition event) counted.

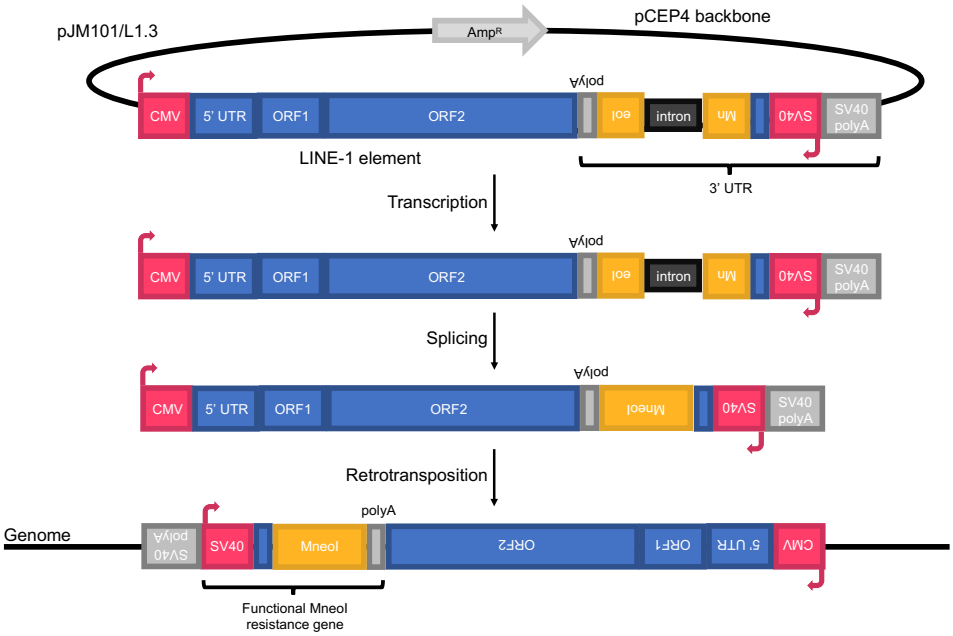


Figure 5.1 Retrotransposition assay construct. The pJM101/L1.3 plasmid contains a LINE-1 element including its 5' UTR, ORF1 and ORF2 under the control of a constitutive CMV promoter. In its 3' UTR, in reverse orientation is an *mneoI* neomycin resistance cassette split by an intron with an SV40 promoter. This transcriptional unit is transcribed from the plasmid, spliced to remove the intron, and then retrotranscribed into the genome, at which point the *mneoI* cassette is expressed in its complete form, conferring neomycin resistance to cells in which a retrotransposition event has occurred. Abbreviations: CMV = cytomegalovirus promoter; MneoI = neomycin resistance cassette; ORF = open reading frame; SV40 = simian virus 40.

HeLa-HA[650], an immortal cell line permitting both LINE-1 and Alu retrotransposition, was used for the assay. Cells were grown in HeLa-HA medium[627] and subcultured one flask to five every two days at approximately 80% confluency. For each individual replicate, 70000 cells were seeded into each well of a 6-well plate 16 hours (h) before transfection. Immediately prior to transfection, medium was replaced with 2 mL antibiotic-free HeLa-HA medium. Each transfection mixture was prepared as follows and added to the well: 100 µL Opti-MEM (Gibco #31985-062), 3 µL FUGENE 6 (Promega #E2691), 1 µL expression plasmid (1 µg/µL). The FUGENE 6 volume was increased to 5 µL if two plasmids were to be transfected. One day post-transfection, transfection medium was aspirated and replaced with full HeLa-HA medium. At three days post-transfection, G418 selective antibiotic (Gibco #10131-035) was added to the medium at 400 µg/mL and replaced daily until 15 days post-transfection. This leads to the death of all cells in which

no retrotransposition event has occurred. Surviving cells were then fixed in 4% paraformaldehyde (ChemCruz #sc-281692) for 20 minutes, stained with 0.1% crystal violet (Sigma-Aldrich #V5265) for 1 h and washed four times with sterile water. Each remaining colony of cells consists of a founding cell in which a retrotransposition event occurred and its daughters; these appear as purple foci and can be counted to quantify the number of retrotransposition events occurring. Colonies were counted using a Scan 500 colony counter (Interscience) followed by manual correction.

In this experiment, TPRX1 was selected for testing based on its consistent downregulation of retrotransposition-inhibiting genes. The production of a V5-tagged human TPRX1 expression plasmid was described previously[90]. Several key control treatments are required in addition to the treatments consisting of the pJM101/L1.3 plasmid with and without the TPRX1 expression plasmid (Table 5.1). ARRB1 does not affect the rate of retrotransposition: LINE-1 plus ARRB1-transfected samples are used as the baseline for the number of retrotransposition events[627]. APOBEC3A (A3A) is a positive control; it is known to reduce the rate of retrotransposition[667]. In addition, a possible confounding effect of the protein of interest on cellular survival must be accounted for. For instance, if the LINE-1 plus TPRX1 samples show fewer retrotransposition events than the LINE-1 plus ARRB1 samples, this could be because TPRX1 truly results in a reduction in the rate of retrotransposition or it could be because TPRX1 is cytotoxic and therefore there are fewer live cells in which a retrotransposition event can occur. To account for this, the pU6iNEO plasmid is also transfected into cells with the genes of interest; this plasmid carries the same antibiotic resistance cassette but without the need for retrotransposition. The effect of the genes of interest on retrotransposition is then corrected to account for cytotoxicity. In addition, each gene of interest is transfected alone to ensure that this does not lead to cellular survival in the absence of retrotransposition. Finally, a modified pJM101/L1.3 plasmid which carries a D702A mutation that abolishes its competence for retrotransposition is used to ensure

that the unintegrated LINE-1 plasmid does not lead to cellular survival. Transfection efficiency was quantified with an immunocytochemistry stain against the V5 tag of TPRX1, as described in Chapter 4 section 4.7.2. Each treatment was performed in triplicate (three technical replicates) on three occasions (three biological replicates) for a total of nine repeats. Calculation of retrotransposition efficiency and accounting for cellular toxicity is performed in accordance with the previously published protocol[627].

Treatment no.	Transfection	Purpose
1	L1 only	Baseline for number of L1 events
2	L1 + ARRB1	Baseline for number of L1 events with protein that does not affect rate of retrotransposition
3	L1 + A3A	Positive control for effect on retrotransposition - reduces rate
4	L1 + TPRX1	Test if TPRX1 effects rate of retrotransposition
5	pU6iNEO + ARRB1	To control for protein reducing cellular survival (not retrotransposition)
6	pU6iNEO + A3A	To control for protein reducing cellular survival (not retrotransposition)
7	pU6iNEO + TPRX1	To control for protein reducing cellular survival (not retrotransposition)
8	Mock transfection	Negative control for survival - check cells die with selection
9	ARRB1 only	Negative control for survival - check this protein does not lead to cell survival
10	A3A only	Negative control for survival - check this protein does not lead to cell survival
11	TPRX1 only	Negative control for survival - check this protein does not lead to cell survival; estimate transfection efficiency with immunocytochemistry stain
12	L1Ψ	Negative control for survival - check that unintegrated L1 does not lead to survival

Table 5.1 List of conditions for retrotransposition assay.

5.4.4 Differential expression of transposable elements

Data from a published dataset in which ETCHbox genes were ectopically expressed in fibroblasts were used to test whether the presence of ETCHbox proteins affects the expression of TEs. Human *ARGFX*, *DPRX*, *LEUTX* and *TPRX1* genes plus an empty expression construct were expressed in human adult fibroblasts and RNA-sequencing conducted[90]. Raw RNA-seq data were downloaded from NCBI's Sequence Read Archive (SRA) (accession SRP073307). Quality control was completed using FastQC version 0.11.8[404] and MultiQC version 1.8[586] and reads trimmed with Trimmomatic version

0.39[587]. SalmonTE version 0.4[661] was used to quantify TE expression in each sample. SalmonTE uses the Salmon[408] pseudoaligner for quantification before calling DESeq2[411] to perform differential expression analysis. The false discovery rate method (FDR = 0.1) was used to correct for multiple testing. Downstream analysis was performed using R version 4.1.0[466]. BioVenn[590] was used to create Venn diagrams.

5.4.5 TE enrichment

To test whether ETCHbox genes have an enrichment of TEs around their loci, RepeatMasker version 4.1.1[668] was run on every gene on chromosome 19 to annotate TEs and calculate the percentage of the sequence composed of TEs. For each gene, the examined sequence included 10 kb upstream and downstream of the coding region, and Welch's t-test was used to compare ETCHbox loci to all other chromosome 19 genes. Dfam[669] was used to create images of TEs at ETCHbox loci.

5.5 Results

5.5.1 Human ETCHbox proteins regulate genes involved in TE repression

As an initial test for a link between ETCHbox genes and TEs, I explored whether any TE-regulating genes are differentially expressed (DE) in response to ETCHbox ectopic expression. Maeso et al. (2016)[90] expressed the human ETCHbox genes *ARGFX*, *DPRX*, *LEUTX* and *TPRX1* in human dermal fibroblasts and performed DE analysis on RNA-seq data but did not investigate TE-regulating genes. I first searched for KZNFs in the DE gene sets. DPRX, LEUTX and TPRX1 downregulate 6, 42 and 32 KZNFs respectively (Figure 5.2A; Supplementary Table S5.2); for each gene, this is a significant enrichment of KZNFs in the DE gene set (Fisher's exact tests, Supplementary Table S5.3). In contrast, the ARGFX-downregulated gene set has a significant underrepresentation of KZNF genes. There is a significant overlap between the KZNFs downregulated by DPRX, LEUTX and TPRX1 (Figure 5.2B, Supplementary Table S5.4);

the large overlap between LEUTX and TPRX1 KZNFs reflects a broader overlap in function between the two genes[90]. There is also a significant overlap between the ARGFX up KZNFs and LEUTX and TPRX1 down KZNFs (Figure 5.2B, Supplementary Table S5.4), consistent with the previously observed antagonistic function of ARGFX against LEUTX and TPRX1[90].

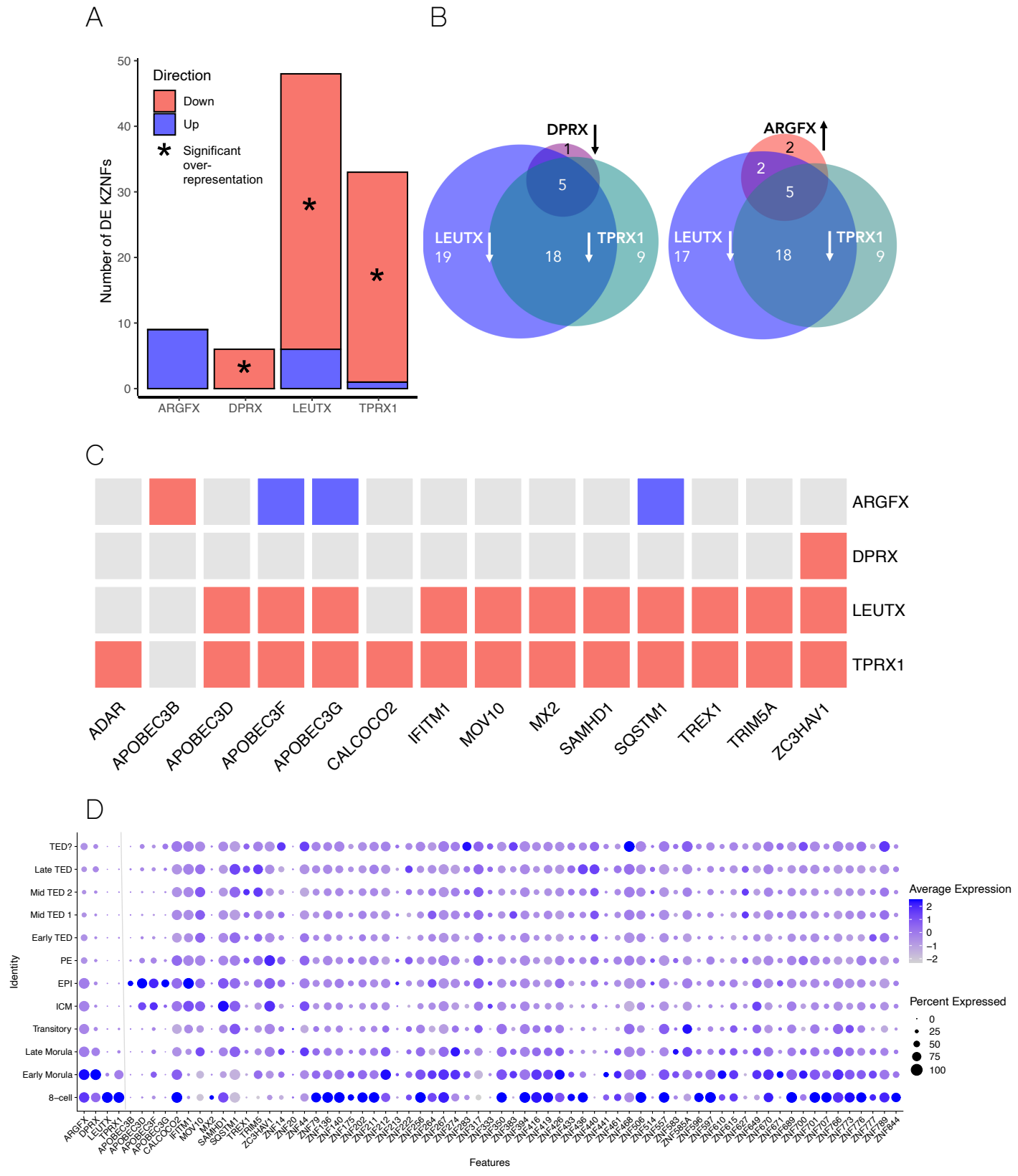


Figure 5.2 Human ETCHbox genes regulate genes that repress transposons. (A) Bar chart shows number of KZNFs DE in response to ectopic expression of human ETCHbox genes. Asterisks highlight a significant overrepresentation of KZNFs in the DE gene sets. **(B)** Venn diagrams show overlap of DE KZNFs between ETCHbox genes.

Circle areas and overlaps are proportional to the number of genes in each segment. **(C)** Non-KZNF TE-repressive genes DE in response to ectopic expression of human ETCHbox genes. Genes that affect TE expression by piRNA or epigenetic mechanisms are excluded. Grey = not DE, red = downregulated, blue = upregulated. **(D)** Expression of ETCHbox genes and ETCHbox-regulated TE-repressive genes in an scRNA-seq dataset from the human preimplantation embryo. Point size is proportional to the number of cells expressing a gene, and point colour reflects expression level.

Furthermore, I find that genes that act at a post-transcriptional level in myriad further TE-repressive pathways are also regulated by human ETCHbox proteins (Figure 5.2C). These include multiple *APOBEC3* genes plus *ADAR*, *CALCOCO2*, *IFITM1*, *MOV10*, *MX2*, *SAMHD1*, *SQSTM1*, *TREX1*, *TRIM5A* and *ZC3HAV1*. With only three exceptions, all such genes are downregulated by both LEUTX and TPRX1; this consistent response suggests a central role for these two ETCHbox proteins in downregulating TE response pathways. *APOBEC3F*, *APOBEC3G* and *SQSTM1* are also upregulated by ARGFX, again consistent with antagonism of ARGFX versus LEUTX and TPRX1[90].

Since this data is obtained from ectopic expression in fibroblast culture, I explored whether the DE genes are realistic embryonic targets. The vast majority of DE KZNFs and all of the 14 non-KZNF TE-repressive genes are expressed in the preimplantation embryo at the time of native ETCHbox expression (Figure 5.2D), making them plausible in vivo targets. The majority of KZNFs are most highly expressed at the 8-cell stage and then subsequently downregulated; this coincides with the timing of the expression of *LEUTX* and *TPRX1*, which were found to downregulate them in cell culture. Two of the three non-KZNF genes upregulated by ARGFX (*APOBEC3G* and *APOBEC3F*) show specificity to the ICM and/or epiblast (Wilcoxon rank sum tests; Supplementary Table S5.5); this is notable because ARGFX also exhibits such a specificity (fully discussed in Chapter 6). Overall, human LEUTX and TPRX1 downregulate expression of a large suite of genes that repress TEs.

5.5.2 Regulation of TE-repressive genes by ETCHbox proteins occurs across species

To test whether the regulation of TE-repressive genes by ETCHbox proteins is human-specific or a more ancient gene property, we searched for TE-regulating genes in the bovine ARGFX and LEUTX DE datasets produced in Chapter 3. *ARGFX* and *LEUTX* ectopic expression resulted in the downregulation of 27 and 26 KZNFs respectively in bovine foetal fibroblast cells; just one KZNF was upregulated by *LEUTX* and none by *ARGFX* (Figure 5.3A; Supplementary Table S5.2). These are significant underrepresentations of KZNFs in both upregulated gene sets and an overrepresentation in the ARGFX-downregulated gene set (Supplementary Table S5.3). There is a significant overlap in the KZNFs downregulated by *ARGFX* and *LEUTX* (Figure 5.3B; Supplementary Table S5.4), consistent with their overlapping role identified in Chapter 3. We find a minimal overlap in the KZNFs regulated by human and bovine ETCHbox genes (Fisher's exact tests, $p > 0.05$ for all tests; Supplementary Figure S5.1, Supplementary Table S5.4), suggesting that while the functional role of KZNF-regulation has been conserved, the identity of the regulated KZNFs has diverged, consistent with the genes' fast-evolving nature.

As in human, bovine ETCHbox genes downregulate a multiplicity of further TE-repressive genes (Figure 5.3C). LEUTX acts similarly in human and cattle, broadly downregulating many TE-repressive genes. Of these, *CALCOCO2*, *IFITM1*, *MOV10*, *MX2*, *SQSTM1*, *TREX1* and *TRIM5* are all downregulated by LEUTX in both species. Interestingly, ARGFX appears to act in opposite directions: in humans it upregulated *APOBEC3F*, *APOBEC3G* and *SQSTM1*, but in cattle it broadly downregulates TE-repressive genes in a similar fashion to LEUTX. This is consistent with the results of Chapter 3, which suggested that human and cattle ARGFX proteins perform opposing roles. Bovine APOBEC genes were not expressed in the fibroblast culture (Chapter 3 Supplementary Table S3.1) so ARGFX or LEUTX downregulation of those genes would not be detected in this analysis.

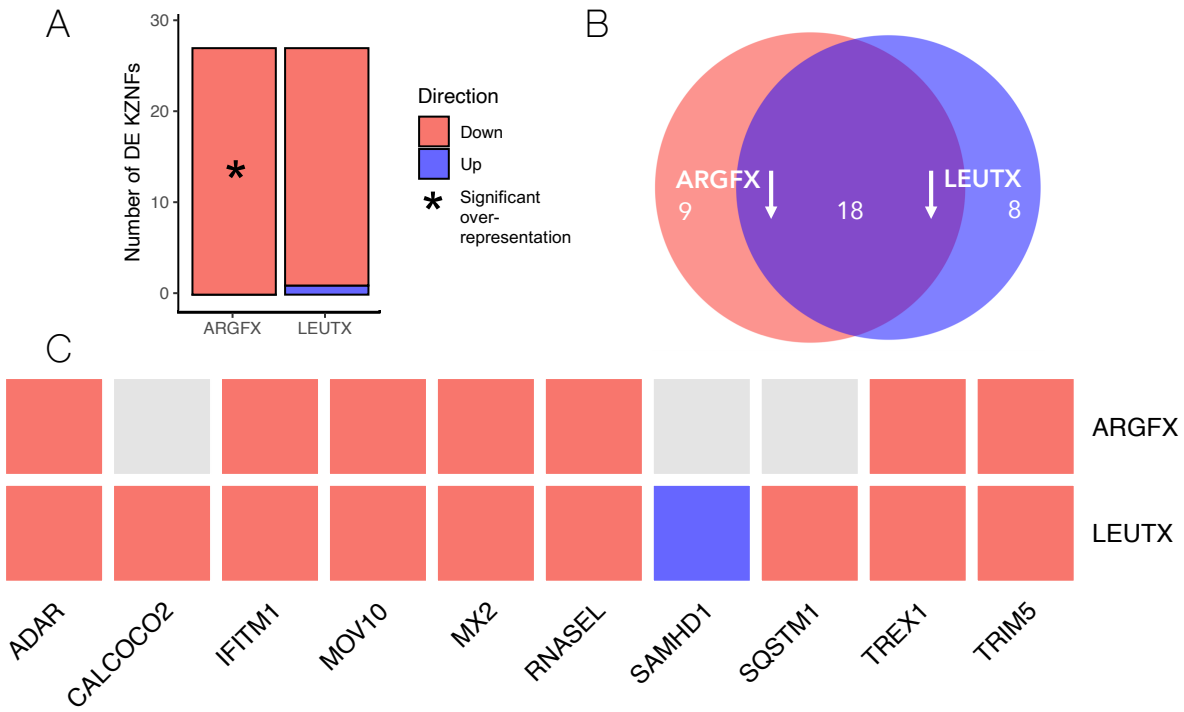


Figure 5.3 Bovine ETCHbox genes regulate genes that repress transposons. (A) Bar chart shows number of KZNFs DE in response to ectopic expression of bovine ETCHbox genes. Asterisks highlight a significant overrepresentation of KZNFs in the DE gene sets. **(B)** Venn diagram shows overlap of DE KZNFs between ETCHbox genes. Circle areas and overlaps are proportional to the number of genes in each segment. **(C)** Non-KZNF TE-repressive genes DE in response to ectopic expression of bovine ETCHbox genes. Genes that affect TE expression by piRNA or epigenetic mechanisms are excluded. Grey = not DE, red = downregulated, blue = upregulated.

I also performed GO analyses on the DE gene sets. Based on the antagonistic relationship of ARGFX with LEUTX and TPRX1 observed above with respect to TE-repressive genes and by Maeso et al. (2016)[90] more generally, in human I took a subset of the data, restricting the results to terms that were enriched in the gene sets upregulated by ARGFX and downregulated by LEUTX and TPRX1 (124 genes, Supplementary Table S5.6). For GO biological process terms, we find that 9 terms related to cellular viral defence pathways appear in this category (Table 5.2 and Supplementary Table S5.7). Because of the dual role many proteins possess in viral defence and TE defence, many of the genes that caused these terms to be enriched are also involved in TE defence (e.g. *IFITM1*, *APOBEC3D*, *APOBEC3F*, *APOBEC3G*, *MOV10* etc.). GO analysis for bovine data was performed as part of Chapter 3 (Chapter

3 Supplementary Table S3.3). As in human, we find numerous enriched GO terms related to the antiviral response (Table 5.2). This supports the above conclusions that, in both species, ETCHbox genes regulate more TE-repressive genes than would be expected by chance.

Overall, the differential expression of many TE-repressive genes from a wide diversity of pathways in response to ETCHbox protein expression across both humans and cattle suggests that ETCHbox proteins regulate this network in the preimplantation embryo, and that the evolution of this function predates the divergence between the Euarchontoglires and Laurasiatheria.

Antiviral-related GO term	**Human *ARGFX* up, *LEUTX* down and *TPRX1* down DE gene set FDR-adjusted *p*-value**	**Bovine *ARGFX* down FDR-adjusted *p*-value**	**Bovine *LEUTX* down FDR-adjusted *p*-value**
cellular response to virus (GO:0098586)	8.82E-03	NS	NS
defense response to virus (GO:0051607)	1.79E-07	4.94E-13	8.03E-10
negative regulation of viral entry into host cell (GO:0046597)	NS	2.44E-02	4.97E-03
negative regulation of viral genome replication (GO:0045071)	1.81E-03	3.27E-07	7.33E-05
negative regulation of viral life cycle (GO:1903901)	NS	3.41E-02	NS
negative regulation of viral process (GO:0048525)	1.71E-03	5.59E-09	0.000236
regulation of defense response to virus (GO:0050688)	5.30E-03	NS	NS
regulation of viral entry into host cell (GO:0046596)	NS	3.69E-05	NS
regulation of viral genome replication (GO:0045069)	8.85E-03	2.86E-05	3.30E-03
regulation of viral life cycle (GO:1903900)	9.10E-03	7.18E-08	6.83E-05
regulation of viral process (GO:0050792)	1.74E-02	9.23E-08	1.73E-04
response to virus (GO:0009615)	1.84E-07	2.04E-09	3.06E-07

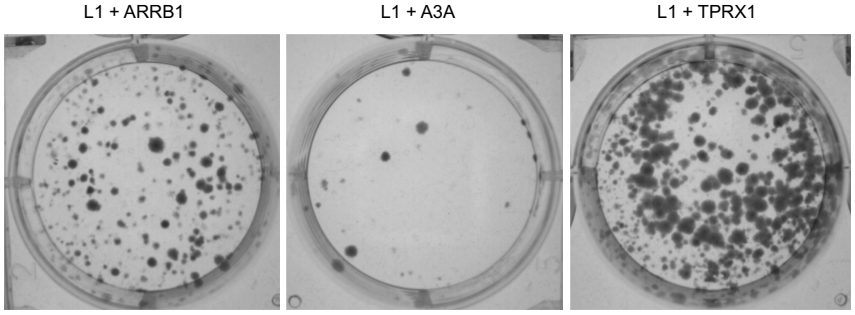
Table 5.2 GO biological process terms related to the antiviral response which are enriched in ETCHbox DE gene sets. In human, the tested genes are those which are upregulated by ARGFX and downregulated by LEUTX and TPRX1. In bovine, the ARGFX and LEUTX DE gene sets were tested separately. *P*-values are from binomial tests with FDR corrections for multiple testing. Abbreviations: FDR = false discovery rate; GO = gene ontology; NS = not significant.

5.5.3 Effect of ETCHbox proteins on the rate of retrotransposition

The proteins encoded by *ADAR*, *APOBEC3B*, *APOBEC3F*, *APOBEC3G*, *MOV10*, *MX2*, *SAMHD1*, *TREX1*, *TRIM5* and *ZC3HAV1*, all of which are downregulated in response to LEUTX and/or TPRX1 expression in human, act to inhibit retrotransposition[645,648–650,670–674]. In light of this, I questioned whether the ectopic expression of ETCHbox proteins and subsequent downregulation of inhibitors of retrotransposition was sufficient to modify the rate of retrotransposition events occurring in a given cell population. To test this, I implemented a LINE-1 cultured cell retrotransposition assay[627,666]. This assay tests whether the expression of an exogenous protein changes the rate of retrotransposition events. Briefly, a retrotransposition-competent LINE-1 transposon, complete with antibiotic resistance selective marker in its 3' UTR, is transfected into a HeLa cell culture with and without the protein of interest. The antibiotic resistance cassette is only expressed if a retrotransposition event occurs, so the number of clones surviving post-selection is used to quantify the rate of retrotransposition.

I tested whether the human ETCHbox protein TPRX1 affects the rate of LINE-1 retrotransposition. To achieve this, the mean number of colonies in samples transfected with the LINE-1 plasmid plus an ARRB1 expression plasmid, which is known to have no effect on retrotransposition rate, is taken as a baseline and designated as a retrotransposition efficiency of 100%. After correction for cellular toxicity, the mean retrotransposition efficiency of the positive control (LINE-1 plus APOBEC3A [A3A], which is known to restrict retrotransposition) was 9.7%, showing that the assay functioned as expected. The retrotransposition efficiency of LINE-1 plus TPRX1 was 154.4%; the

presence of TPRX1 in a cell increased the rate of retrotransposition by approximately 50% (Figure 5.4; Supplementary Table S5.8). This suggests that the downregulation by TPRX1 of retrotransposition-restricting genes like those of the *APOBEC3* group results in a detectable increase in the rate of retrotransposition.



Condition	% Retro. efficiency
ARRB1	100.0
A3A	9.7
TPRX1	154.4

Figure 5.4 Effect of TPRX1 on LINE-1 retrotransposition. ARRB1 is the negative control, it does not affect the rate of retrotransposition – it sets the baseline rate. A3A is the positive control, it is known to reduce the rate of retrotransposition. TPRX1 is the protein of interest.

Two qualifying factors must be considered when interpreting this result. First, this result is drawn from one biological replicate consisting of three technical replicates and is therefore not statistically robust. Though I performed two more biological replicates, the transfection efficiency of TPRX1 in the second and third replicates was poor (Supplementary Figure S5.2) and therefore the results are uninformative. Given these factors, the above result should be interpreted with caution as preliminary work, with additional experiments required to test the results.

Second, samples transfected with only the TPRX1 construct consistently have between 1 and 6 apparently antibiotic-resistant colonies surviving after antibiotic selection (Supplementary Figure S5.3). This is not the case for ARRB1 or A3A. The TPRX1 expression plasmid does not have a neomycin resistance gene and, in any case, the rate of survival would be much higher if this was the reason for the surviving colonies. I do not have an explanation for this effect, but the current data analysis procedure should account for it regardless. The calculation of retrotransposition efficiency includes a

correction for cytotoxicity of the tested protein. This is achieved by calculating a ratio of how many more (or less) colonies there are in the LINE-1 plus TPRX1 treatment versus the pU6iNEO plus TPRX1 treatment (see section 5.4.3), and subsequently comparing this to the corresponding figure for the ARRB1 samples in which retrotransposition is unaffected by the tested protein. This controls for the effect of cellular toxicity of TPRX1 but also controls for TPRX1 having the opposite effect, with *extra* colonies surviving. Since both of these conditions should be subject to this unexplained increase, its effect would be eliminated. Despite this, I still urge a measure of caution in interpreting this result until this effect is understood.

5.5.4 ETCHbox proteins have a minimal effect on TE expression

TE activity can be restricted not only by preventing retrotransposition but also prior to this by transcriptional repression. I hypothesised that either directly or due to their effects on proteins like KZNFs which act to restrict TE transcription, ETCHbox proteins may act at a transcriptional level to modulate the *expression levels* of TEs. Using data from a published study that expressed the human ETCHbox genes *ARGFX*, *DPRX*, *LEUTX* and *TPRX1* in human fibroblasts and performed RNA-seq[90], I quantified the expression of TEs and tested for differential expression between ETCHbox-transfected and control samples. Very few subfamilies of TEs are DE (Figure 5.5A; Supplementary Table S5.9); ARGFX, DPRX, LEUTX and TPRX1 protein expression causes the differential expression of 9, 2, 29 and 12 TE subfamilies, respectively. Given a false discovery rate (FDR) of 0.1 was used to account for multiple testing, this is likely to be data noise rather than a true signal. No clear pattern is observed in the direction of change of expression (Figure 5.5B), and TEs from all major clades are affected (Figure 5.5C). This again points to DE subfamilies being data noise. However, I do find a large overlap between the TEs downregulated in response to LEUTX and TPRX1 expression (Figure 5.5D). This is consistent with the reported overlapping effect of LEUTX and TPRX1 on gene expression[90], suggesting this may be a true signal.

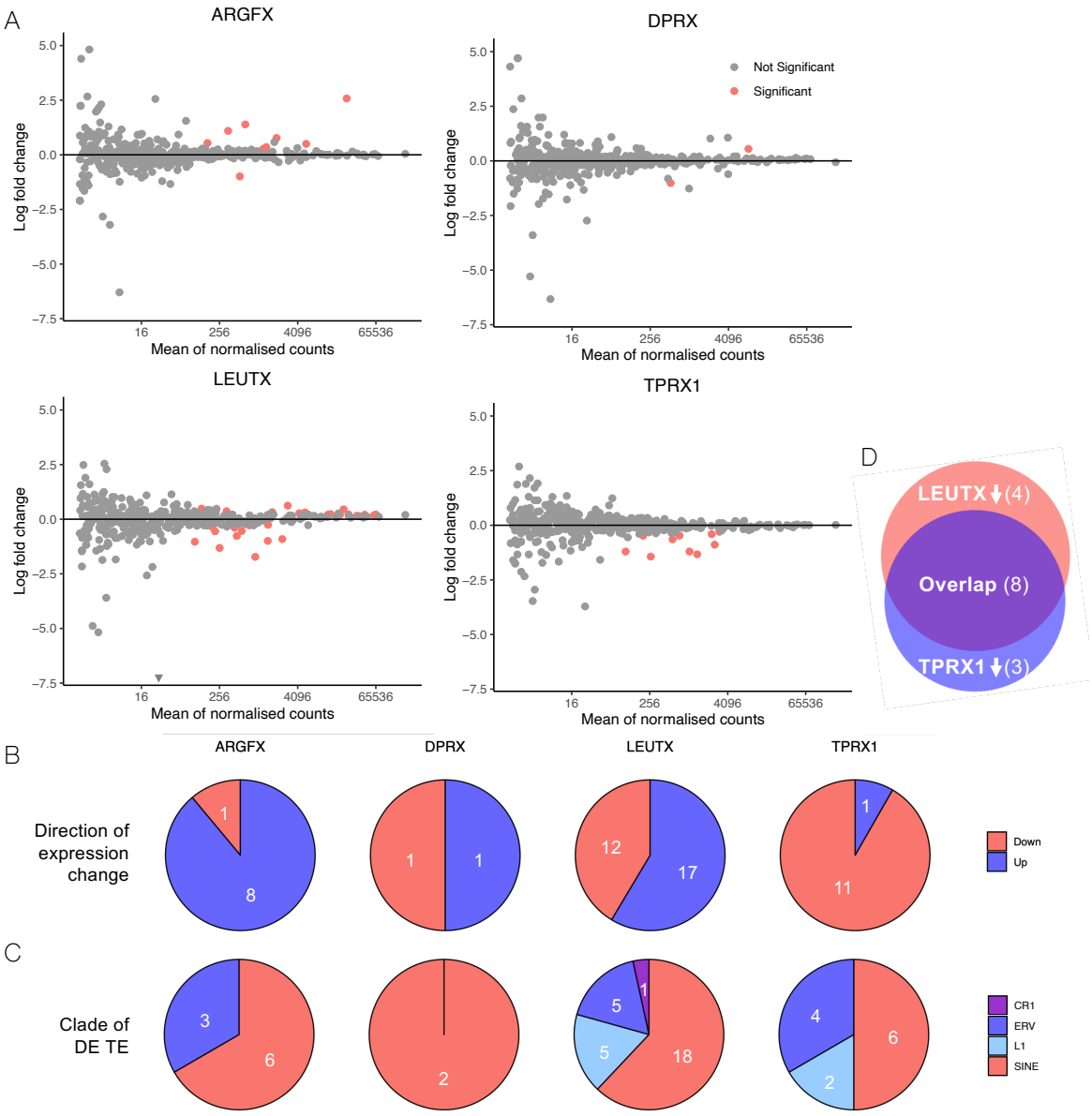


Figure 5.5 Effect of ETCHbox proteins on TE expression. (A) MA plots showing fold change of TE subfamilies in ETCHbox-transfected versus control samples. Red points show DE TEs. **(B)** Direction of change of DE TEs. **(C)** Clades to which DE TEs belong. **(D)** Overlap between TEs downregulated in response to LEUTX and TPRX1 expression.

5.5.5 ETCHbox loci are enriched in SINEs

As the final part of my exploration of the relationship between ETCHbox proteins and TEs, I switched the focus from the idea of ETCHbox proteins regulating TEs to TEs regulating ETCHbox proteins. TEs have been integrated into the GRNs governing preimplantation development as key *cis*-regulatory elements[61,65,377], so I questioned whether there is an enrichment of TEs around ETCHbox loci, which may point towards their involvement in the regulation of ETCHbox genes. I find that a higher proportion of

ETCHbox loci (versus all other genes on chromosome 19) consists of TEs (Welch's t-test p = 6.60 x 10^{-4}) and that this is driven by a higher density of SINEs (Welch's t-test p = 3.92 x 10^{-2}) (Figure 5.6A); The genomic region of *ARGFX*, which includes up to 57 SINE elements, is annotated and shown as an example (Figure 5.6B). This analysis hints at a possible role for SINEs in the regulation of ETCHbox expression.

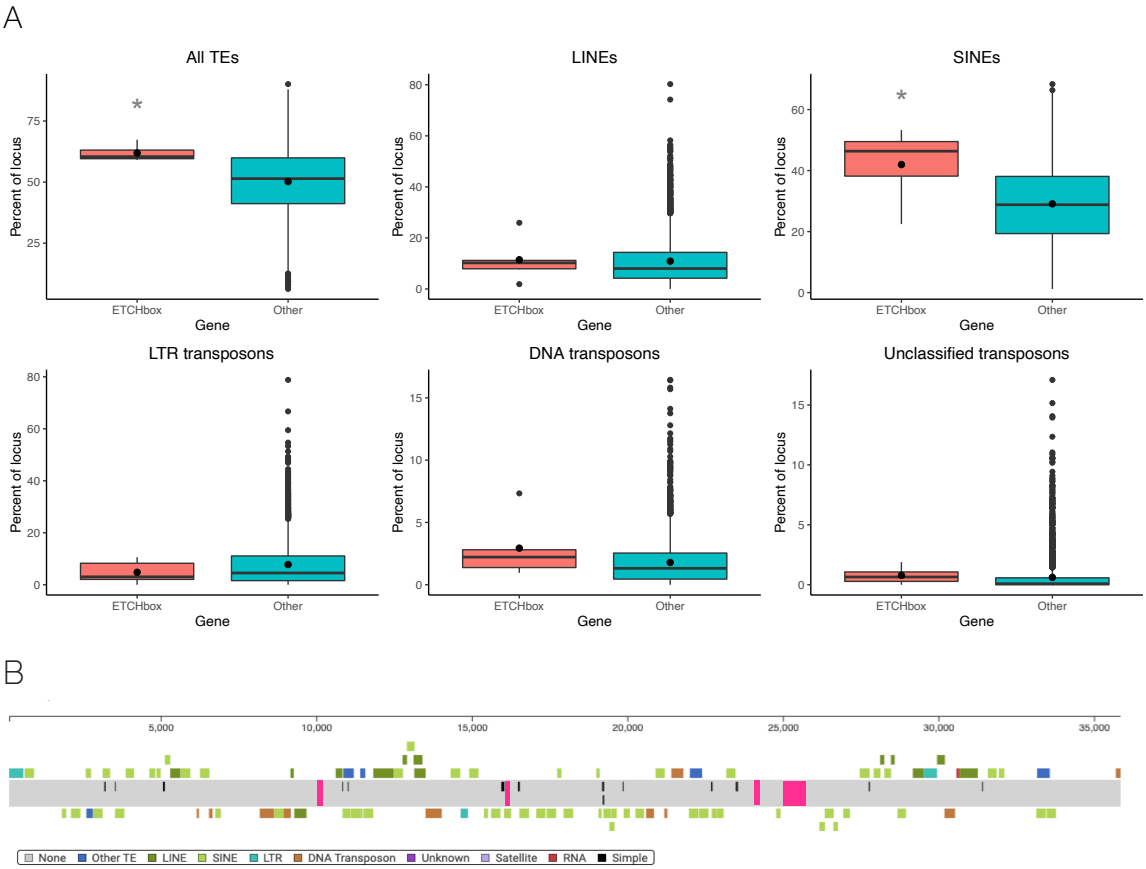


Figure 5.6 TE enrichment at ETCHbox loci. (A) Plots show the percent of each locus composed of TEs for ETCHbox genes versus all other chromosome 19 genes. Asterisks denote a significantly higher percentage at ETCHbox loci. **(B)** Graphical representation of repeat sequences present at the human *ARGFX* locus. Pale green bars represent SINEs, which are overrepresented at ETCHbox loci. Pink bars show ARGFX protein-coding exons, the second of which overlaps with a SINE because it is composed of an AluJb element.

5.6 Discussion

The simultaneous occurrence of ETCHbox expression and the rise in the activity of TEs during preimplantation development provoked the idea that ETCHbox proteins may play a role in regulating TEs in the embryo. The initial idea for ETCHbox proteins as TE regulators emerged as an explanation for their rapid evolution. Consider a scenario in which ETCHbox proteins directly bind to transposons in the genome and coordinate their repression. The resulting selection on TEs to escape repression and ETCHbox proteins to maintain it would generate an arms race that could result in rapid sequence evolution. However, two results presented in this chapter suggest that this is not the case. First, the regulation of numerous TE-repressive genes by ETCHbox proteins suggests that any effect on TEs is indirect and mediated by intermediate factors, meaning no direct binding of TEs, no arms race and no rapid evolution. Second, and most conclusively, ETCHbox expression in cell culture has a negligible effect on TE expression, suggesting that the proteins do not function as transcriptional regulators of TEs.

To summarise, my data controvert the initial hypothesis of direct TE regulation but show compellingly that human LEUTX and TPRX1 downregulate a multitude of genes that are known to repress transposons. Moreover, data from the retrotransposition assay suggest that, at least in the case of TPRX1, this results in an increase in the rate of retrotransposition (though, as emphasised above, caution is urged when interpreting this result). Naturally, this leads us to ask why, at a point in development pre-specification of the germline, and therefore in which the mutagenic insertion of a TE would be heritable, embryos have been selected to downregulate the genes that protect them from aberrant TE activity. One possibility is that the downregulation of TE-repressive genes by ETCHbox proteins does not result in an increase in retrotransposition in vivo, and that this is a product of a cell culture environment which is a poor analogue of the preimplantation embryo. The chromatin state, methylation levels, transcriptome and

proteome of HeLa cells are highly divergent to that of an 8-cell blastomere, and it is conceivable that the cellular environment of the preimplantation embryo consists of factors that prevent an ETCHbox-mediated increase in the rate of retrotransposition even though this is not the case in vitro.

If for a moment we lay this possibility aside and assume that the increase in the rate of retrotransposition in cell culture is indeed reflective of the effect of ETCHbox expression in the embryo, it is interesting to consider the possible evolutionary causes of this phenomenon. Given the assumed fitness cost of an elevated rate of retrotransposition, it seems most likely that this effect would be a by-product rather than the selected function of ETCHbox proteins. For instance, the ADAR protein, which restricts retrotransposition and is downregulated by TPRX1, functions through RNA editing by deaminating adenosine to inosine, which is then read by ribosomes as guanosine and thus changes the protein sequence[675]. The fitness benefit of reducing ADAR expression and therefore preventing the aberrant editing of key protein-coding mRNAs at this critical point in development may outweigh the fitness cost of a slight uptick in the rate of retrotransposition. In this way, the increase in retrotransposition rate is not directly adaptive but is a by-product of an adaptive function.

Is it possible that increasing the rate of retrotransposition is an adaptive function of ETCHbox genes? Like there is an optimum rate of mutation[191,676–678], a balance between the creation of genetic novelty and the fitness costs of deleterious changes, I wonder if there is a non-zero optimum rate of retrotransposition (from the perspective of the 'host' genome). TEs are a source of genetic novelty, and their exaptation as *cis*-regulatory elements, non-coding RNAs or even part of protein-coding sequences is a consistent feature of mammalian evolution[344,376,679]. This is especially pertinent to the preimplantation embryo, in which TF binding sites in TEs have allowed the rewiring of gene regulatory networks (GRNs) and coordination of stage-specific

transcription[65,344,376,680,681]. The dispersal of TF-binding motifs throughout the genome is dependent on retrotransposition[679], and it is therefore possible that selection on the host genome acts to increase this. However, given the strong selection pressure on 'selfish' TEs themselves to duplicate, it seems unlikely that it would be necessary for selection to favour host factors to facilitate this.

One final point of interest emerges from this analysis. Human ARGFX upregulates APOBEC3F and APOBEC3G (in contrast to the LEUTX and TPRX1 downregulation). These two proteins repress TE retrotransposition by cytosine deamination[649,682,683], and scRNA-seq data show that expression of both is specific to the epiblast in the preimplantation embryo. In the next chapter, I find that human ARGFX itself is also specific to the epiblast, suggesting it upregulates these transposon-restricting genes in this cell lineage. The epiblast is the blastocyst lineage fated to become the embryo itself[336], as opposed to extra-embryonic membranes, and therefore aberrant TE activity in this location would have the biggest evolutionary impact. It has been suggested in the past that the mammalian embryo makes a concerted effort to protect the epiblast specifically from TE-induced mutagenesis[340], and this result is consistent with that hypothesis.

The above discussion considers TE-repressive genes that act at a post-transcriptional level, for example by RNA editing. However, large numbers of KZNFs, transcriptional repressors of TEs, are also downregulated in response to human and bovine ETCHbox expression. ScRNA-seq data (Figure 5.2D) shows a broad downregulation of these DE KZNFs after the 8-cell stage of human development. This is consistent with the timing of ETCHbox expression and highlights the genes as realistic targets of ETCHbox function. Despite this, it is noticeable that TE quantification in RNA-seq data revealed only minimal changes induced by ETCHbox expression, suggesting that the change in KZNF expression had little impact of TE transcription. However, the same qualifying factors as

above apply here. This analysis used data from a fibroblast ectopic expression experiment; it should maybe be unsurprising that the epigenetic state of fibroblasts does not permit TE transcriptional changes in a similar way to the preimplantation embryo, and it is therefore difficult to judge the significance of this negative result. For instance, KZNFs function by guiding histone methylation apparatus[636], and while repression can be transient[684], they can also induce stable, irreversible repression of TE loci[685–687], and therefore the downregulation of KZNFs in fibroblasts will inevitably have little effect on such permanently repressed TEs. In future, in vivo knockdown experiments followed by quantification of TE expression would be the ideal test of this question.

In addition, as discussed in section 5.2, the properties of TEs create challenges for bioinformatic analyses. One drawback of my use of SalmonTE[661] for TE quantification is that its resolution is limited to the subfamily level, which does not permit the analysis of specific TE loci. This could mask a specific effect of ETCHbox proteins on one or two individual loci, and does not allow us to distinguish a direct effect on TEs from their expression changing because they are present in a DE gene. In light of this, I also ran the locus-specific TE pipeline SQuIRE[688] on the same data, but a recent comparison of TE-quantification methods found it to produce large numbers of false positives[660], and thus I did not consider the results reliable. A repeat of this analysis at the locus-specific level once tools for this purpose improve is necessary to test the result that ETCHbox proteins have a minimal effect on TE expression levels.

Finally, I found an enrichment of SINE elements at ETCHbox loci compared with those of other genes on chromosome 19. TEs have been integrated into the GRNs controlling mammalian preimplantation development, and function for example as stage-specific transcription initiation sites with binding sites for key developmental TFs like Dux (for a more complete discussion see Chapter 1 section 1.6.4). The exaptation of SINEs to act as enhancers driving stage- or location-specific transcription has occurred on several

occasions, including in mammals[689–692]; in future it would be interesting to investigate whether the observed overrepresentation at ETCHbox loci has a biological function such as this. Incidentally, the presence of repetitive elements increases the likelihood of unequal crossing over (i.e. recombination between paralogous sequences) and therefore of gene duplications and losses. The large proportion of ETCHbox loci made up of SINEs may therefore have contributed to their elevated rates of copy number change during mammalian evolution.

5.7 Conclusions

This chapter offers an exploration of the relationship between ETCHbox proteins and TEs. My first result draws a link between these two apparatuses of the preimplantation embryo, demonstrating that several human and bovine ETCHbox genes act to broadly downregulate genes that restrict TE retrotransposition. Cell culture retrotransposition assays suggest that this results in an increase in the rate of TE genomic insertions, but the implications of this for early development remain uncertain. In addition, human and bovine ETCHbox proteins downregulate a suite of KZNFs, which are transcriptional repressors of TEs, but quantification of TE transcripts in ETCHbox ectopic expression experiments suggests that this does not translate to a difference in TE expression. Again, the relevance of these cell culture experiments to the in vivo functions of ETCHbox proteins requires additional effort. Overall, this work points to an interaction between ETCHbox proteins and TE dynamics, but further work is required to establish the significance of this to the biology of the preimplantation embryo.

5.8 Supplementary figures

Figure S5.1



Figure S5.1 Overlaps between human and cow KZNF genes DE in response to ETCHbox gene expression. Corresponding Fisher's exact test scores available as Supplementary Table S5.4.

Figure S5.2

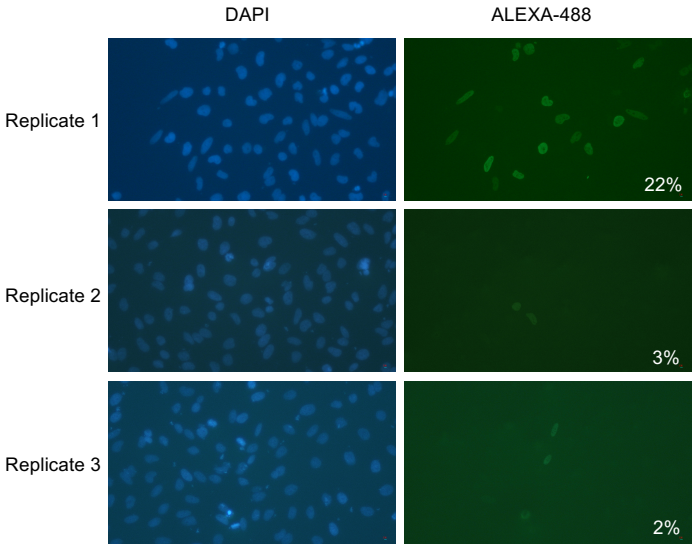


Figure S5.2 Transfection efficiencies of replicates 1, 2 and 3. Figure shows immunocytochemistry stain of TPRX1-transfected samples from each replicate. White numbers show sample transfection efficiency. Efficiency of replicate 1 was approximately 10 times that of replicates 2 and 3.

Figure S5.3

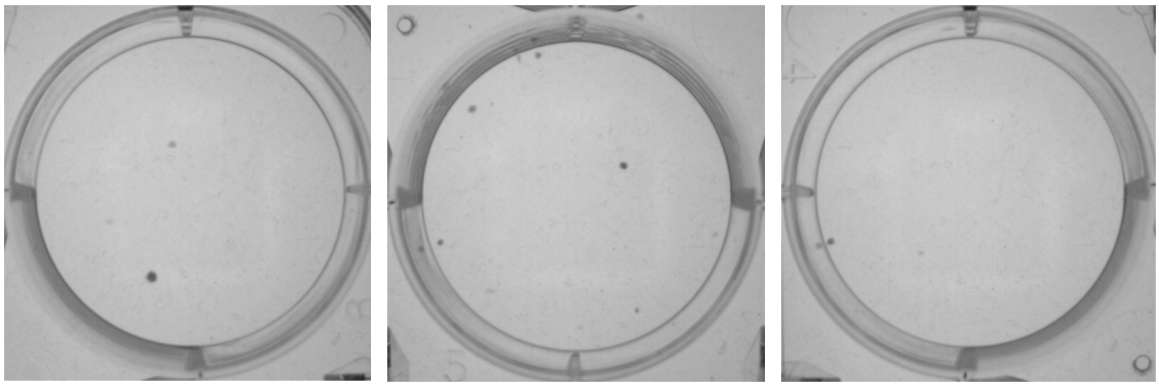


Figure S5.3 Colonies in TPRX1 negative control samples.

Electronic supplementary material

ESM5.1 Supplementary Tables S5.1 – S5.9.

Chapter 6: Spatiotemporal expression of ETCHbox genes in the preimplantation embryo

Statement of authorship contributions

This project was conceived by discussions between Thomas D. Lewin and Peter W. H. Holland. Initial advice on single-cell RNA-sequencing data analysis was provided by Jordi Solana (Oxford Brookes University). Peter Mulhair (University of Oxford) provided advice on the writing of custom python scripts. Data analysis, data interpretation and writing were carried out by TDL. PWHH supervised the work.

6.1 Abstract

The advent of single-cell protocols has revolutionised the type of questions that RNA-sequencing data can be used to answer, and one of the key advances has been the ability to explore spatial expression patterns at high resolution. A critical question left open by previous work on ETCHbox genes but now able to be addressed is whether their expression in the morula and blastocyst is universal or specific to certain cell lineages. In this work, I analyse published single-cell RNA-sequencing (scRNA-seq) datasets spanning preimplantation development in humans, mice and pigs. In humans, the genes peak in expression in the order *LEUTX*, *TPRX1*, *DPRX*, *ARGFX*; during the latter part of its expression, *ARGFX* is highly specific to the inner cell mass (ICM) and then epiblast, the lineages fated to become the embryo proper. It also upregulates several genes that themselves show ICM and epiblast-specificity, including *BAMBI*, a known ICM marker; these data suggest a role for ARGFX in the formation of the ICM, which is a eutherian novelty. Several other notable conclusions emerge from these analyses: firstly, ETCHbox genes are consistently activated during the major wave of embryonic genome activation (EGA) across humans, mice and pigs, and differences in their timing of activation between species reflect differences in the timing of EGA. Secondly, in humans, DUXA is identified as a putative activator of *ARGFX* and *LEUTX* expression in the preimplantation embryo. Thirdly, the highly duplicated ETCHbox genes of mice show some degree of temporal compartmentalisation of function, suggesting that subfunctionalisation has been a contributing factor in their retention. Finally, I find no evidence in support of the suggestion that a role in placentation has driven rapid ETCHbox evolution, as they show no expression in the lineage of the embryo fated to become the placenta.

6.2 Introduction

Despite over 15 years passing since their discovery, the functional role that ETCHbox genes perform during preimplantation development remains only partially understood. In humans, their expression occurs at a time of major embryonic upheaval: not only are there sequential waves of TE reactivation (Chapter 5), but embryonic genome activation (EGA) is occurring and, notably, the first cell fate decisions are being made. The major wave of EGA occurs in humans at the 4 to 8-cell stages (embryonic day 3) and is followed by compaction of the blastomeres to form the morula (day 4)[289,291,302]. The first cell fate decisions then occur during the initial formation of the blastocyst (day 5), separating trophectoderm (TED) cells from those of the inner cell mass (ICM) (Figure 6.1)[297,302,693]. A further cell fate decision then divides the ICM (day 6), separating the epiblast, which will go on to form the embryo itself, from the primitive endoderm, which is fated to form the yolk sac[302].

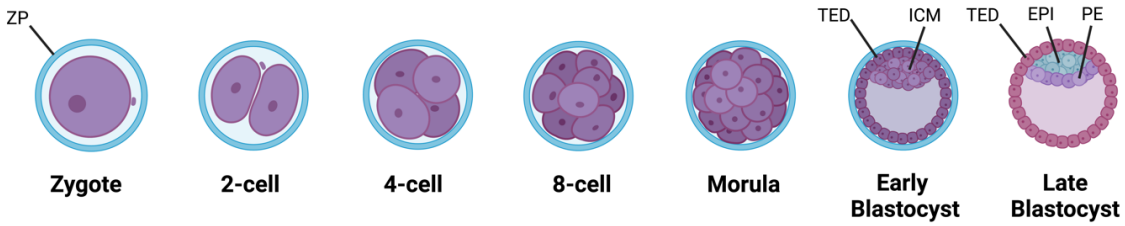


Figure 6.1 Preimplantation embryonic development in humans. Abbreviations: EPI = epiblast; ICM = inner cell mass; PE = primitive endoderm; TED = trophectoderm; ZP = zona pellucida. Figure produced using BioRender.

These morphological stages of preimplantation development are well conserved across eutherians, including the house mouse (*Mus musculus*) and the domestic pig (*Sus scrofa*), but the genetic networks underlying them show divergence, and the timing of key events varies[302,335,336,339,694]. For example, mouse preimplantation development is accelerated compared to that of human, with major EGA occurring at the 2-cell stage and the three lineages of the blastocyst specified by day 4[285,302,306,331,332].

The point at which ETCHbox expression peaks, just prior to the time at which the first cell fate decisions are made, poses the question of whether the genes play an active role in these lineage specification events. Such a function may be indicated by the specific expression of ETCHbox genes in one or more lineages of the blastocyst, or even in a subset of morula cells if their fates have already been specified. One method of testing whether this is the case would be *in situ* hybridisation, which visualises mRNA transcripts using labelled complementary nucleic acid probes[695–697]. However, the application of these procedures to delicate early preimplantation embryos is highly challenging, and most protocols do not offer quantitative results. The advent of single-cell RNA sequencing (scRNA-seq) and recent publication of scRNA-seq datasets from mammalian preimplantation embryos[339,605,663,698–701] now offers the opportunity to answer this question using transcriptomics; by using marker genes to identify the lineage of blastocyst cells and quantifying the expression of ETCHbox genes in each population, we can test whether ETCHbox genes show differential expression between cell lineages.

In addition, oft-overlooked in previously published work but critical for a complete understanding of ETCHbox genes' contribution to preimplantation development is knowledge of the upstream factors that activate their expression. Such information can help to build a more comprehensive picture of where the genes sit in the gene regulatory networks (GRNs) underlying the very first stages of embryonic development. Once again, scRNA-seq data can be exploited, with thousands of individual cells providing gene co-expression data at a resolution sufficient for GRN inference and the computational prediction of ETCHbox-regulatory factors[702].

On top of exploring the regulation and embryonic functions of the genes themselves, high-resolution data on the expression of ETCHbox genes in preimplantation embryos present an interesting opportunity to study gene duplication. The duplication of genes is a critical process in evolution because it creates new genetic material on which selection,

drift and mutation can act, which may facilitate the evolution of increased complexity[187,190,703]. While the processes which generate duplicates are mostly well understood[190,198], the mechanisms that control the evolutionary fate of duplicate genes remain the subject of much discussion[228]. Since most duplicate genes are pseudogenised within the first few million years of duplication[191], the retention of ETCHbox genes since their origin in the ancestor of eutherians makes them a good model to provide insights into such mechanisms.

The duplication events that have produced species' current ETCHbox repertoires can be considered to have occurred in two main phases. The first is the initial duplication(s) of *CRX* which spawned the six ETCHbox genes in the ancestor of the eutherians. After their birth, the ETCHbox genes appear to have been retained by evolution due to subfunctionalisation: the *CRX* gene of metatherians (the sister group to eutherians) is deduced to perform both retinal and embryonic roles, whereas in eutherians the role of *CRX* is retina-only and ETCHbox genes have taken on the embryonic function[92]. The logical question arising from this is why multiple genes in eutherians are required to perform a function for which one gene is sufficient in metatherians[92]. It is thought that lineage-specific transcription factors (TFs) can facilitate the evolution of lineage-specific innovations[50,190] and, indeed, the presence of an ICM at the blastocyst stage is specific to the eutherian lineage and is completely absent in monotremes and marsupials[127,341]. It is possible that ETCHbox genes evolved as part of the network that forms the ICM.

The second phase of duplication events is equally intriguing: many eutherian groups have undergone further lineage-specific duplications of individual ETCHbox genes, and some of them, such as rabbit *LEUTX* and mouse *Obox*, are present in giant tandem arrays of highly similar genes[96,525]. How and why so many copies are maintained by selection in what is in some cases almost the complete absence of sequence divergence is still to be understood, and one possibility is that their functions are compartmentalised

by expression timing and location. Such changes can occur rapidly, and appear to be key to the retention of developmental genes[190,704]. For example, it was recently shown that, in the purplish bifurcate mussel (*Mytilisepta virgata*), the highly-duplicated early developmental Spiralian TALE (SPILE) homeobox genes each show exquisitely specific expression patterns at the 8-cell and 16-cell stages[47], and that these expression domains contribute to the specification of cell fate[50].

In this chapter, I use publicly-available scRNA-seq data to profile the expression dynamics of ETCHbox genes in human, porcine and murine preimplantation embryos. I explore the hypothesis that the evolution of ETCHbox genes is intertwined with that of the ICM, and find that human *ARGFX* shows ICM-specificity. Importantly, the data also show that ETCHbox genes are consistently activated at EGA across humans, mice and pigs, suggesting that they are conserved EGA genes, and that in humans this is at least in part DUXA-mediated. I also ask whether the *Obox* genes of mice show spatiotemporal variation in their expression patterns, exploring this as a possible explanation for their retention by selection.

6.3 Aims

1. Do ETCHbox genes show lineage-specificity in the morula or blastocyst?
2. Are ETCHbox expression patterns conserved between species?
3. Which genes activate ETCHbox expression?
4. Do the highly duplicated mouse *Obox* genes exhibit spatiotemporal variation in their expression profiles?
5. Are ETCHbox expression patterns consistent with a role in placentation?

6.4 Materials and methods

6.4.1 Analysis of human preimplantation embryo scRNA-seq data

The scRNA-seq dataset of Petropoulos et al. (2016)[663], including 1,529 cells derived from 88 embryos from the 8-cell stage to blastocyst, was downloaded from ArrayExpress as raw counts (E-MTAB-3929). Integrating cells from a dataset including earlier stages was unnecessary as no lineage specification has occurred at this point and previous bulk analyses have quantified ETCHbox expression at these stages[90]. The scRNA-seq pipeline of Seurat version 3.1.5[665,705] was used in R version 4.0.0[466] to process scRNA-seq data. Briefly, quality control filtered out cells with >5% mitochondrial genes (an indicator of poor cell quality[706]), and over 20 million RNA counts (an indicator of a potential doublet, where two cells are sequenced together). Data were then normalised using Seurat's NormalizeData function, and the top 2,000 most variable genes identified. Data were scaled so the mean expression across cells is zero and the variance across cells is one; this prevents the most highly expressed genes dominating downstream analyses. Seurat's graph-based clustering algorithm was then implemented to cluster cells by their transcriptomes using the top 15 principal components. The data were initially over-clustered (resolution = 0.9) to separate clusters with more similar transcriptomes (as was required to separate PE and epiblast lineages) and then the most similar clusters merged.

Uniform manifold approximation and projection (UMAP) non-linear dimensionality-reduction[707] was used to visualise cells on a 2D plane; cells with similar transcriptomes occur together in 2D space. Wilcoxon rank sum tests were used to find markers for each cluster, and clusters were identified based on previously reported marker genes[336,339,663,708–712]. Tests for differential expression of genes between clusters were performed using Seurat's FindMarkers function to implement multiple comparison-corrected Wilcoxon rank sum tests. To verify our approach, the identity of each cell was

compared to that of a previous analysis of this data[710]. Identities were largely consistent, with the exception of primitive endoderm (PE), for which 56% of cells were assigned differently. However, the expression of PE marker genes is strong and highly specific to our PE cluster (Supplementary Figure S6.1), suggesting that it is accurately identified. Despite its separation from the main TED clusters, the cluster labelled 'TED?' is classified as a divergent TED lineage because clustering at a lower resolution places it with TED (Supplementary Figure S6.2), and this is consistent with its position in the pseudotime analysis, which places it in the mid-blastocyst. The absence of expression of ICM, PE and EPI markers and broadly similar expression of TED markers to other TED clusters confirms this. All 38 cells in this cluster were assigned by Stirparo et al. (2018)[710] as TED, supporting this identity. To search for early lineage separation in the morula, a data subset including only morula cells was created and then the clustering and UMAP processes repeated.

The unsupervised algorithm of Monocle 3 version 1.0.0[713,714] was used to construct trajectories of cells and order cells in pseudotime to measure each cell's progression through the developmental trajectory. Branched trajectories locate the point at which cell-fate decisions are made. The UMAP method of non-linear dimensionality-reduction[707] was used for this process.

Single-Cell rEgulatory Network Inference and Clustering (SCENIC) version 1.3.1[715] was implemented in R to reconstruct GRNs from the human scRNA-seq dataset. The rapid GRN inference algorithm of GRNboost2[716] and the TF binding motif database of RcisTarget[717] were used for this process. The t-distributed stochastic neighbour embedding (t-SNE) method of dimensionality-reduction[718] was used to distribute cells in 2D space based on their activity of identified regulons.

6.4.2 Analysis of pig preimplantation embryo scRNA-seq data

To characterise ETCHbox expression in the domestic pig (*Sus scrofa*), a scRNA-Seq data analysis pipeline was applied to two separate datasets. For Dataset 1, containing 90 cells from the zygote to early blastocyst, raw sequencing reads were downloaded from NCBI BioProject (Accession: PRJNA580004)[701]. Dataset 2 consisted of 220 cells from morula to late blastocyst, with raw sequencing reads also downloaded from NCBI BioProject (Accession: PRJNA445862)[699]. Dataset integration did not identify shared cell populations so the two datasets are analysed independently. For both datasets, quality control was performed with FastQC[404] and MultiQC[586], and transcripts were then quantified with Kallisto version 0.46.1[407]; *S. scrofa* ETCHbox sequences were determined previously (Chapter 2[525]) and included in the input transcriptome. Kallisto equivalence classes (ECs) containing multi-mapped reads were retained for clustering analyses. Read count matrices were processed as above, but with a clustering resolution of 0.5. Clusters were identified using previously-reported marker genes[699,701]. In Dataset 1, some embryonic stages are represented by very few cells (e.g. 2 oocyte cells), and are therefore not identified as individual clusters by Seurat. In these specific cases, cells are manually assigned a separate cluster.

6.4.3 Analysis of mouse preimplantation embryo scRNA-seq data

To determine expression profiles of the divergent ETCHbox genes of mice (*Mus musculus*) at single-cell resolution, raw sequencing reads from a dataset covering zygote to blastocyst[700] were downloaded from NCBI BioProject (Accession: PRJNA195938) and processed as above. Clusters were identified by marker genes[339,719] and cells' previously published developmental stage. Full coding sequences of all ETCHbox genes with a complete homeodomain (three *Crxos* [*Tprx1*] transcripts and 28 *Obox* [*Tprx2*] genes) were obtained from Royall et al. (2018)[96] and included in the mouse transcriptome before analysis. The *Obox* nomenclature proposed by Royall et al.[96] is used throughout. For *OboxA*, *OboxB* and *OboxC* gene groups, only uniquely mapped reads were considered,

yielding gene-specific expression profiles. The 16 *OboxD* genes show extremely high similarity, meaning that many of the reads originating from them cannot be mapped uniquely to a specific locus; for these genes, we consider not only uniquely mapped reads but also Kallisto ECs containing reads mapping to multiple loci. Clustal Omega[454] was used to create an alignment of *Obox* genes and then MrBayes[451,452] used to build a Bayesian gene tree.

6.5 Results

6.5.1 Clustering of human preimplantation embryo cells

To map the spatiotemporal expression patterns of ETCHbox genes in the human preimplantation embryo, we analysed publicly-available sequencing data produced by Petropoulos et al. (2016)[663]. Quality control left a dataset of 1,525 cells (Supplementary Figure S6.3), the top 2,000 most variable genes were then identified (Supplementary Figure S6.4), and cells clustered using Seurat's graph-based clustering algorithm, resulting in the identification of 12 clusters (Figure 6.2A). Each cluster's top 50 marker genes were determined (Supplementary Table S6.1) and clusters were assigned identities using previously published marker genes (Supplementary Figure S6.1); as described previously[339], we find that cells within a given cluster show heterogeneity in their levels of marker gene expression.

The progression of development can be visualised as an arc from right to left in a UMAP non-linear dimensionality reduction plot (Figure 6.2A). Cell fate decisions can be located in the plot; the first dividing ICM from TED and the second dividing PE from epiblast. We also identify a cluster of cells in a transitory state between the morula and blastocyst. To verify identities assigned to clusters, trajectory analysis was completed using Monocle 3[713]. Monocle learns the sequence of transcriptomic changes that occur during cellular transitions in the dataset, estimates a developmental trajectory, orders cells along this

trajectory, and then measures each cell's progress through the trajectory using pseudotime (Figure 6.2B). We find that cluster pseudotimes estimated by Monocle are consistent with the identities assigned above. The Monocle trajectory recapitulates the ICM/TED cell fate decision and also identifies further lineage splits within the TED. Different TED subpopulations may represent a further cell fate specification event separating mural and polar TED[663]. The epiblast/PE lineage split is not present on the pseudotime plot because the branch to the PE is too short to avoid pruning, but both lineages are assigned the correct pseudotime.

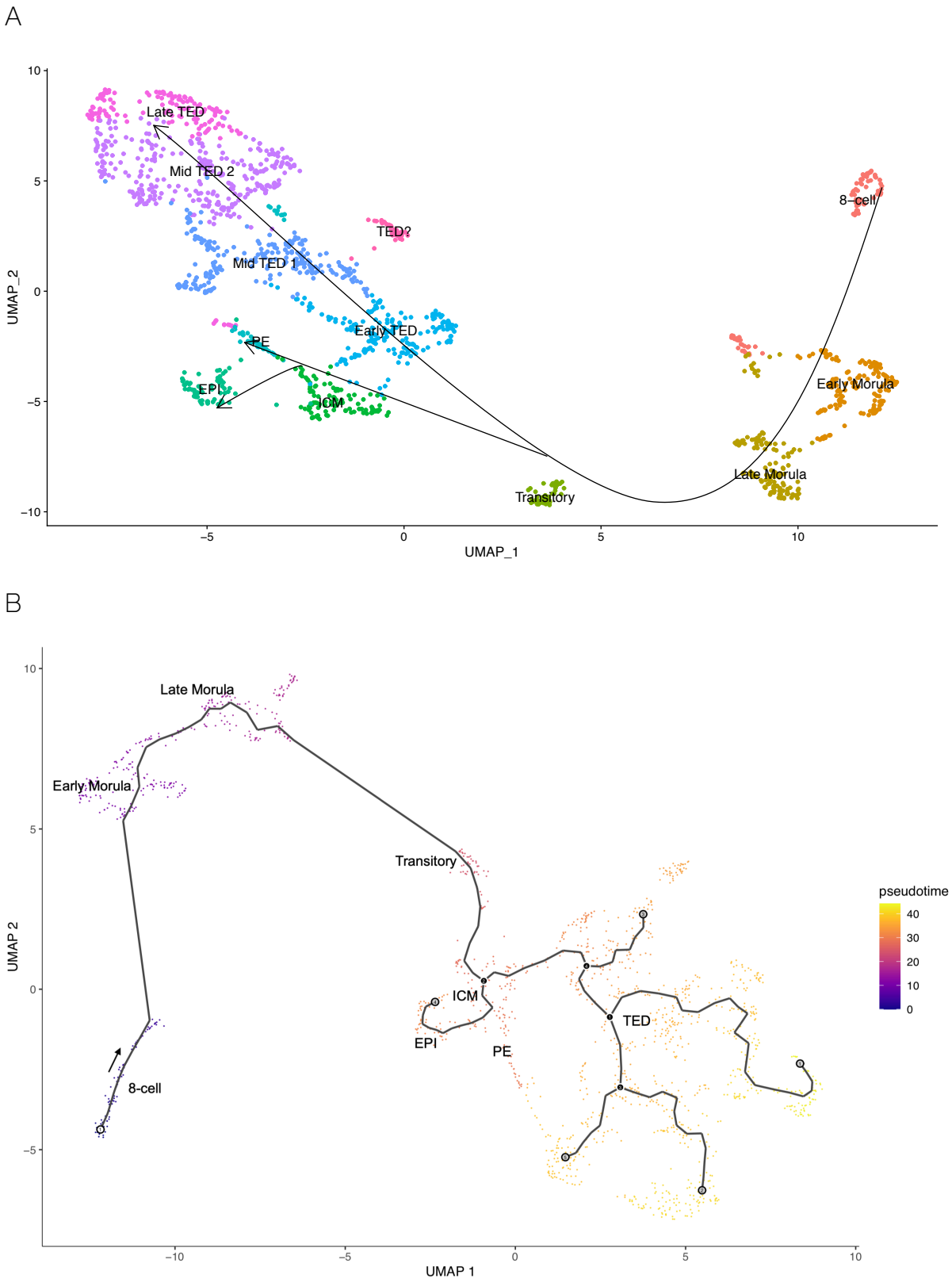


Figure 6.2 Clustering of 1,525 human preimplantation embryo cells based on their transcriptome. (A) UMAP visualisation of the transcriptomes of individual cells. Each point represents a cell, and cells with similar transcriptomes group together. Cells are coloured by their cluster identity. The arrows show the two cell fate decisions: the first

separating the TED from ICM, and the second separating the PE from epiblast. **(B)** Developmental trajectory of cells in the dataset estimated using Monocle. The black arrow shows the direction of the trajectory. Cells are coloured by their pseudotime, which represents their progression through the trajectory. Node 2 represents the split of ICM and TED. Abbreviations: EPI = epiblast; ICM = inner cell mass; PE = primitive endoderm; TED = trophectoderm.

6.5.2 Expression of human ETCHbox genes at single-cell resolution

Quantifying the expression of ETCHbox genes in the cells of each cluster reveals that they exhibit small but notable variations in their timing and specificity of expression (Figure 6.3A, B; Supplementary Table S6.2). The use of Monocle for developmental trajectory estimation facilitates analysis at a greater temporal resolution (Figure 6.3C). In this dataset, *LEUTX* is highly specific to 8-cell blastomeres, with minimal expression even by the early morula, and expression completely eliminated by the formation of the blastocyst. Like *LEUTX*, *TPRX1* expression peaks at the 8-cell stage, but continues into the morula. *DPRX* is also expressed at the 8-cell stage, with expression levels increasing as cells progress through this stage. Its expression peaks sharply in the early morula and persists into the late morula. Like *DPRX*, *ARGFX* expression increases throughout the 8-cell stage and peaks in the early morula but, importantly, it also persists into the blastocyst. Strikingly, *ARGFX* expression in the blastocyst is specific to the ICM (versus early TED, Wilcoxon rank sum test adjusted $p = 1.2 \times 10^{-28}$) and then the epiblast (versus PE, Mid TED 1 and Mid TED 2, Wilcoxon rank sum test adjusted $p = 3.16 \times 10^{-11}$, 6.73×10^{-32} and 3.72×10^{-38}, respectively), the lineages that go on to form the embryo proper, and it is not expressed in the PE or TED. No other ETCHbox gene shows lineage specificity in the blastocyst (Wilcoxon rank sum test adjusted $p > 0.05$). Overall, human ETCHbox genes peak in expression in the order *LEUTX*, *TPRX1*, *DPRX*, *ARGFX*; and *ARGFX* expression is ICM then epiblast-specific.

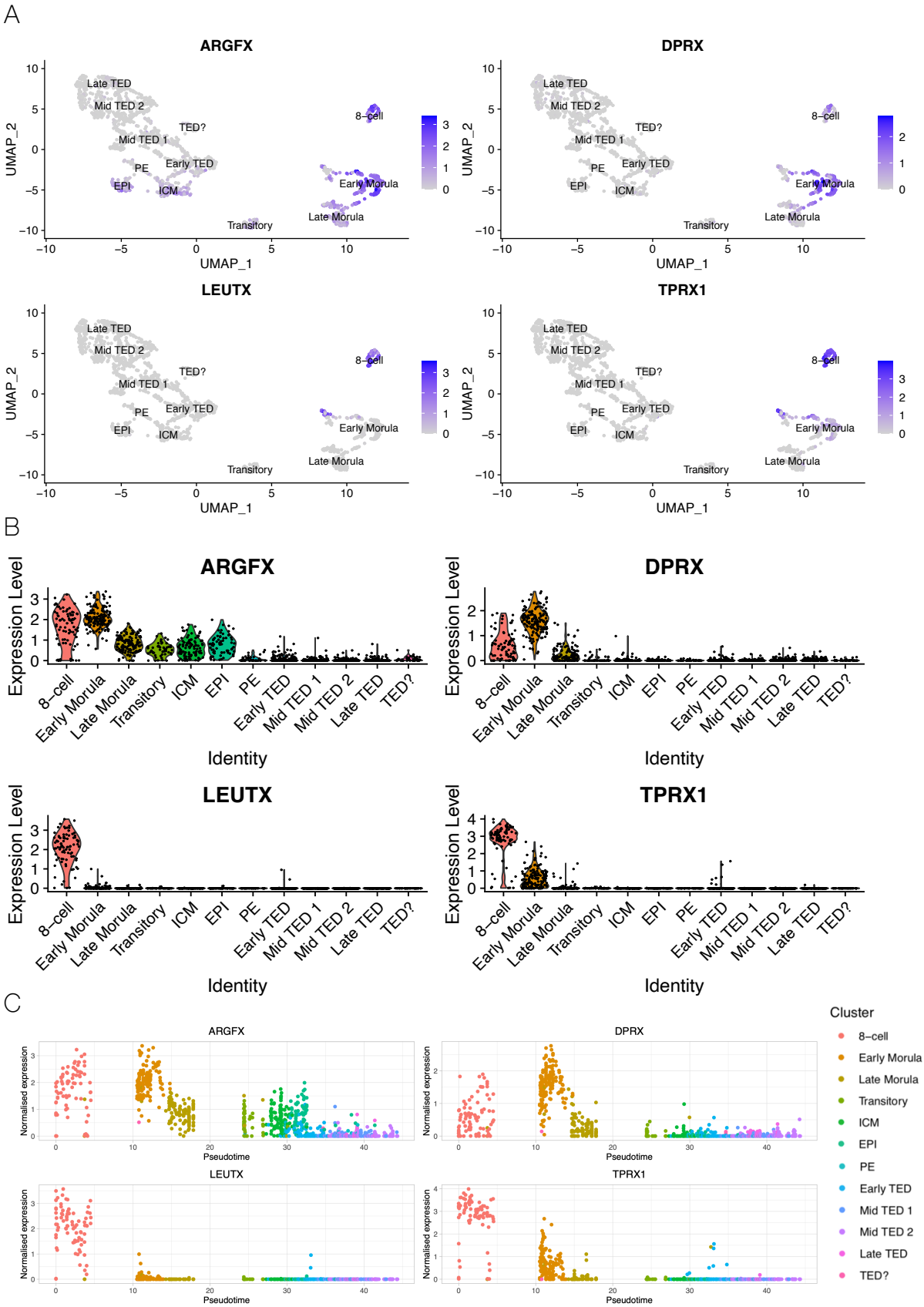


Figure 6.3 ETCHbox gene expression in human single-cell RNA-seq data. (A) UMAP plots with cells coloured by their expression of ETCHbox genes: grey = no expression, and darker colours indicate increased normalised expression levels. **(B)**

Scaled, normalised expression of ETCHbox genes in each cluster. Each point represents a cell. **(C)** Scaled and normalised expression of ETCHbox genes in cells ordered by pseudotime. Each point represents a cell. Cells are coloured by their cluster identity. Abbreviations: EPI = epiblast; ICM = inner cell mass; PE = primitive endoderm; TED = trophectoderm.

6.5.3 *ARGFX* expression peaks before that of ICM and epiblast lineage markers

Given the specificity of *ARGFX* to the ICM and then epiblast, we sought to understand whether its expression follows a similar pattern to marker genes of these lineages, which might suggest a similar role for *ARGFX*. To this end, we plotted the expression of *ARGFX* and key ICM and epiblast markers across pseudotime (Figure 6.4). As expected, the expression of ICM and epiblast markers (e.g. *NANOG* and *DPPA2*, respectively) peaks in the ICM and epiblast respectively. In contrast, *ARGFX* expression peaks in the early morula.

Since *ARGFX* expression peaks just prior to that of epiblast and ICM lineage markers, but continues specifically into those tissues, we hypothesised that *ARGFX* may be an early marker of ICM cell fate. This hypothesis predicts *ARGFX* to be expressed heterogeneously across cells in the morula, depending on their future cell fate, marking cells that are fated to become ICM. Heterogeneity of *ARGFX* expression is present in the morula (Figure 6.4, black box), making this plausible.

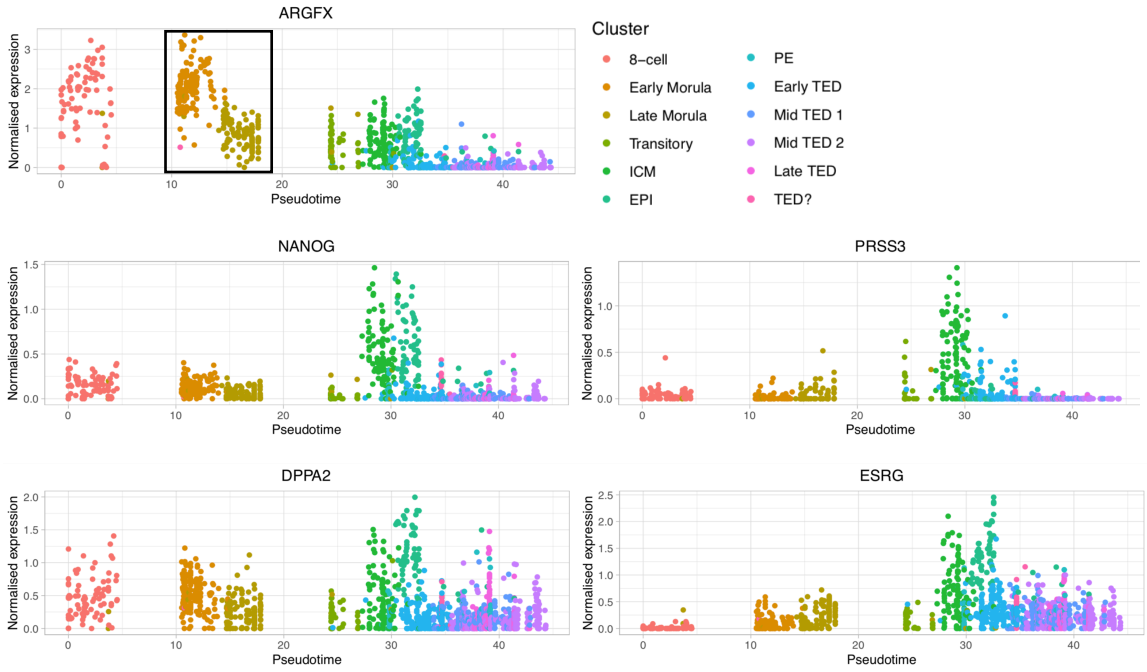


Figure 6.4 Expression of *ARGFX* and key ICM and epiblast marker genes. Plots show expression of *ARGFX* alongside ICM markers *NANOG* and *PRSS3* and epiblast markers *DPPA2* and *ESRG* in human preimplantation embryos. *NANOG* marks both ICM and epiblast. Normalised, scaled gene expression is plotted against pseudotime. Each dot represents a cell, and each cell is coloured by its cluster identity. The black box highlights heterogeneity of ARGFX expression in the morula. Abbreviations: EPI = epiblast; ICM = inner cell mass; PE = primitive endoderm; TED = trophectoderm.

6.5.4 Early lineage specification in the human morula?

Given the above result that ETCHbox expression occurs just prior to that of traditional lineage markers, we questioned whether a signal of lineage specification could be detected earlier, in the morula or transitory populations, and if so, whether ETCHbox genes would show any specificity. To test this, a data subset (n = 335 cells) was created including only morula and transitory populations, and then the clustering and UMAP processes repeated (Figure 6.5A). Three clusters were present in each of the early morula and late morula populations, but the transitory population remained as one cluster only. ICM, epiblast, PE and TED markers used above are expressed invariably across clusters at both the early and late morula stages, suggesting that lineage specification is not detectable at this stage (Supplementary Figure S6.5). This is consistent with previous reports that no lineage specification is detectable in the human morula[663]. Furthermore, ETCHbox expression is invariable across clusters at each

morula stage (i.e. invariable across early clusters A, B and C, and invariable across late clusters A, B and C), showing that the genes are not pre-specifying future cell lineages at this point in development (Figure 6.5B; Supplementary Table S6.3). Of all comparisons (Early A vs. Early B, Early A vs. Early C, Early B vs. Early C, Late A vs. Late B, Late A vs. Late C, Late B vs. Late C) for all ETCHbox genes, each shows no difference in expression except *DPRX* (higher in Early C than Early B, Wilcoxon rank sum test adjusted p = 3.23 x 10^{-3}) and *TPRX1* (higher in Early B than Early C, Wilcoxon rank sum test adjusted p = 1.08 x 10^{-9}). Violin plots (Figure 6.5B) suggest that these are unlikely to be biologically important differences. Overall, no early lineage segregation is detected in cells of the human morula.

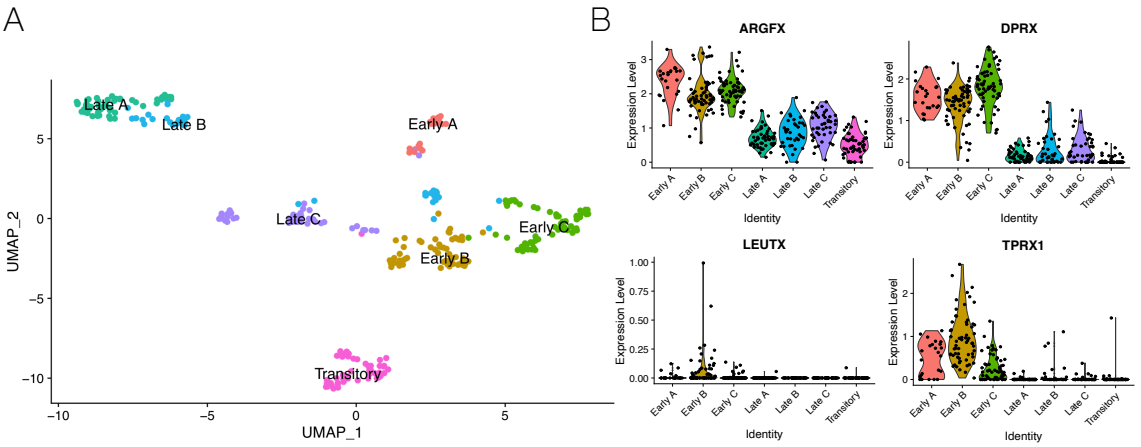


Figure 6.5 Clustering of human morula and transitory cells. (A) UMAP visualisation of the transcriptomes of morula and transitory cell populations. The early and late morula cells are each split into three clusters, while the transitory cell population shows less variation and remains as one cluster. Cells are coloured by their cluster identity. **(B)** Scaled, normalised expression of ETCHbox genes in each cluster of morula and transitory cells.

6.5.5 *ARGFX* upregulates ICM-specific genes

Given the specificity of *ARGFX* expression to the ICM and epiblast, we investigated whether its target genes also show such specificity. We use the *ARGFX*-upregulated gene set produced by Maeso et al. (2016)[90] and, to focus our search on realistic embryonic target genes, restrict the dataset to a specific group of key genes identified by the authors. Their data showed that the set of genes regulated by *ARGFX*, *LEUTX*

and *TPRX1* are enriched for genes with a specific expression pattern, in which genes have no or low expression from the oocyte to 4-cell stage, then are expressed highly in the 8-cell, morula and early blastocyst stages before expression again drops off (Figure 6.6A).

Genes upregulated by *ARGFX* and fitting this profile (n = 26 genes, Supplementary Table S6.4) were identified in the scRNA-seq data, then differential expression tests were performed to test for ICM and epiblast specificity. Of 26 genes upregulated by *ARGFX*, six genes showed specificity for ICM over Early TED (*ARRDC2*, *BAMBI*, *DDIT3*, *ID1*, *IER3*, *PMAIP1*), one for epiblast over PE (*PMAIP1*), seven for EPI over Mid TED 1 (*ARRDC2*, *BAMBI*, *DDIT3*, *ID1*, *IER3*, *PIM1*, *PMAIP1*) and four for EPI over Mid TED 2 (*ARRDC2*, *BAMBI*, *ID1*, *PMAIP1*) (Figure 6.6B, C; Supplementary Table S6.4). In contrast, only one gene showed specificity against ICM or EPI in any of these comparisons (*RHOB* for Early TED vs. ICM). This shows that key embryonic genes upregulated by ARGFX have higher expression in the ICM and epiblast, to which *ARGFX* is lineage-specific, than tissues in which *ARGFX* is not expressed. Overall, *ARGFX* upregulates lineage-specific genes in the ICM and epiblast.

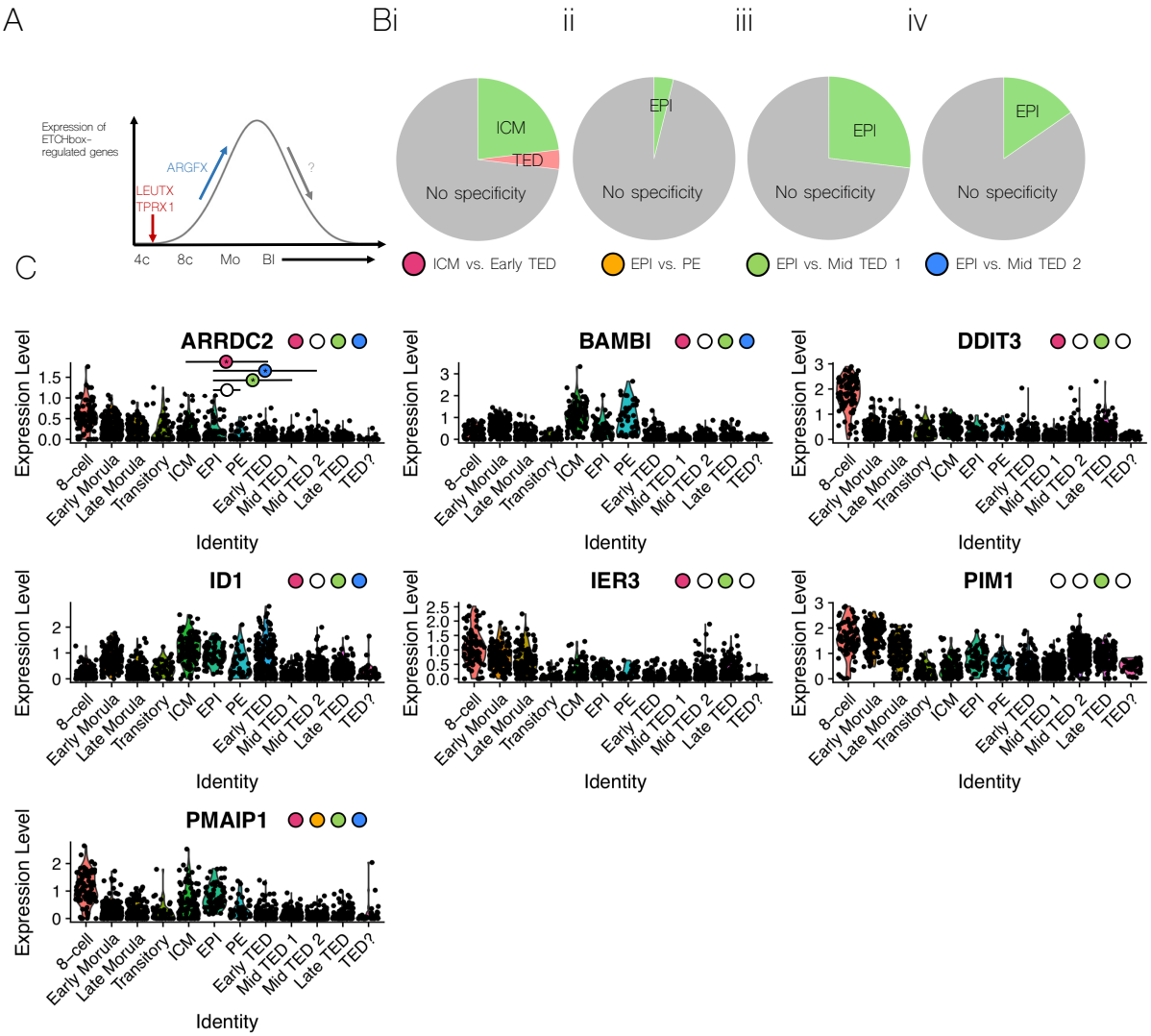


Figure 6.6 Lineage-specificity of genes regulated by ARGFX. (A) Representation of the temporal gene expression profile which is overrepresented in gene sets upregulated by ARGFX and downregulated by LEUTX and TPRX1[90]. **(B)** Pie charts showing percentage of ARGFX-upregulated genes that show lineage specificity in the blastocyst. **(C)** Expression of ARGFX-regulated lineage-specific genes in scRNA-seq data. Each plot has four circles above it, each representing a differential expression test. 1 (red) = ICM vs. Early TED; 2 (yellow) = EPI vs PE; 3 (green) = EPI vs Mid TED 1, 4 (blue) = EPI vs. Mid TED 2. Circles are coloured if the gene has higher expression in the ICM or epiblast (Wilcoxon rank sum test adjusted $p < 0.05$), and white if not. Colour-coded bars with asterisks indicating significance are shown in the first plot (*ARRDC2*) to facilitate understanding of the rest of the plots. For example, *ARRDC2* has higher expression in ICM than Early TED, in epiblast than Mid TED 1, and in epiblast than Mid TE 2, but not higher in epiblast than PE. Abbreviations: Bl = blastocyst; c = cell; EPI = epiblast; ICM = inner cell mass; Mo = morula; TED = trophectoderm.

6.5.6 GRN reconstruction reveals DUXA as a putative activator of *ARGFX* and *LEUTX*

Significant effort has been expended by myself and others attempting to elucidate the targets of ETCHbox genes' TF function. However, in order to facilitate a more rounded understanding of the role they play in preimplantation development, it is also necessary to uncover the factors that activate their expression. In short, what genes are upstream of ETCHbox genes in the cascades that control preimplantation development? To this end, I used the SCENIC pipeline[715] to perform GRN reconstruction from the human scRNA-seq data. Briefly, the GRNboost2 algorithm is implemented to find sets of genes that are coexpressed with TFs. These 'coexpression modules' are then pruned to include only direct targets through *cis*-regulatory motif analysis to produce bona fide regulons (one TF and the genes it directly regulates); the AUCell algorithm[715] then quantifies the activity of each regulon in each cell and uses this for clustering and t-SNE[718] non-linear dimensionality reduction processes.

For both *ARGFX* and *LEUTX*, a single TF is identified with high confidence as an activator in the human preimplantation embryo dataset. In both cases, this TF is DUXA, which acts via a highly similar binding site present multiple times in the region surrounding both genes' transcription start sites (TRAYNTAATCA, where R is either purine, Y is either pyrimidine and N is any nucleotide; Figure 6.7A). The DUXA regulon, consisting of genes directly activated by DUXA, contains 91 genes including *ARGFX* and *LEUTX* (Supplementary Table S6.5). The regulon is most active at the 8-cell stage, with activity decreasing throughout the morula stage and minimal upon blastocyst formation (Figure 6.7B, C). This is consistent with the expression timing of DUXA itself (Figure 6.7B), and the timing of expression of the putative DUXA-regulated genes (Supplementary Figure S6.6). Though the earliest cells in this dataset are drawn from the 8-cell stage, it is known from previous work that DUXA expression peaks between the 4 and 8-cell stages[94]. This is consistent with it being an activator of *ARGFX* and *LEUTX*, the expression of which peaks just after *DUXA*, at the 8-cell and morula stages

respectively (Figure 6.7D). Overall, these data suggest that DUXA activates *ARGFX* and *LEUTX* expression in the human preimplantation embryo.

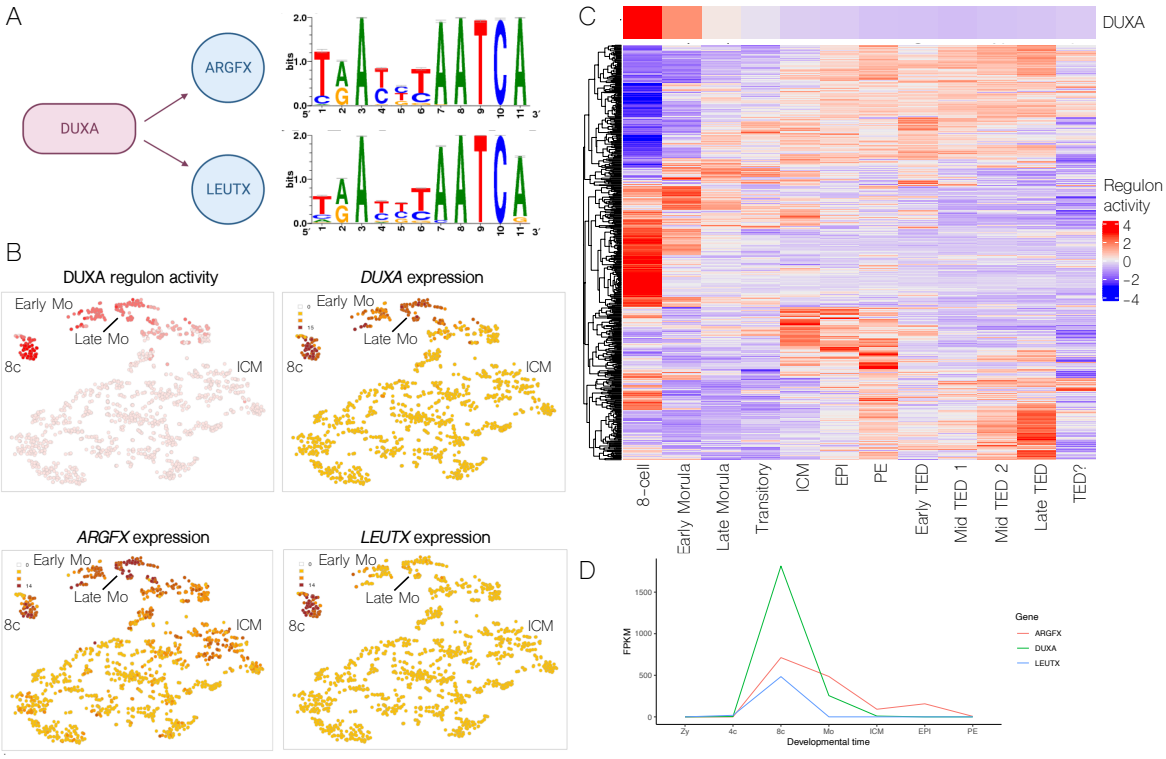


Figure 6.7 GRN reconstruction reveals DUXA as an activator of *ARGFX* and *LEUTX*. (A) Schematic of a section of the DUXA regulon determined by SCENIC. DUXA activates *ARGFX* and *LEUTX* expression in the preimplantation embryo. Motif logos show the binding motifs used by DUXA to bind the *ARGFX* and *LEUTX* loci. **(B)** t-SNE dimensionality reduction plots where each dot is a cell and cells with similar expression of regulons identified by SCENIC occur together. The first panel shows DUXA regulon activity (pale red = low activity, right red = high activity), and the following panels show *DUXA*, *ARGFX* and *LEUTX* gene expression (mRNA) levels, respectively (pale yellow = no expression, dark yellow/brown = high activity). **(C)** Activity of all regulons (n = 606) in each Seurat cluster. DUXA is displayed separately for clarity; it is most active in the 8-cell cluster, with activity decreasing over time and minimal by the onset of blastocyst formation. **(D)** Timing of *DUXA*, *ARGFX* and *LEUTX* expression during human embryonic development. Raw data for part D is taken from the Stirparo et al. (2018)[710] scRNA-seq dataset. Abbreviations: Bl = blastocyst; c = cell; EPI = epiblast; ICM = inner cell mass; Mo = morula; Oo = oocyte, PE = primitive endoderm; TED = trophectoderm.

6.5.7 Clustering of porcine preimplantation embryo cells

To understand whether the expression patterns observed in humans show conservation across mammals, we expanded our analysis to domestic pig (*Sus scrofa*). Pigs belong to the superorder Laurasiatheria; their distant relatedness to humans combined with the availability of high-quality omics data due to their agricultural importance makes them a good model to answer this question. Two single-cell datasets were available, one spanning from oocyte to early blastocyst[701], referred to herein as Dataset 1, and another covering morula to late blastocyst[699], referred to herein as Dataset 2, which contains a greater number of blastocyst cells. Following quality control, 87 and 219 cells were retained for the two datasets, respectively (Supplementary Figure S6.7), and the top 2,000 most variable genes (and Kallisto equivalence classes composed of two or more genes [see 6.4 Materials and methods]) identified (Supplementary Figure S6.8). Cells were then clustered using Seurat's graph-based clustering algorithm (Figure 6.8A, B); each cluster's top 50 marker genes were determined (Supplementary Tables S6.6, S6.7); and clusters were assigned identities using previously published marker genes (Supplementary Figures S6.9, S6.10).

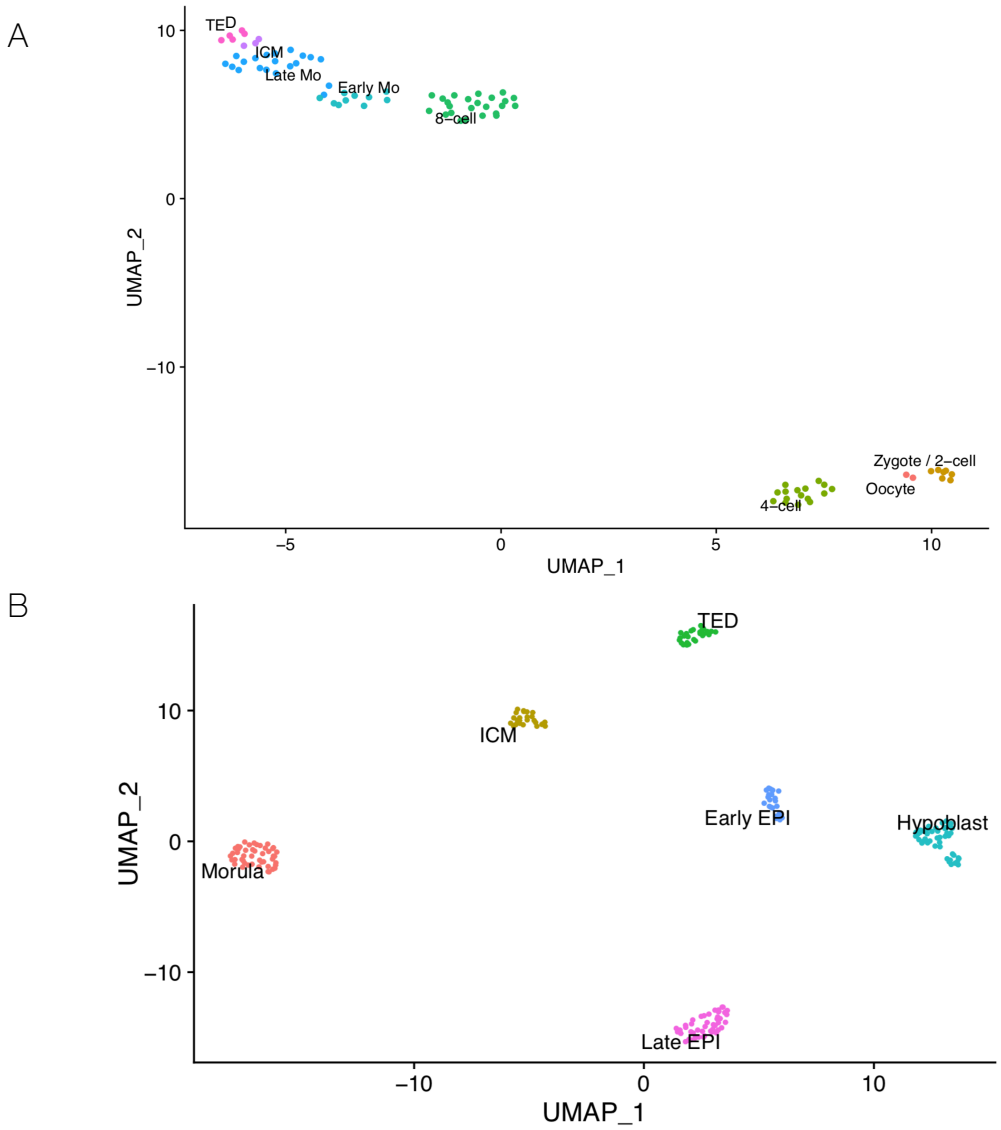


Figure 6.8 Clustering of porcine preimplantation embryo cells based on their transcriptome. Parts **(A)** and **(B)** correspond to Datasets 1 and 2, respectively**.** UMAP visualisation of the transcriptomes of individual cells. Cells are coloured by their cluster identity. Abbreviations: EPI = epiblast; ICM = inner cell mass; Mo = morula; TED = trophectoderm.

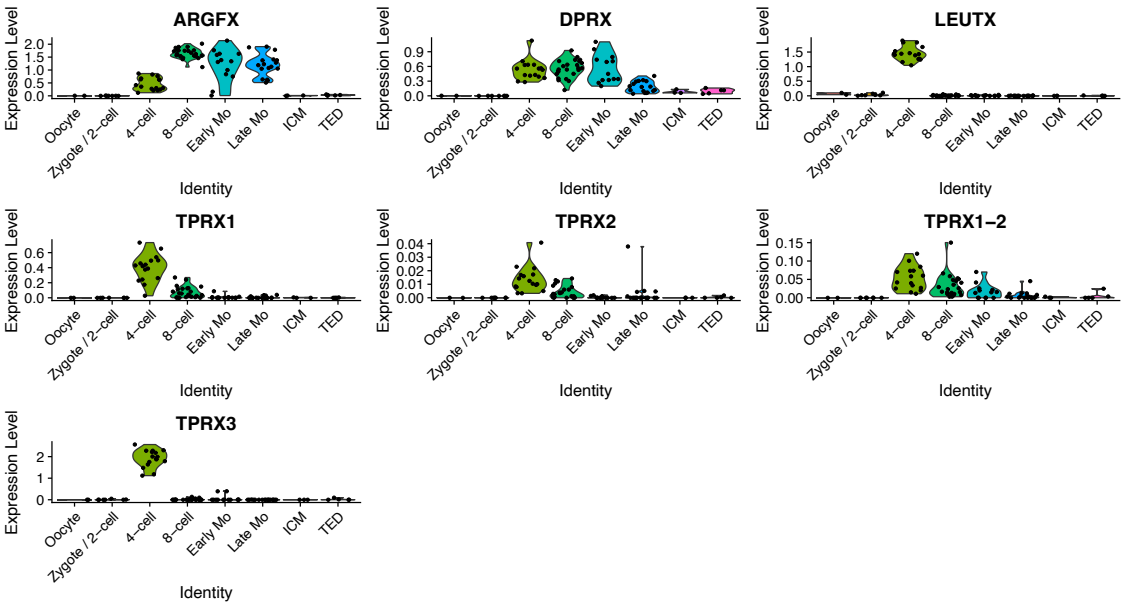
6.5.8 ETCHbox expression in the pig preimplantation embryo

Dataset 1 (Figure 6.9A, B; Supplementary Table S6.8) reveals that all porcine ETCHbox genes are simultaneously activated at the 4-cell stage of development, coinciding with the major wave of EGA[694,720]. *ARGFX* and *DPRX* are expressed until the morula stage, and then *ARGFX* expression is extinguished by the early blastocyst; *DPRX* expression may continue at a low level into the blastocyst, but the sample size is too small to confirm this. *TPRX1* and *TPRX2* expression peaks at the 4-cell stage, but persists at a lower

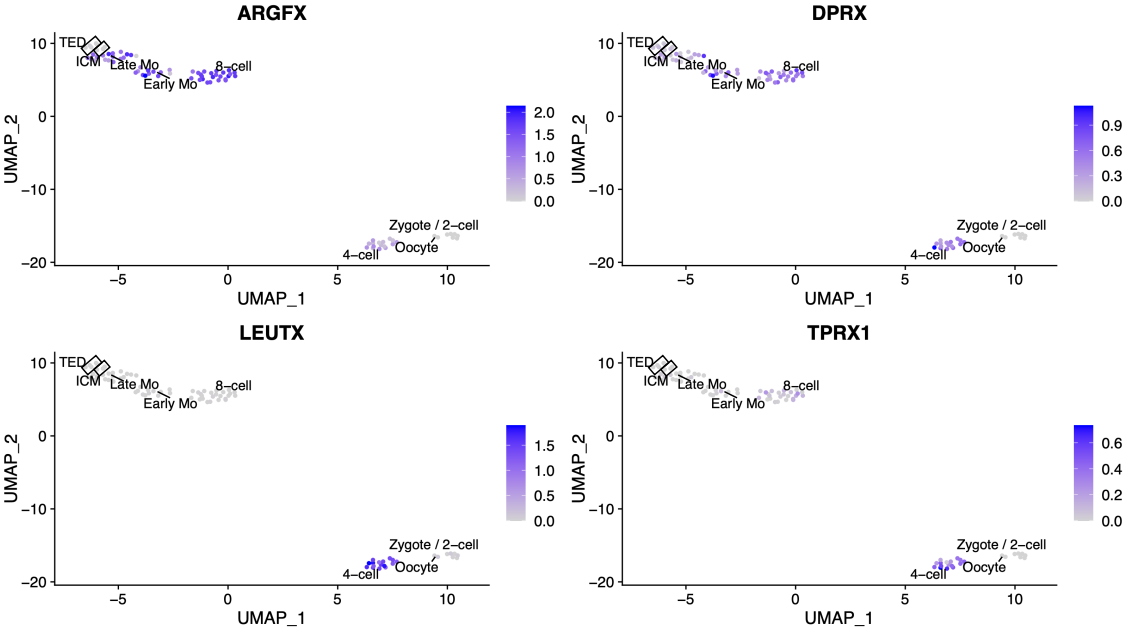
level into the 8-cell stage. *LEUTX* is highly specific to the 4-cell stage, similar to its 8-cell specificity in human. The Artiodactyla-specific *TPRX* duplicate, *TPRX3*, also exhibits this highly 4-cell specific expression, which is more similar to that of *LEUTX* than it is to its paralogues *TPRX1* or *TPRX2*.

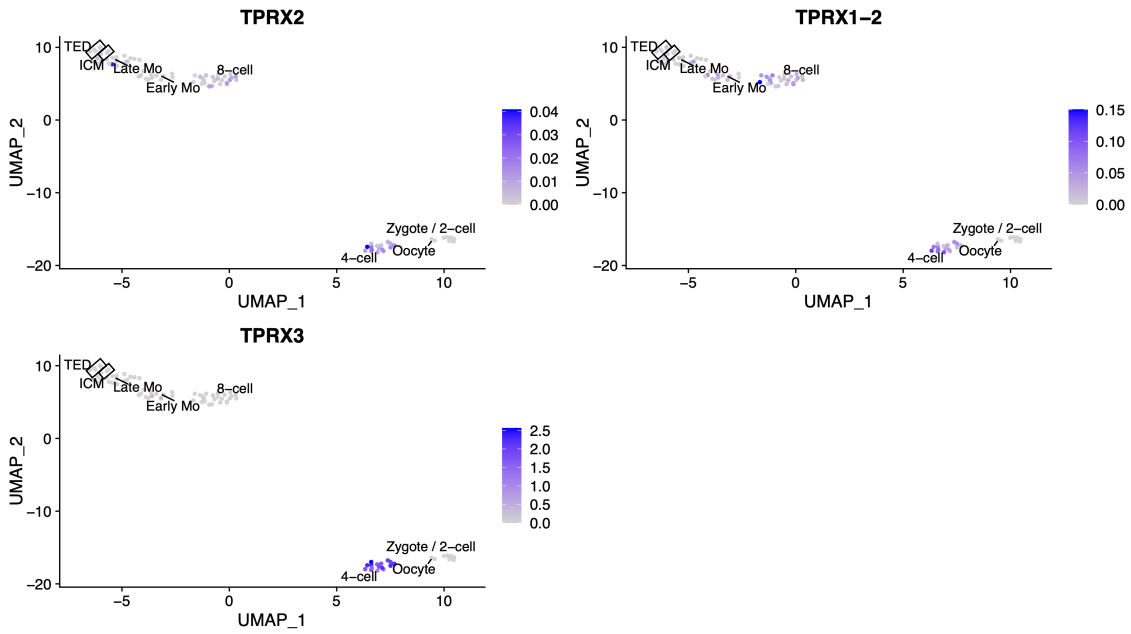
Dataset 2 (Figure 6.9C, D; Supplementary Table S6.9), in which sampling begins in the morula, supports the results of Dataset 1: minimal *LEUTX*, *TPRX1*, *TPRX2* and *TPRX3* expression is detected, as their expression period has ended, but strong *ARGFX* and *DPRX* expression is observed in the morula. Dataset 2 has deeper sampling of the blastocyst, and confirms that, with the exception of *DPRX*, all ETCHbox expression is eliminated by the end of the morula stage. The Dataset again suggests that *DPRX* expression does continue at a low level into the early blastocyst. At the peak of ETCHbox gene expression, the genes show clear heterogeneity in their expression between cells (Supplementary Figure S6.11), suggesting the potential for lineage specificity. However, the expression period of all genes except *DPRX* has ended by the onset of blastocyst formation, and therefore shows no clear lineage specificity in pigs, in contrast to humans.

A

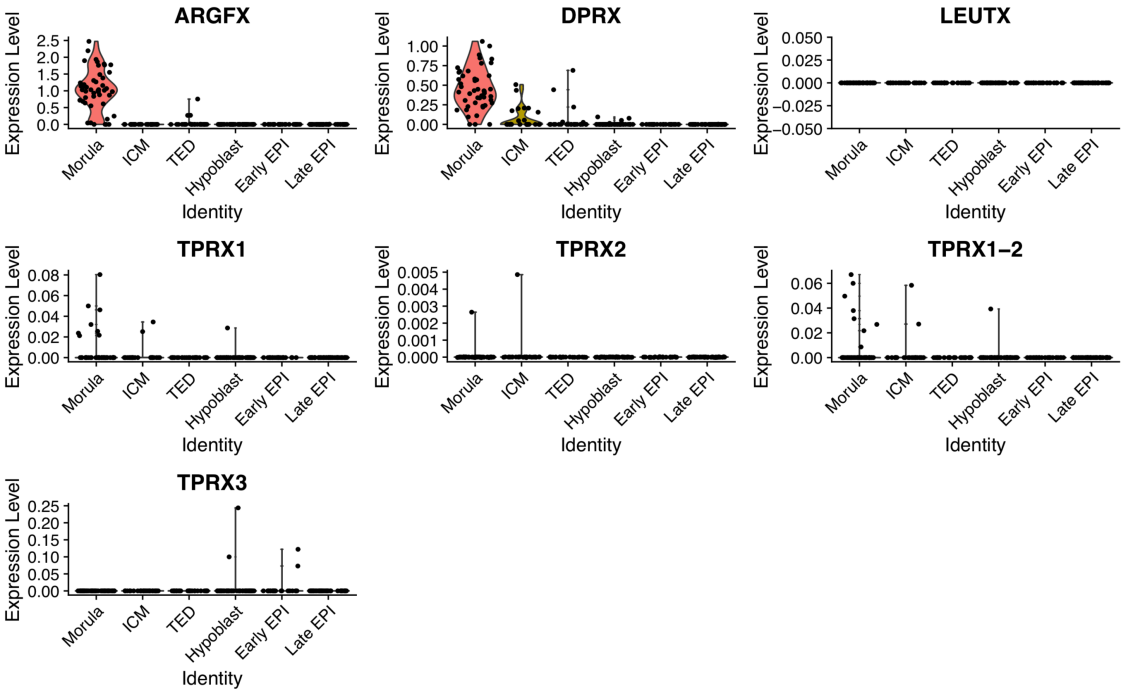

ARGFX
DPRX
LEUTX
TPRX1
TPRX2
TPRX1–2
TPRX3
Expression Level
Identity
Oocyte
Zygote / 2–cell
4–cell
8–cell
Early Mo
Late Mo
ICM
TED

B


ARGFX
DPRX
LEUTX
TPRX1
UMAP_1
UMAP_2
TED
ICM
Late Mo
Early Mo
8–cell
Zygote / 2–cell
4–cell
Oocyte


TPRX2
TPRX1–2
TPRX3
UMAP_1
UMAP_2
TED
ICM
Late Mo
Early Mo
8–cell
4–cell
Zygote / 2–cell
Oocyte

C


ARGFX
DPRX
LEUTX
TPRX1
TPRX2
TPRX1–2
TPRX3
Expression Level
Identity
Morula
ICM
TED
Hypoblast
Early EPI
Late EPI

D

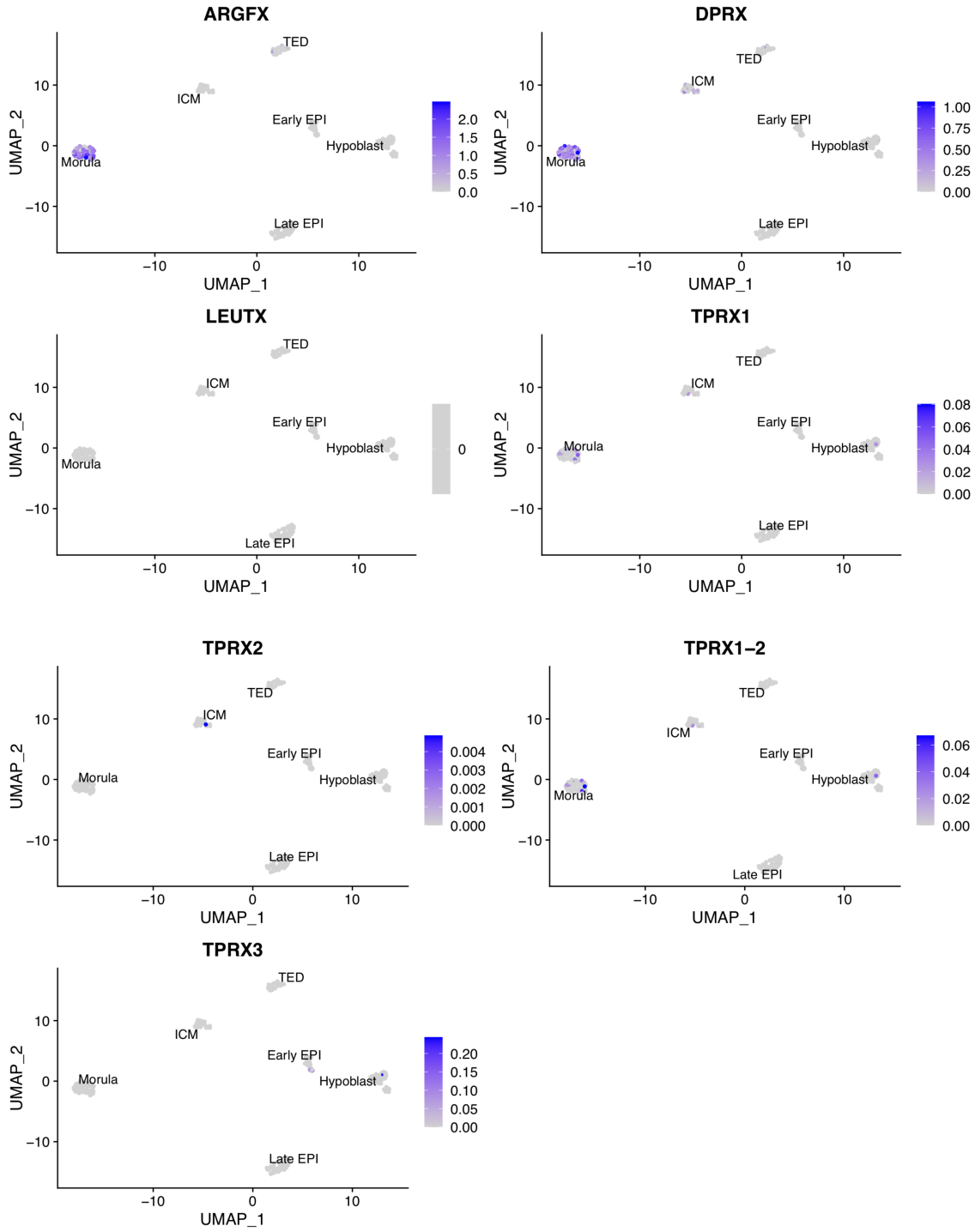


Figure 6.9 ETCHbox expression in porcine scRNA-seq data. In all plots, TPRX1-2 represents the Kallisto equivalence class containing reads that map to both *TPRX1* and *TPRX2*; this is necessary due to the genes' highly similar sequences. **(A)** Dataset 1: scaled, normalised expression of ETCHbox genes in each cluster. **(B)** Dataset 1: UMAP plots. Colours show expression of the ETCHbox genes: grey = no expression, and darker colours indicate increased normalised expression levels. **(C)** Dataset 2: scaled, normalised expression of ETCHbox genes in each cluster. **(D)** Dataset 2: UMAP plots. Colours show expression of the ETCHbox genes: grey = no expression, and darker

colours indicate increased normalised expression levels. Abbreviations: EPI = epiblast; ICM = inner cell mass; Mo = morula; TED = trophectoderm.

6.5.9 Early lineage specification in the porcine morula?

Kong et al. (2020)[701] and more recently Zhi et al. (2022)[721] have suggested that three sub-clusters are detectable within morula cells in pig: those fated to become TED, those fated to become ICM, and those whose fates remain undetermined, which express 8-cell marker genes to higher levels. Given our above finding that *ARGFX* and *DPRX* are highly expressed in the morula with some cell-cell variation, we hypothesised that they might show higher expression in the cells fated to become ICM, since *ARGFX* shows ICM specificity in humans. To test this, we created a morula-only subset of each Dataset and then repeated the clustering and UMAP processes.

In Dataset 1, our clustering separates the early morula cells expressing 8-cell markers from later cells, but does not separate ICM and TED-fated clusters (Supplementary Figure S6.12). No evidence for early lineage specification is found, though this could be due to the very small sample size. In addition, ETCHbox genes show no coexpression with ICM or TE marker genes. In Dataset 2, however, ICM- and TED-fated clusters can be tentatively assigned based on the marker genes published by Zhi et al. (2022)[721] (Figure 6.10). Though the small sample size and early stage of specification reduce the certainty of these identities, the expression of *DAB2* (TED marker) and *PDGFRA* (ICM marker) in the putative TED- and ICM-fated clusters is similar to that observed previously in specified morula cells[721]. However, neither *ARGFX* nor *DPRX* are significantly DE between the ICM- and TED-fated clusters (Wilcoxon rank sum tests $p > 0.05$), so we find no evidence that either gene is involved in lineage specification at this stage. Despite this, it is possible that the sample size is simply too small to detect this pattern, and future analysis of a dataset with a larger sample size is required to confirm these results.

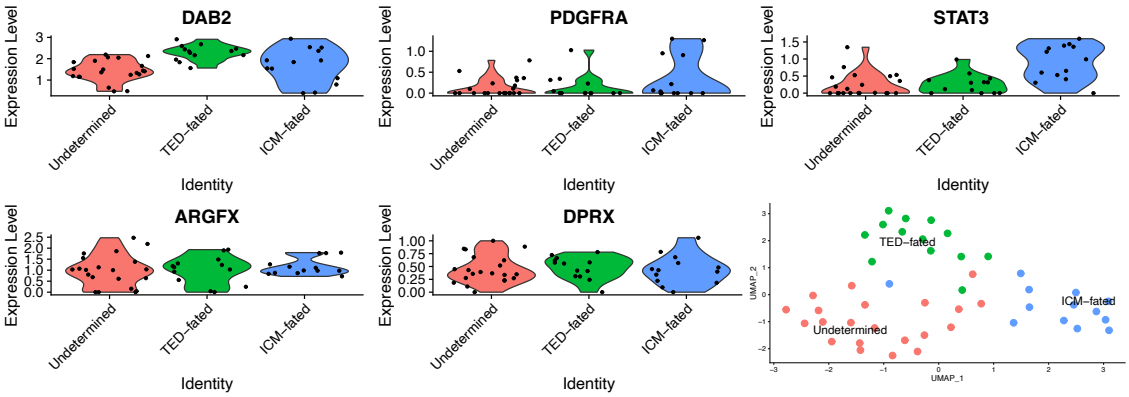


Figure 6.10 Lineage specification in the porcine morula (Dataset 2). UMAP plot separates clusters of cells fated to become ICM and TED from those whose fate is yet to be determined. Violin plots show *DAB2* and *PDGFRA*, identified by Zhi et al. (2022)[721] as early lineage markers for TED and ICM respectively and *STAT3*, another ICM marker. ETCHbox genes *ARGFX* and *DPRX* are also shown; neither show specificity to ICM- or TED-fated clusters.

6.5.10 Clustering of mouse preimplantation embryo cells

To understand ETCHbox expression in the mouse embryo, we utilised a published single-cell dataset of 317 cells spanning the zygote to the early blastocyst[700]. Following quality control (Supplementary Figure S6.13), the top 5,000 most variable genes were identified (Supplementary Figure S6.14) and then cells clustered using Seurat (Figure 6.11). Each cluster's top 50 marker genes were determined (Supplementary Table S6.10), and ICM and TED blastocyst cells identified by known lineage markers (Supplementary Figure S6.15). As reported previously for this dataset[339], we find that the ICM has not yet undergone the specification of epiblast and PE. Indeed, many ICM and TED markers are not exclusive to their lineage and are expressed at a low level in cells of the other lineage, suggesting that the cells were sampled at an early stage of specification. The dataset includes a small number of liver and fibroblast cells, which we retain as a negative control for ETCHbox expression. Clustering accurately resolves cells of each developmental stage, and reflects known features of mouse development. For example, early 2-cell blastomeres cluster with zygotes, while mid-late 2-cell blastomeres form a separate cluster, closer to the 4-cell cluster. This reflects the fact that the major

EGA event in mice occurs at the mid to late 2-cell stage[722]: the 2-cell blastomeres clustering with the zygote are pre-EGA and those nearer the 4-cell cluster are post-EGA.

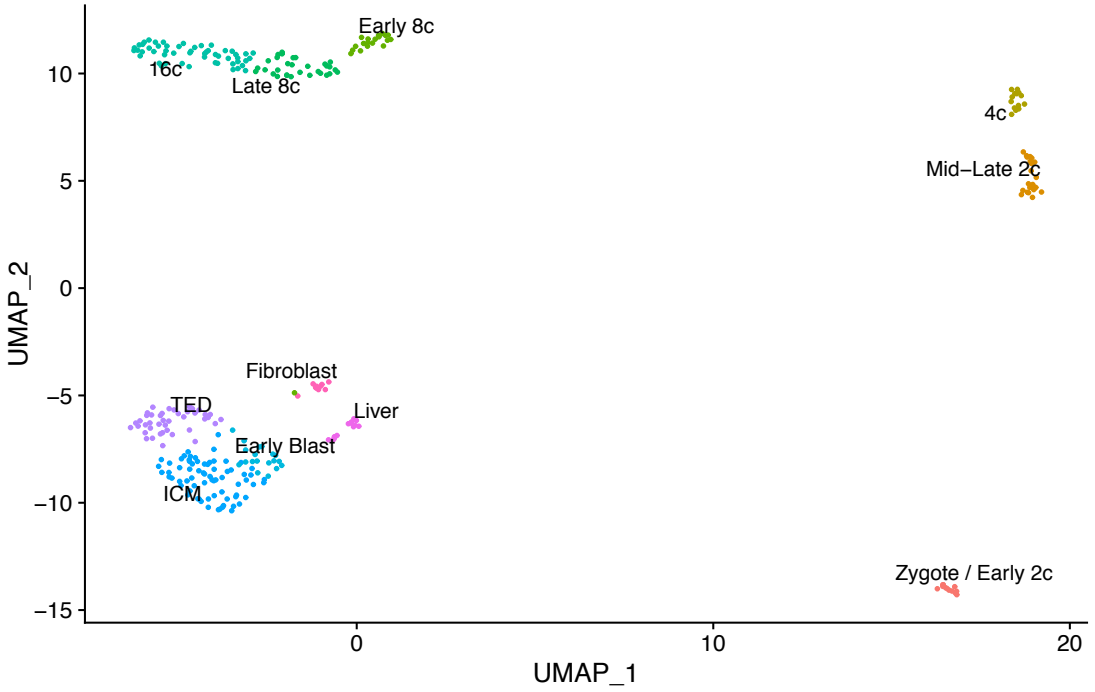


Figure 6.11 Clustering of mouse preimplantation embryo cells based on their transcriptome. UMAP visualisation of the transcriptomes of individual cells. Each point represents a cell, and cells with similar transcriptomes group together. Cells are coloured by their cluster identity. Abbreviations: Blast = blastocyst; c = cell; ICM = inner cell mass, TED = trophectoderm.

6.5.11 Expression profiles of highly duplicated ETCHbox genes

The availability of single-cell transcriptomes of mouse preimplantation embryos provides an opportunity to look at the expression of ETCHbox genes in a species in which they have been highly duplicated. The mouse ETCHbox repertoire consists of two *Crxos* (*Tprx1*) loci (producing 3 transcripts) and 66 *Obox* (*Tprx2*) loci, 28 of which encode a complete homeobox, while *Argfx*, *Dprx* and *Pargfx* have been completely lost and *Leutx* is a highly degraded pseudogene[90,96]. This situation offers the chance to ask whether the genes present in these large tandem arrays show temporal or spatial compartmentalisation of their expression.

We profiled the expression of every *Obox* gene with a complete homeobox. The sequences of *OboxA*, *OboxB* and *OboxC* genes are divergent enough that their expression profiles can be determined based on uniquely mapped reads, and each of these genes conforms to one of four broad expression profiles (Figure 6.12). Notably, we find that genes that have the most similar sequences also have the most similar expression profiles. Focusing on the *OboxA* genes, *OboxA1* is the outgroup and has the most divergent expression profile, being highly expressed from EGA until the 16-cell stage, while all other *OboxA* genes are expressed only up to the 4-cell stage. The remaining *OboxA* genes form two clades: clade A, consisting of *OboxA2*, *OboxA4*, *OboxA6* and *OboxA12*, is made up of genes that are most highly expressed in the zygote/pre-EGA 2-cell cluster, with expression decreasing by the mid-late 2-cell stage and minimal by the 4-cell stage. The outgroup of clade B, *OboxA8*, exhibits the same expression pattern as the genes of clade A, but the other four genes (*OboxA3*, *OboxA5*, *OboxA7*, *OboxA10*) have undergone a change of pattern and are not expressed pre-EGA, with expression highly activated at EGA, lower by the 4-cell stage and absent by the 8-cell stage. Overall, *Obox* genes conform to one of several distinct expression profiles, with highly similar sequences tending to exhibit the most similar expression patterns; those with more divergent sequences (*OboxA1*, *OboxC*) have more different expression profiles.

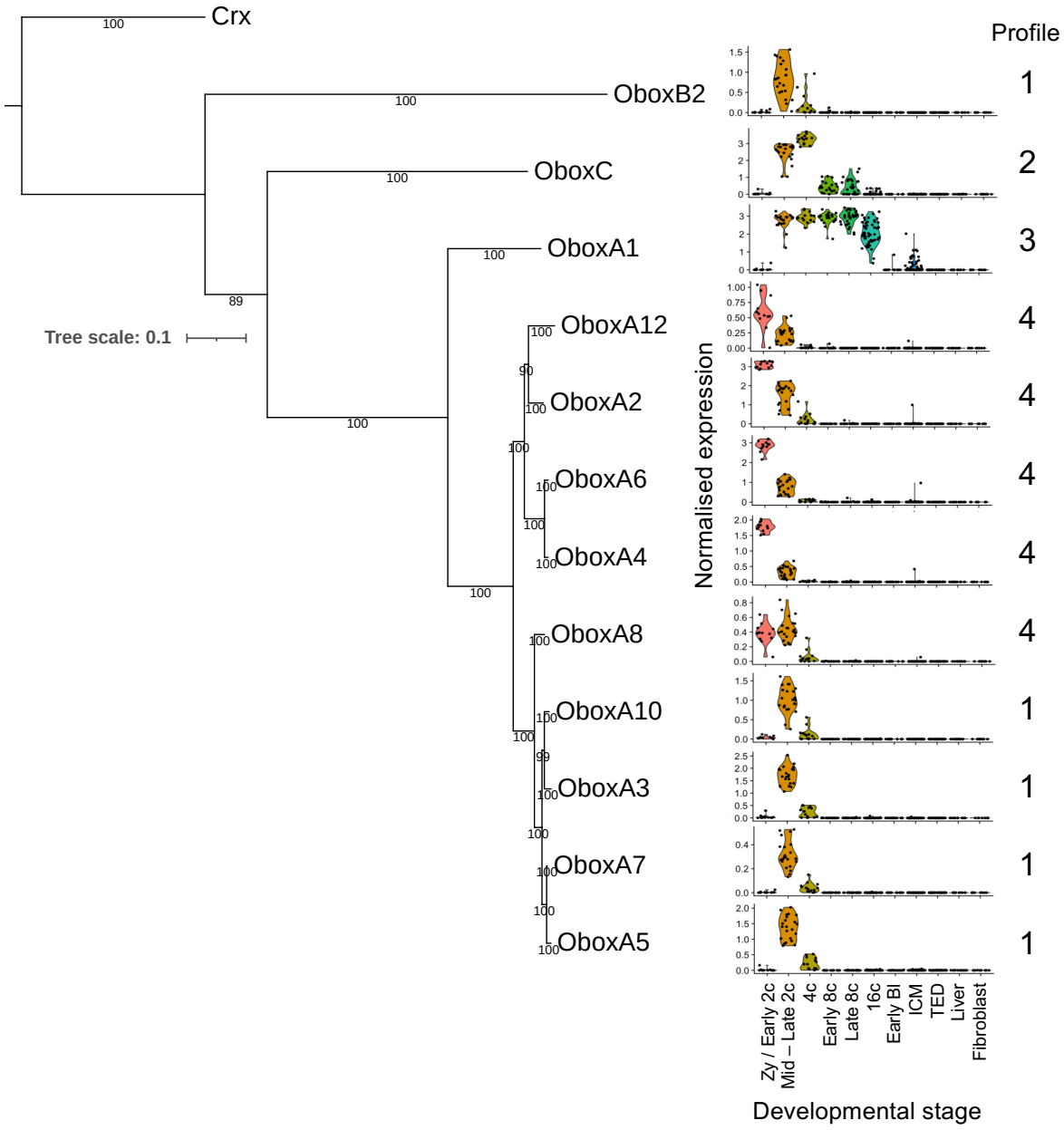
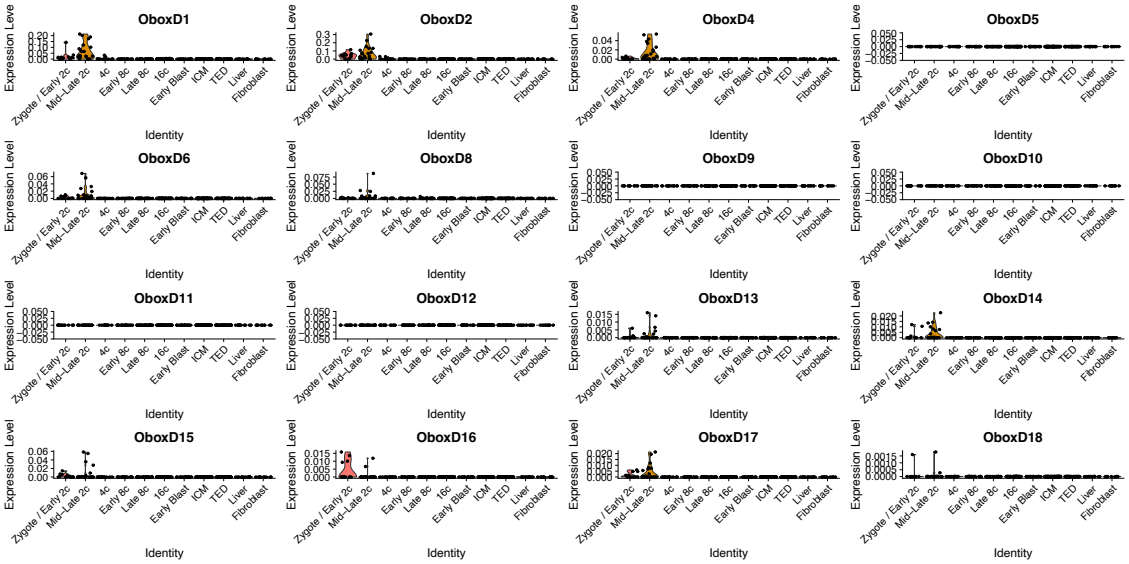


Figure 6.12 Expression of *Obox* genes in scRNA-seq data. Figure shows Bayesian gene tree of all mouse genes from the *OboxA*, *OboxB* and *OboxC* groups which encode a complete homeobox, accompanied by violin plots showing each gene's expression in each cluster of the single-cell dataset. Numbers on the tree show posterior probabilities; *Crx* is the outgroup. Violin plots show scaled, normalised expression level. Genes are assigned an expression profile number from 1 to 4 based on which one of the four profiles they conform to. Abbreviations: Bl = blastocyst; c = cell; ICM = inner cell mass, TED = trophectoderm; Zy = zygote.

The 16 putatively functional *OboxD* genes (those with a complete homeobox) show between 97.7 and 99.9% identity to the *OboxD* consensus sequence (Figure 6.13); this extreme similarity means that most reads originating from them cannot be mapped back to one specific locus. As such, many of the Kallisto equivalence classes (ECs) containing reads mapping exclusively to one *OboxD* gene contain very few reads, and indeed five such classes (*OboxD5*, *OboxD9*, *OboxD10*, *OboxD11*, *OboxD12*) contain no uniquely mapped reads (Figure 6.13). Since these are the genes with the least divergence from consensus (99.9% identity), this is likely to be because all reads originating from these loci map to multiple sequences, not because they are not expressed. Indeed, 664 reads map to an EC containing only *OboxD9*, *OboxD10*, *OboxD11* and *OboxD12,* showing that at least one of these genes with no unique reads must be expressed. While these data are imperfect, we can conclude that at least 11 of the 16 putatively functional *OboxD* genes are expressed in the preimplantation embryo. An expression profile of low expression pre-EGA, a peak in the mid-late 2-cell stage and then nothing for the rest of preimplantation development is followed by all of these genes that have enough reads uniquely mapped to determine a putative expression profile. *OboxD1* and *OboxD2* have the most uniquely mapped reads, which corresponds to them being the most divergent *OboxD* genes, and both of these genes show this profile.



OboxD gene	1	2	4	5	6	8	9	10	11	12	13	14	15	16	17	18
Identity to consensus (%)	98.50	97.65	99.67	99.94	99.74	99.87	99.94	99.94	99.94	99.94	99.94	99.87	99.74	99.54	99.74	99.74

Figure 6.13 Expression of *OboxD* genes in scRNA-seq data using uniquely mapped reads only. Table shows percent identities of *OboxD* genes encoding complete homeodomains to the *OboxD* consensus sequence. Plots show expression of each *OboxD* gene in each cluster of the dataset; each dot is a single cell. Plots include only reads mapping uniquely to that gene. Abbreviations: Blast = blastocyst; c = cell; ICM = inner cell mass, TED = trophectoderm.

To extract further meaning from this data, we identified every EC containing reads mapping to one or more *OboxD* genes and no other genes (n = 76, Supplementary Table S6.11). We eliminated those containing fewer than 100 reads, leaving 38 ECs. Of these classes, 11 contain more reads than any of the uniquely mapped classes, and 31 contain more reads than all except *OboxD1* and *OboxD2*, the two most divergent genes. This confirms the above hypothesis that *OboxD* gene expression was underestimated due to multi-mapping of reads. With only a single exception (EC 104306), the 38 ECs containing multi-mapped reads exhibit the same expression pattern described above for the *OboxD* genes based on only uniquely mapped reads: low expression in the zygote, a peak at the mid to late 2-cell stage and then nothing for the rest of preimplantation development (Figure 6.14). Even EC 104306 has expression at the same time, it is just higher at the zygote/early 2-cell stage than the mid-late 2-cell stage; moreover, this is the class with

the second-fewest reads, so the pattern is less reliable. Overall, it can be concluded that, during preimplantation development, *OboxD* genes are exclusively expressed from the zygote to the late 2-cell stage. Since each of the uniquely mapped ECs and all of the multi-mapped ECs exhibit the same pattern, with an expression peak immediately post-EGA, we can infer that most *OboxD* genes are expressed with this temporal expression profile. The presence of transcripts before EGA suggests maternal expression, but the peak post-EGA suggests that expression is upregulated at EGA.



Figure 6.14 Expression profiles of Kallisto ECs containing multiple *OboxD* genes. In this figure, it is not the individual details that are important but the overall pattern. Each plot shows the expression of an EC containing reads that map to multiple *OboxD* genes. All ECs have *OboxD* expression only in the zygote and 2-cell stages. ECs are sorted by their number of reads, with the class with the most reads at the top left. The *OboxD* genes corresponding to each EC are presented as Supplementary Table S6.11. Abbreviations: Blast = blastocyst; c = cell; ICM = inner cell mass, TED = trophectoderm.

6.5.12 Lineage specificity of mouse ETCHbox genes

One key question this chapter aims to answer is whether ETCHbox expression in the preimplantation embryo is lineage-specific. In mice, the expression of all three *Crxos* transcripts plus one *Obox* gene, *OboxA1*, continues into the blastocyst and therefore has the potential for lineage-specificity. There are two *Crxos* genes (5' and 3'), plus an extra transcript isoform that includes the homeoboxes from both genes (5'-3') (Figure 6.15, inset), and this arrangement means that many RNA-seq reads map ambiguously to both the 5'-3' and 5' transcripts or 5'-3' and 3' transcripts. I report expression profiles for reads that map uniquely to each transcript plus those of multi-mapped reads separately, to give a more accurate picture of expression dynamics. Data show no lineage specificity of any of the five *Crxos* read classes; all are expressed at an equal level in the ICM and TED (Figure 6.15, Wilcoxon rank sum test adjusted $p > 0.05$). However, though *OboxA1* expression is at a low level by the formation of the blastocyst, those cells that retain expression are exclusively ICM cells. This is not a significant difference after multiple comparison correction (Figure 6.15, Wilcoxon rank sum test adjusted $p = 0.078$), but it justifies future analysis with a larger dataset to further probe this effect.

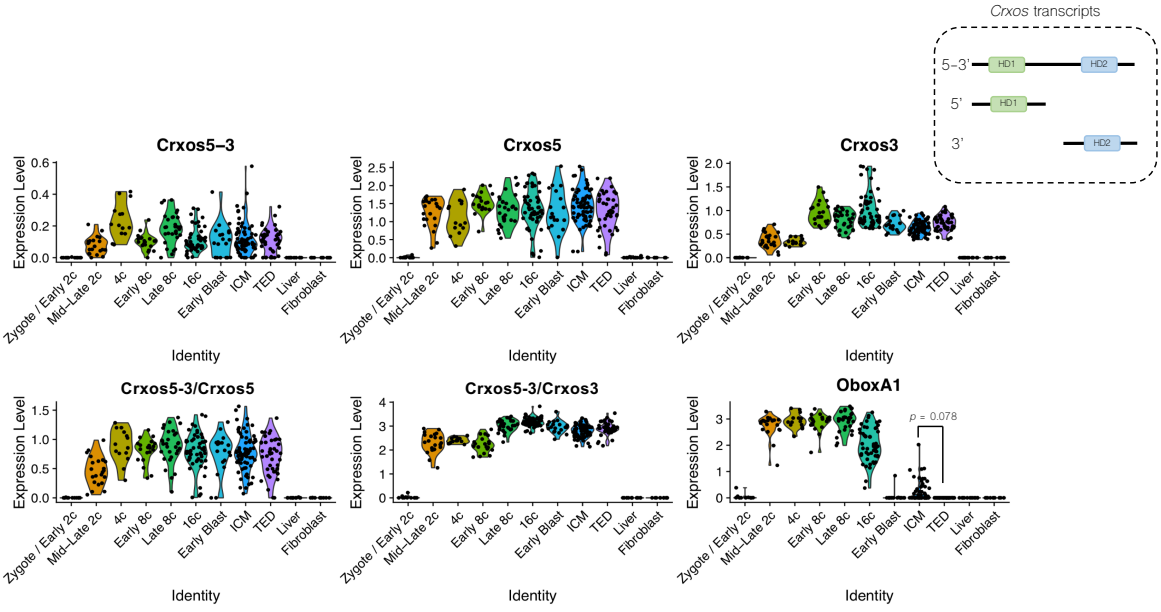


Figure 6.15 Expression of *Crxos* and *OboxA1* genes in the mouse preimplantation embryo. Plots show the normalised expression level of cells in each cluster. The plot labeled Crxos5-3/Crxos5 consists of reads that ambiguously map to either the 5'-3' transcript or the 5' transcript, and the Crxos5-3/Crxos3 plot consists of reads that map

to both the 5'-3' transcript and the 3' transcript. Inset is a diagram showing the arrangement of the two *Crxos* loci: The 5'-3' transcript crosses both loci, the 5' transcript is transcribed exclusively from the 5' locus and the 3' transcript from the 3' locus.

6.6 Discussion

The expression of ETCHbox genes occurs at a tumultuous point in mammalian development, as the very first cell fate decision of an individual's life specifies the cells that will form the ICM and subsequently the embryo itself. Given the fact that both ETCHbox genes themselves and the ICM are synapomorphies of the Eutheria, and based on other cases of lineage-specific TFs underlying novel features of early embryonic development[50], I hypothesised that ETCHbox proteins may play a role in ICM formation or lineage specification. In this chapter, I utilised scRNA-seq data to explore this potential link.

My results show that *ARGFX* is ICM- and then epiblast-specific, supporting a previous result[663], but take this further by suggesting that this is unique amongst human ETCHbox genes. Looking deeper into the properties of human ARGFX, I found that it upregulates several key ICM-specific genes. *BAMBI* is a particularly notable target; it is a known ICM marker[340,663], and is a negative regulator of TGF-β signalling involved in the maintenance of expression of the core pluripotency factor *NANOG* in the ICM[339]. This suggests a role for ARGFX in maintaining *NANOG* expression, and therefore in pluripotency. The specificity of *ARGFX* to the ICM, combined with the fact that it upregulates other ICM-specific genes such as *BAMBI*, suggests a role in ICM formation and possibly the pluripotent state of the ICM. This provides a possible explanation as to why multiple genes in eutherians (ETCHbox) are required for a role ably performed by CRX alone in marsupials: eutherians' more complex blastocyst structure, including the eutherian innovation of an ICM at its core, requires more than one gene for its development. It is tantalising that in pig (*DPRX*) and mouse (*OboxA1*) there are also ETCHbox genes with

hints at ICM specificity, but larger datasets are needed to test this. Overall, my data supports the idea that the lineage-specific ETCHbox genes are involved in the development of the novel eutherian ICM.

Another key question I aimed to answer is whether the genes' expression timing is conserved across species. To explore this, as in previous chapters, I compared expression across members of the superorders Euarchontoglires and Laurasiatheria. From this an intriguing pattern emerges, with the expression of *ARGFX*, *DPRX*, *LEUTX*, *TPRX1* and *TPRX2* all peaking one stage later in human than in pig (Figure 6.16). In humans, the major activation of ETCHbox genes occurs at the 8-cell stage, but in pig it is induced one cell division sooner, at the 4-cell stage. Furthermore, in mice, the vast majority of ETCHbox transcripts are activated at the 2-cell stage. Crucially, in all three species this ETCHbox activation coincides with the major wave of EGA, which occurs at the 8-cell stage in humans[291], 4-cell stage in pigs[720] and 2-cell stage in mice[331]. Based on this, I argue that ETCHbox genes are activated as part of the major wave of EGA and that this is ancestral to at least the Boreoeutheria. Differences in their activation timing across species simply reflect differences in the timing of EGA.

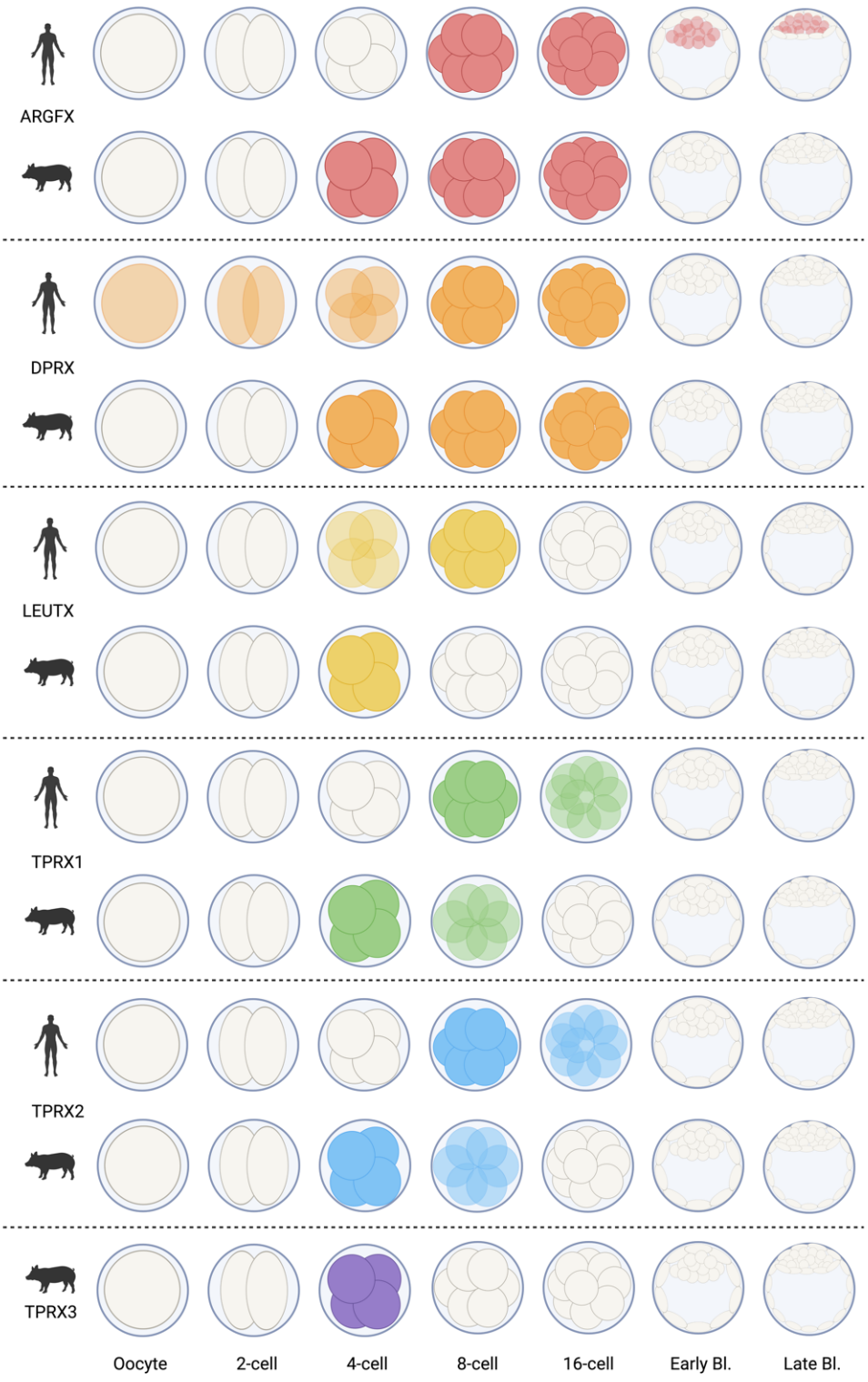


Figure 6.16 Expression of ETCHbox genes in human and pig embryos. Expression is categorised into high (opaque, coloured cells), low (translucent cells) or none (beige cells). Broadly, expression is activated one cell division later in humans than in pigs. Human *TPRX2* was not included in the dataset analysed here and therefore data is drawn from Maeso et al. (2016)[90]. Figure created with BioRender.

Which factors activate ETCHbox genes at EGA? Using GRN reconstruction and *cis*-regulatory motif analysis, I identified the double homeobox protein DUXA as a putative activator of both *ARGFX* and *LEUTX* in the human preimplantation embryo. Several factors lead me to believe that DUXA is a genuine activator of ETCHbox genes in vivo. First, the method used is highly robust: it is specifically designed for scRNA-seq data to make use of each cell as an individual data point for coexpression analysis, and to account for potentially problematic variation like technical dropouts[715]. It also combines multiple data sources, validating the links found by coexpression with motif enrichment analysis. Second, the pipeline is very conservative; for example, no regulators of *DPRX* or *TPRX1* were identified with high confidence, and for *ARGFX* and *LEUTX* the only high-confidence activator reported was DUXA: this lends more support to any link which remains after this process. Third, the DUXA binding motif used is almost identical to one independently predicted and validated for DUX4, a closely related member of the *DUX* group, suggesting that it is likely to be accurate[61]. Fourth, the input data are from preimplantation embryos themselves, rather than over-expression experiments, which should lead to the identification only of links active during this part of development. Fifth, DUXA is a very good candidate to be an activator of ETCHbox genes: its expression (and translation) begins immediately prior to ETCHbox activation, peaking between the 4-cell and 8-cell stages, and it is activated in the minor EGA that occurs at around the 4-cell stage, one cell division before the major EGA at which ETCHbox activation occurs[94,100]. Finally, other work supports this conclusion: most notably, shRNA knockdown of *DUXA* results in *LEUTX* downregulation[723].

The DUXA locus is itself bound and activated by the TF which is considered to be the cornerstone of EGA-initiation, DUX4[61,64,724], suggesting a cascade of DUX4, DUXA and then ETCHbox proteins is active in preimplantation development (Figure 6.17). Indeed, expression of DUX4 in ESCs results in the upregulation of *LEUTX*, and DUX4 knockdown embryos have reduced *ARGFX* expression[64]. Overall, DUXA is an activator

of *ARGFX* and *LEUTX* in the preimplantation embryo. Future work should aim to validate this regulatory link with ChIP-seq data.

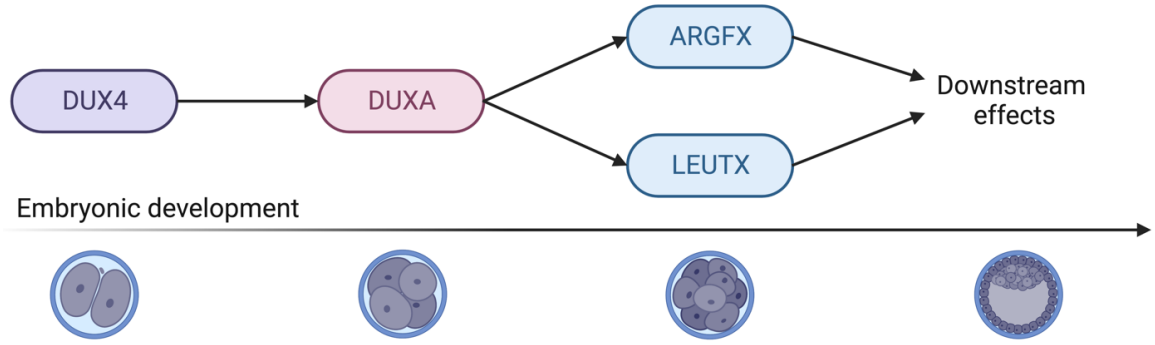


Figure 6.17 Cascade of ETCHbox activation in the preimplantation embryo. DUX4, a key initiator of EGA, activates *DUXA* expression. DUXA then activates *ARGFX* and *LEUTX* expression. Figure created with BioRender.

Based on ectopic expression data, Maeso et al. (2016)[90] postulated a model where human *ARGFX*, *LEUTX* and *TPRX1* coordinate a pre-blastocyst gene expression pulse, with *ARGFX* acting to upregulate and then *LEUTX* and *TPRX1* acting to downregulate a set of genes expressed for just two cell divisions from the 8-cell stage to the morula (Figure 6.18). My data suggest that a refinement to this model is required: *TPRX1* and *LEUTX* peak in their expression at the 8-cell stage, with expression completely extinguished before the blastocyst, while *ARGFX* expression peaks in the morula and continues into the ICM and epiblast. I therefore propose a new model whereby *LEUTX* and *TPRX1* repress a set of genes pre-blastocyst and *ARGFX* mediates the release of this repression specifically in the ICM and epiblast (Figure 6.18). Supporting this is the fact that some of the genes upregulated by *ARGFX* and downregulated by *TPRX1* and *LEUTX* not only peak in expression at the blastocyst stage (not at the 8-cell stage or morula), but also exhibit ICM and epiblast specificity (e.g. *BAMBI, ID1, PMAIP1*).

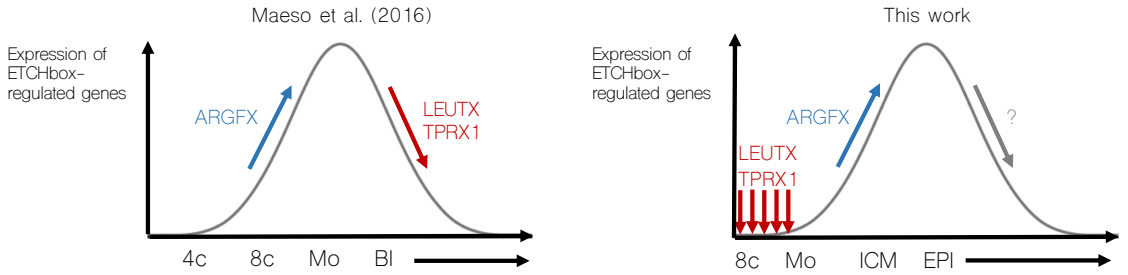


Figure 6.18 An update to the model for ETCHbox function in the human preimplantation embryo. The model previously suggested by Maeso et al. (2016)[90] proposed that ARGFX upregulates a gene set that is briefly expressed in a pulse from the 8-cell to morula stages, and then LEUTX and TPRX1 act antagonistically to downregulate the same gene set. Data analysed in this work are more consistent with a new model, in which LEUTX and TPRX1 repress a set of genes in the 8-cell and morula stages and this is followed by ARGFX-mediated release from repression specifically in the ICM and epiblast. This leaves open the question of the identity of the downregulating factor. 8c = 8-cell; Mo = morula; ICM = inner cell mass; EPI = epiblast; Bl = blastocyst.

The large array of tandem *Obox* duplicates present in the mouse genome provides the opportunity to explore the fate of individual genes when many highly similar copies are present. One important question is why so many duplicates, especially those with almost identical sequences, are retained by evolution. We profiled the expression of each gene in the preimplantation embryo and made gene trees to assess genetic similarity. Across *OboxA*, *OboxB* and *OboxC* groups, four different expression profiles are exhibited, showing that the *Obox* duplications have been followed by some divergence of expression patterns, suggesting subfunctionalisation has occurred[228,231]. Importantly, some genes are maternally expressed, while others are present only post-EGA, suggesting key differences in function. However, this variation of expression is insufficient to explain the retention of such a large cluster of genes, many of which have both highly similar sequences and highly similar expression profiles.

Royall et al. (2018)[96] previously reported the expression profiles of seven *Obox* genes using bulk RNA-seq data. For most genes our new single-cell data is consistent with these results, validating our approach. Interestingly, Royall et al.[96] found no expression of *OboxD2* in the preimplantation embryo, whereas we have strong evidence for

expression in the zygote and at the 2-cell stage. This difference may be due to differences in analysis pipelines; our analysis using Kallisto[407] is better equipped than the previously used STAR[445] and Cufflinks[725] pipeline to process highly similar sequences with many multi-mapped reads.

A further goal of this chapter was to test one of the hypotheses for the cause of ETCHbox genes' rapid evolutionary changes. It has been suggested that this accelerated evolution may be attributed to a role in placentation[92]. Since the process of placenta formation varies significantly between mammals[127,128], it was theorised that ETCHbox genes may have evolved quickly as part of the selection that caused the divergence of placentas. Given the timing of ETCHbox expression, their contribution to placentation would be in the TED cells fated to become the placenta. I find a complete absence of ETCHbox expression in the TED of all three species included in the analysis, firmly demonstrating that ETCHbox proteins are not involved in TED or placenta formation.

6.7 Conclusions

Quantification of gene expression at a single-cell level is a recent addition to the armoury of analyses that can be used to explore a gene's expression dynamics. In this chapter, I used four published scRNA-seq datasets to explore at single-cell resolution the expression patterns of the fast-evolving ETCHbox genes in the eutherian preimplantation embryo. I found that consistently across humans, mice and pigs, ETCHbox genes are activated as part of the major wave of EGA. In humans, this activation is driven at least in part by another PRD-class homeodomain protein, DUXA, which peaks in expression just before major EGA. While the expression of human *DPRX*, *LEUTX* and *TPRX1* is eliminated by the late morula, *ARGFX* expression persists into the blastocyst, where it shows strong specificity to the ICM and then epiblast. Moreover, my data show that ARGFX upregulates genes specific to these lineages, including key ICM markers like

BAMBI, suggesting a role in ICM formation or specification. This supports the idea that the ETCHbox genes, which are specific to the eutherians, are involved in the development of the ICM, which is a eutherian novelty.

6.8 Supplementary figures

Figure S6.1

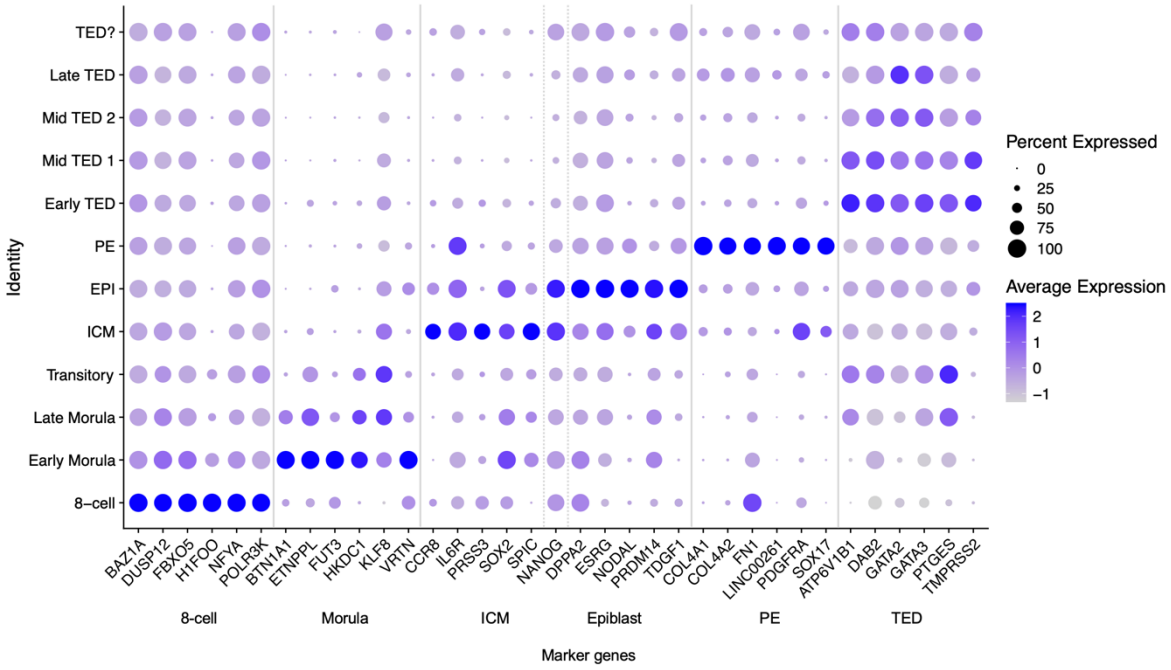


Figure S6.1 Marker genes for each stage and cell lineage in the human preimplantation embryo. Six representative marker genes are shown for each cluster; further genes were used for cluster identification. NANOG is a marker of both the ICM and then epiblast. Point size = percent of cells expressing gene. Point colour = average expression level in each cell; darker colours = higher expression. Abbreviations: EPI = epiblast; ICM = inner cell mass; PE = primitive endoderm; TED = trophectoderm.

Figure S6.2

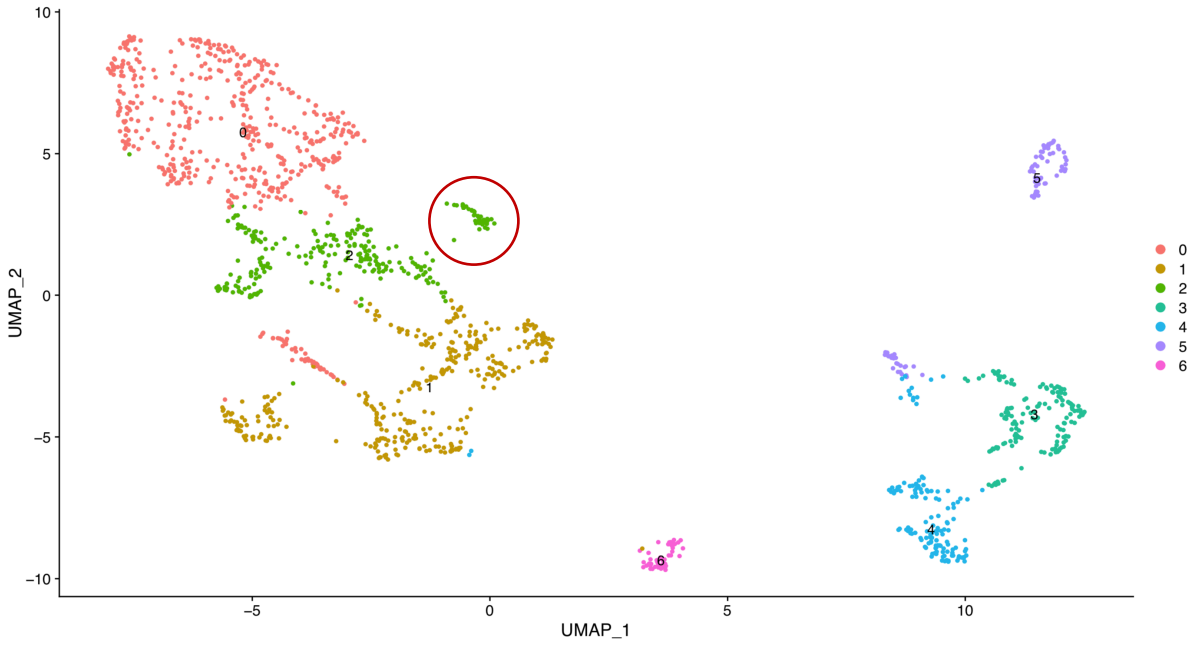


Figure S6.2 Identification of a divergent TED cluster. Re-clustering of human preimplantation embryo cells at lower resolution; the divergent TED group (red ring) clusters with mid-TED cells.

Figure S6.3

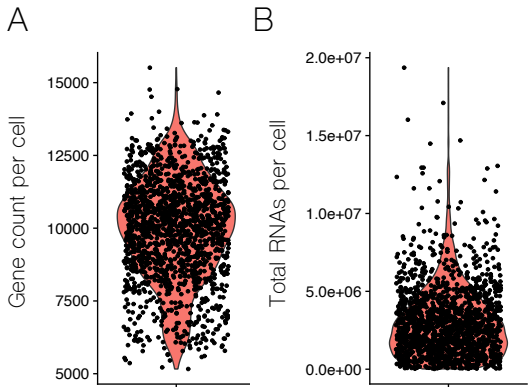


Figure S6.3 Key statistics for cells retained after quality control of human scRNA-seq dataset. (A) number of genes per cell; **(B)** total number of RNAs per cell.

Figure S6.4

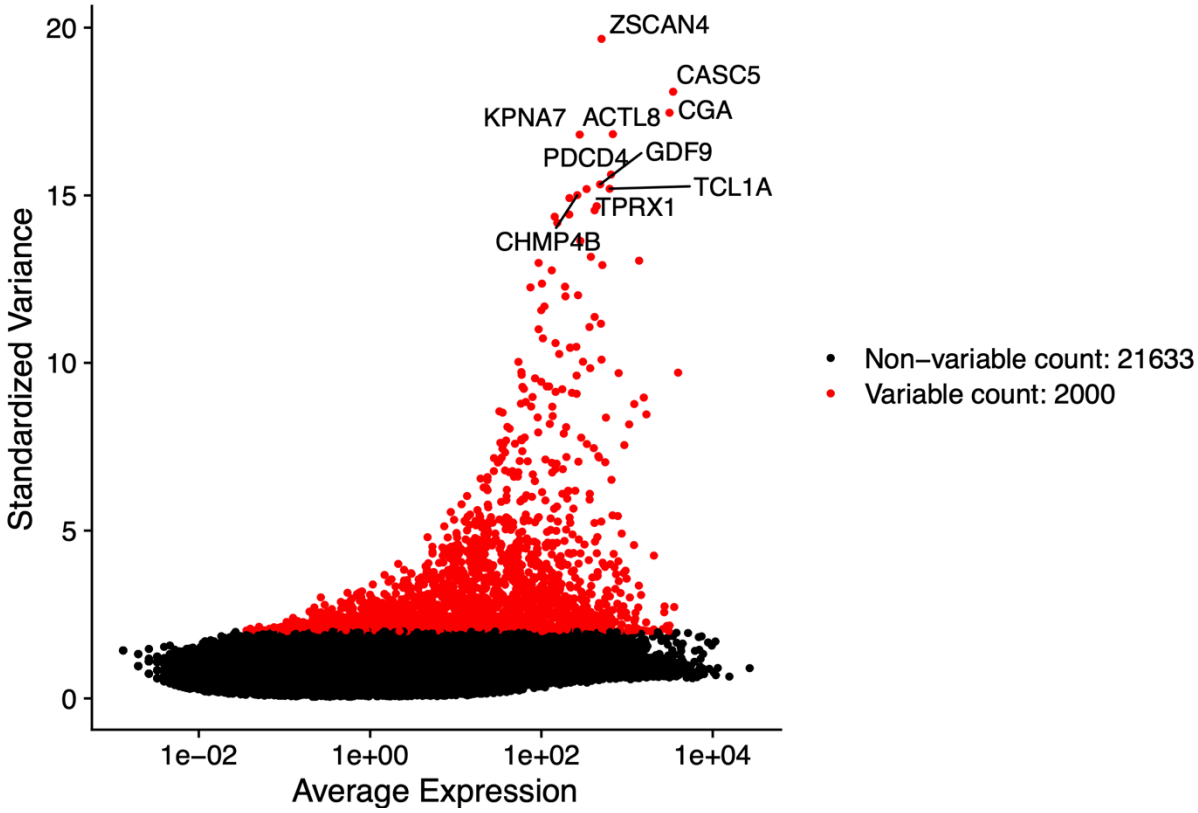
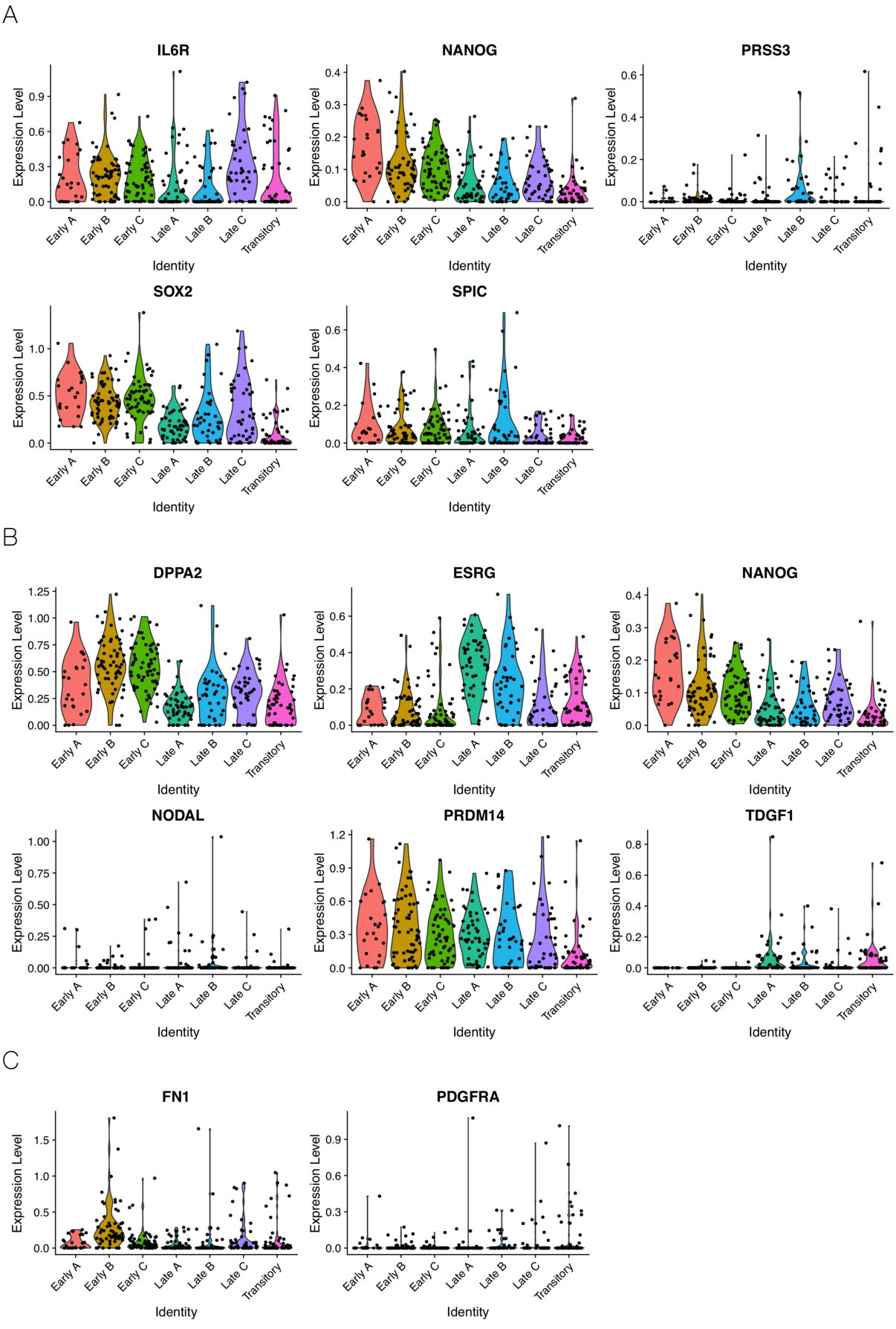


Figure S6.4 Identification of most variable genes in the human preimplantation embryo dataset. The top 2,000 most variable genes are employed for clustering analyses. The top 10 most variable genes are labelled on the plot, and include the ETCHbox gene *TPRX1*.

Figure S6.5


A
IL6R
NANOG
PRSS3
SOX2
SPIC
B
DPPA2
ESRG
NANOG
NODAL
PRDM14
TDGF1
C
FN1
PDGFRA
Expression Level
Identity
Early A
Early B
Early C
Late A
Late B
Late C
Transitory

D

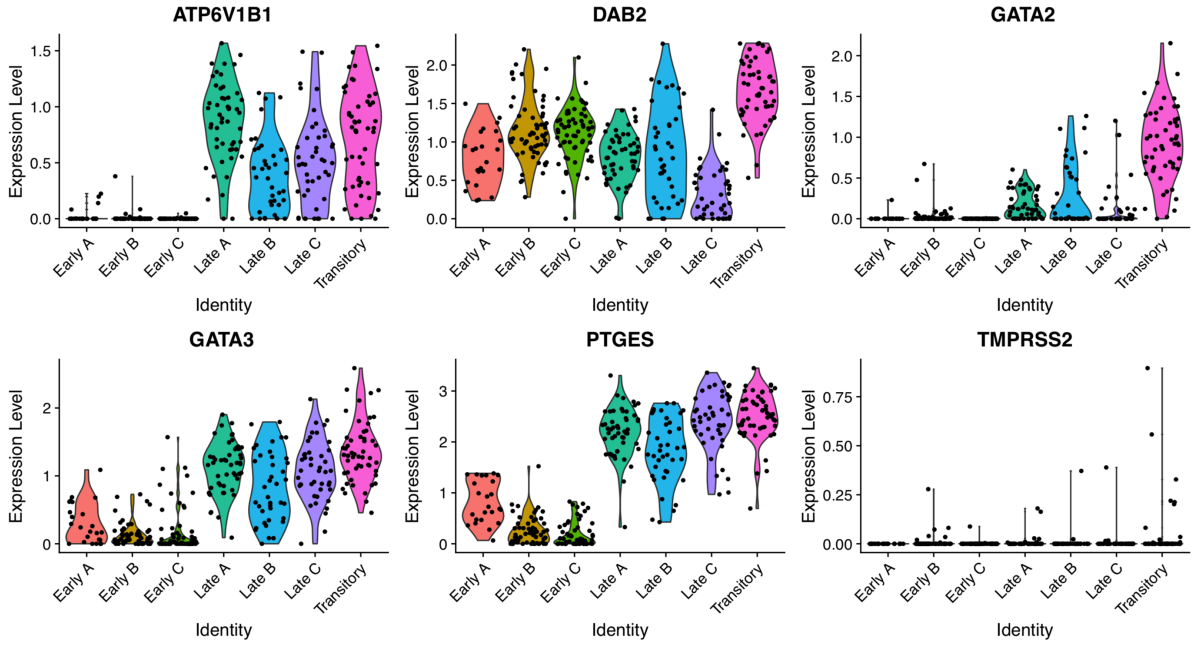


Figure S6.5 Searching for lineage specification of human morula and transitory cell populations. Violin plots show expression of **(A)** ICM, **(B)** epiblast, **(C)** PE and **(D)** TED lineage markers in the clusters of morula and transitory cells. The markers of none of these lineages show clear specificity to certain clusters, showing that no lineage specification is detected in morula cells.

Figure S6.6

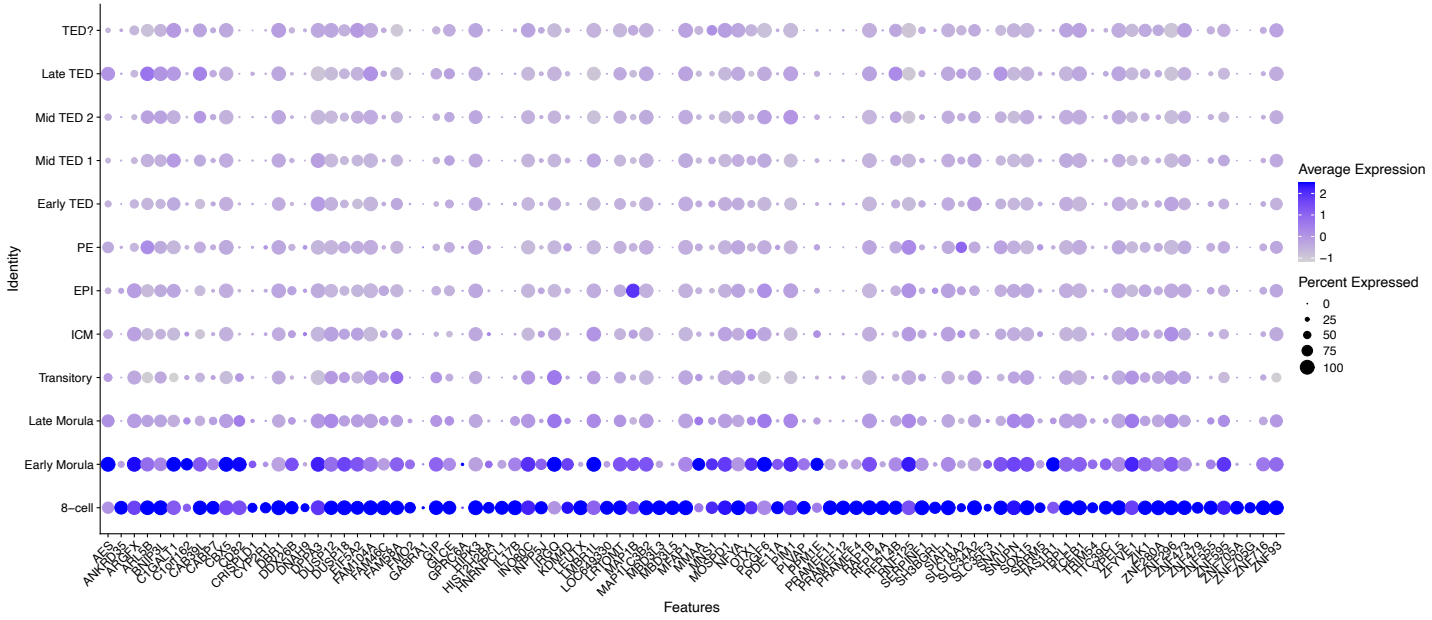


Figure S6.6 Expression profiles of genes in the DUXA regulon. Putative direct targets of DUXA are almost without exception expressed most highly at the 8-cell stage and in the early morula. Point size = percent of cells expressing gene. Point colour = average expression level in each cell; darker colours = higher expression. Abbreviations: EPI = epiblast; ICM = inner cell mass; PE = primitive endoderm; TED = trophectoderm.

Figure S6.7

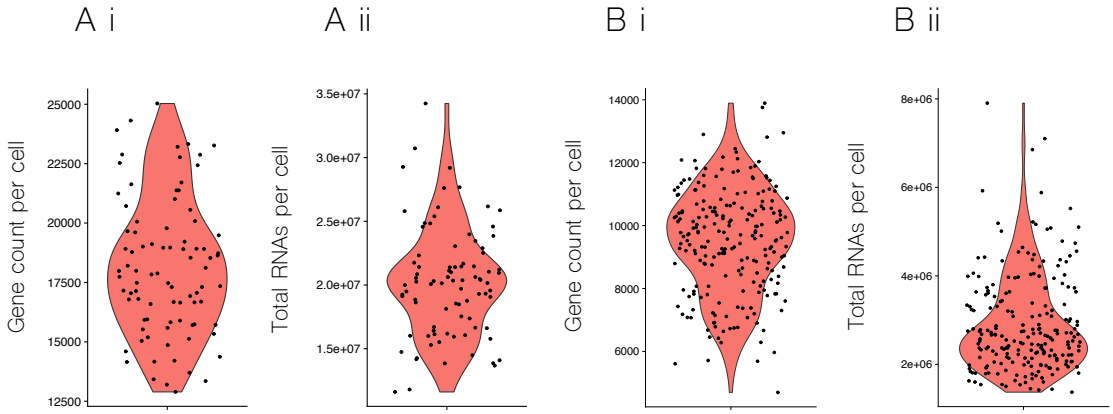


Figure S6.7 Key statistics for cells retained after quality control of pig scRNA-seq datasets. (A) Dataset 1**: (i)** number of genes per cell, **(ii)** total number of RNAs per cell. **(B)** Dataset 2: **(i)** number of genes per cell, **(ii)** total number of RNAs per cell.

Figure S6.8

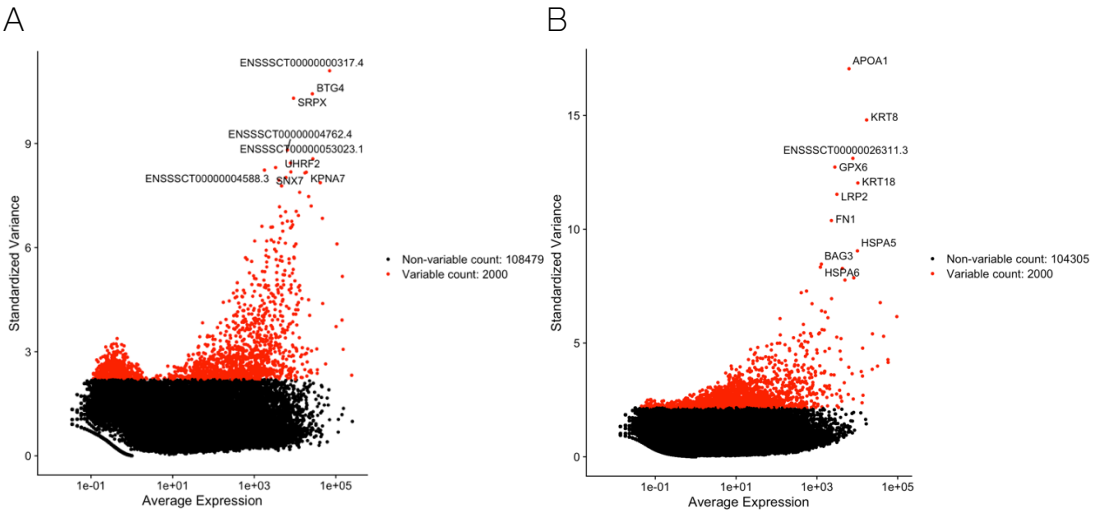


Figure S6.8 Identification of most variable genes in pig preimplantation embryo Datasets 1 and 2. (A) Dataset 1 and **(B)** Dataset 2. The top 2,000 most variable genes are employed for clustering analyses. The top 10 most variable genes are labelled on the plots. Kallisto equivalence classes containing multi-mapped reads are retained for clustering analyses.

Figure S6.9

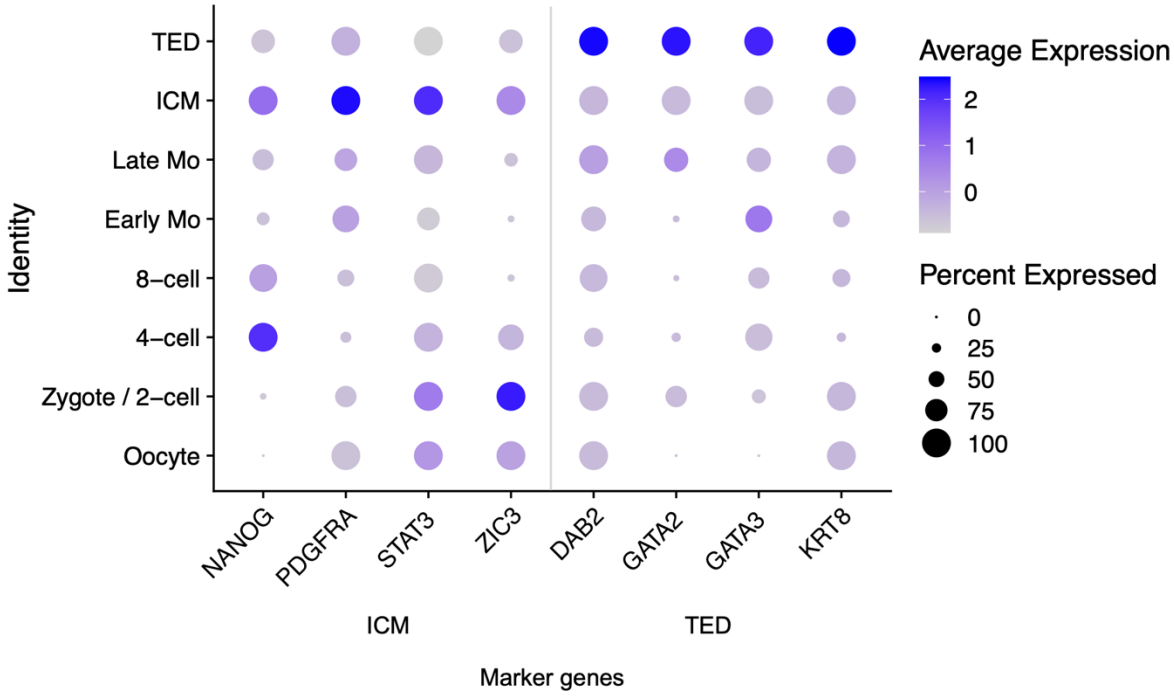


Figure S6.9 Marker genes for blastocyst lineages in pig Dataset 1. ICM is identified by the simultaneous expression of epiblast and hypoblast markers (*NANOG* and *PDGFRA* respectively), and also the expression of *STAT3* and *ZIC3*. Abbreviations: ICM = inner cell mass; Mo = morula; TED = trophectoderm.

Figure S6.10

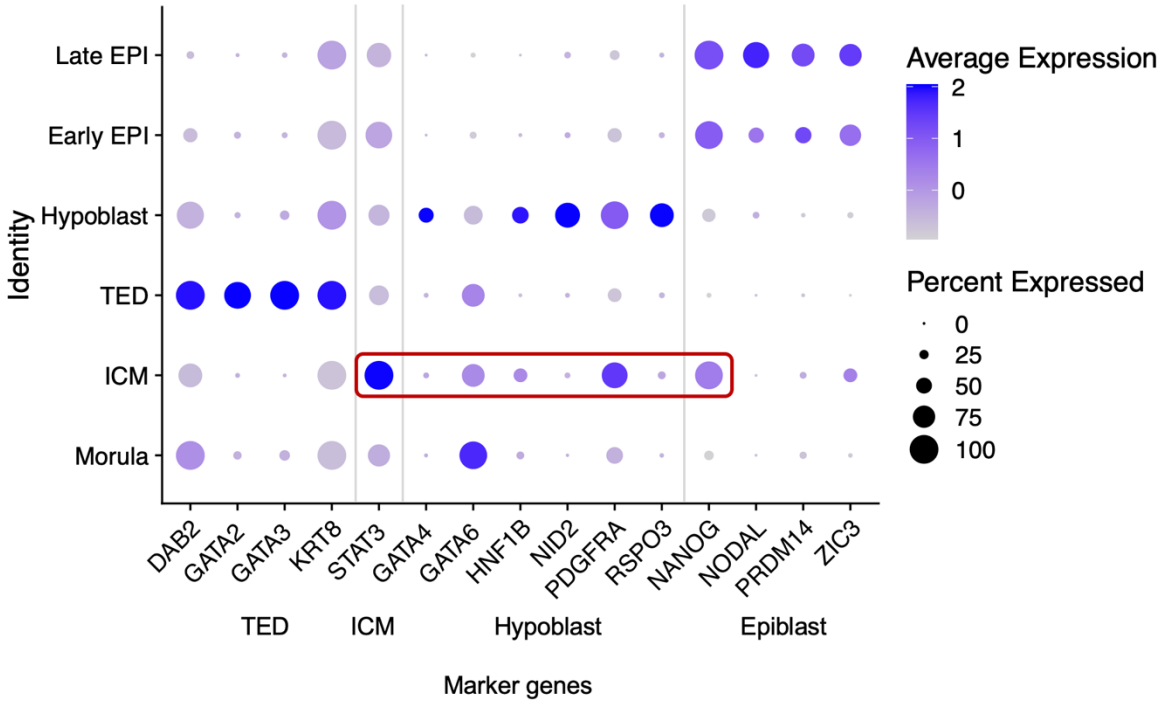


Figure S6.10 Marker genes for blastocyst lineages in pig Dataset 2. ICM is identified by the simulataneous expression of epiblast and hypoblast markers (*NANOG* and *PDGFRA* respectively), and also the expression of *STAT3* (red ring). Abbreviations: EPI = epiblast; ICM = inner cell mass; TED = trophectoderm.

Figure S6.11

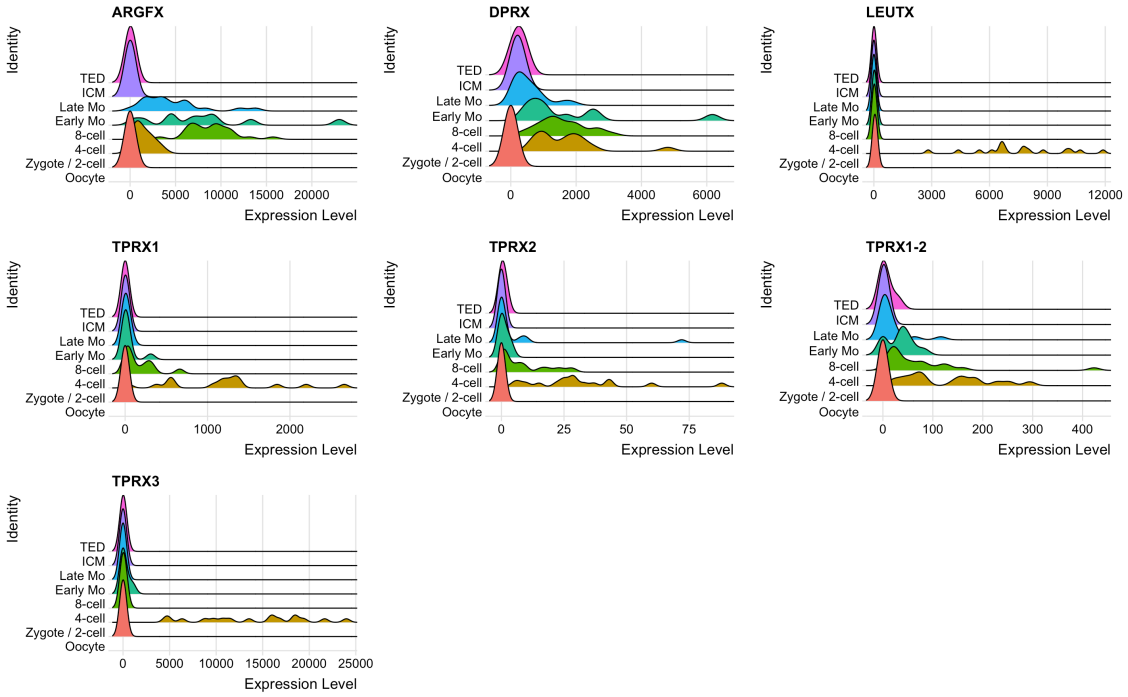


Figure S6.11 Ridge plots of ETCHbox expression. Data from pig Dataset 1. Plots demonstrate that ETCHbox expression is heterogeneous even at its peak. Abbreviations: ICM = inner cell mass; Mo = morula; TED = trophectoderm.

Figure S6.12

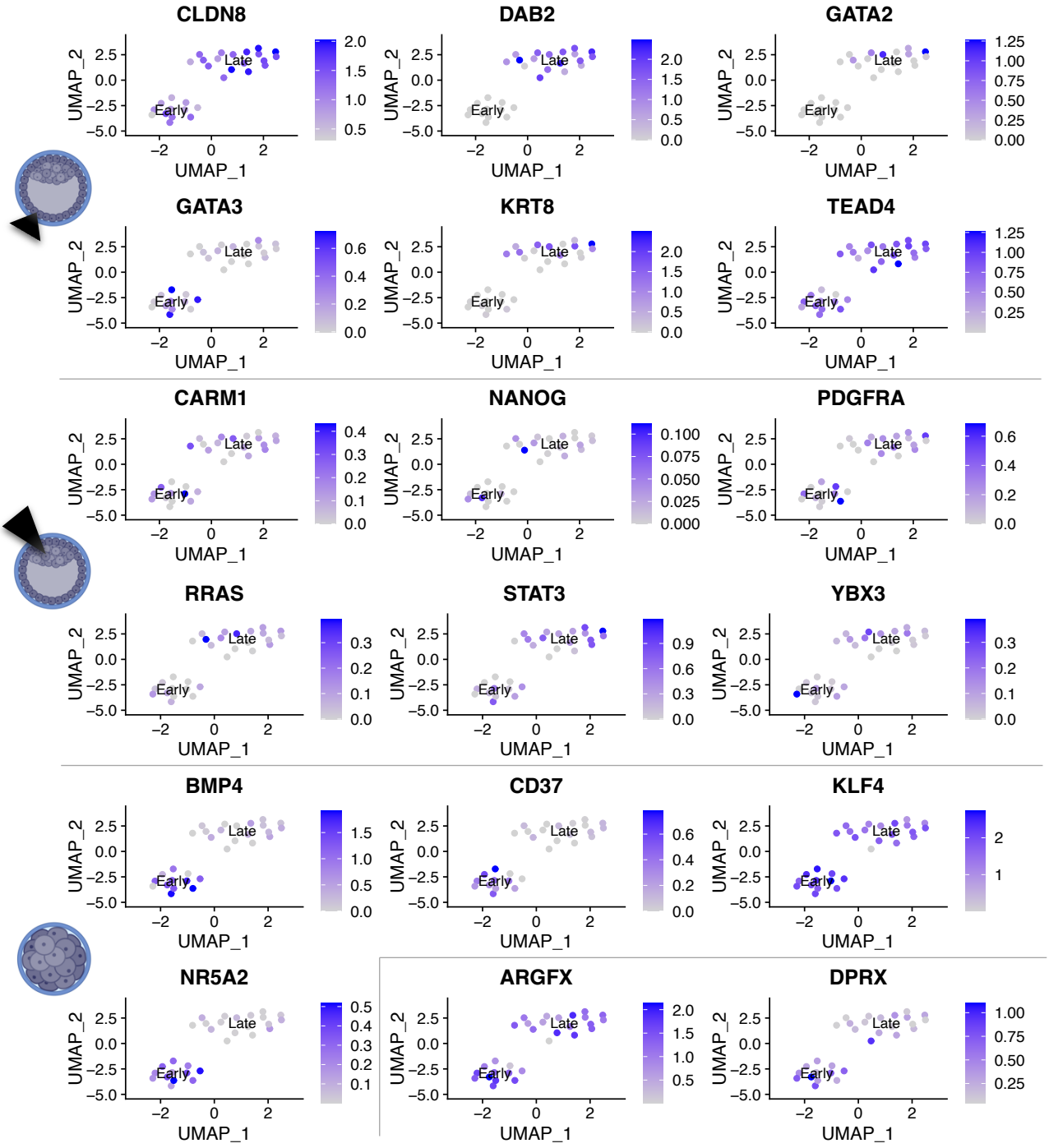


Figure S6.12 Lineage specification in the porcine morula (Dataset 1). UMAP visualisation of the expression of genes reported[701] to be specific to morula ICM-fated (*CARM1*, *RRAS*, *YBX3*) and TED-fated (*DAB2*, *CLDN8*, *TEAD4*) cells, alongside the pig marker genes used above for ICM (*NANOG*, *PDGFRA*, *STAT3*) and TED (*GATA2*, *GATA3*, *KRT8*). ETCHbox genes *ARGFX* and *DPRX* are also shown. Embryos cartoons and arrows indicate which marker genes are shown.

Figure S6.13

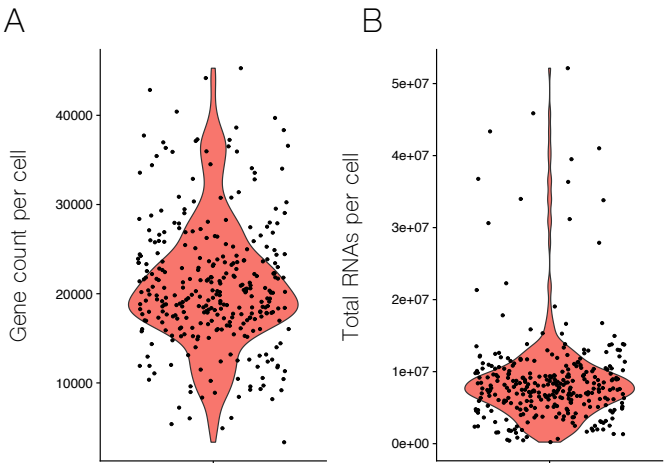


Figure S6.13 Key statistics for cells retained after quality control of mouse scRNA-seq dataset. **(A)** number of genes per cell; **(B)** total number of RNAs per cell.

Figure S6.14

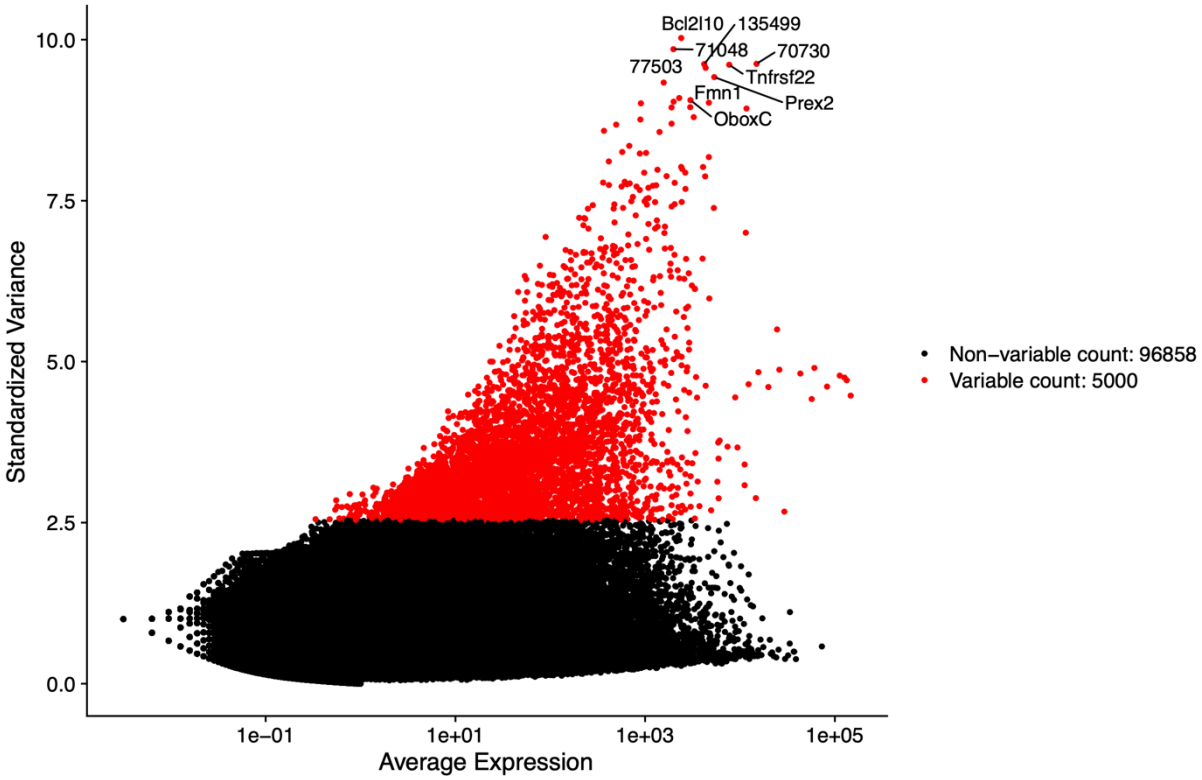


Figure S6.14 Identification of most variable genes in mouse preimplantation embryo dataset. The top 5,000 most variable genes are employed for clustering analyses. The top 10 most variable genes are labelled on the plot, including the ETCHbox gene *OboxC*. Kallisto equivalence classes containing multi-mapped reads are retained for clustering analyses.

Figure S6.15

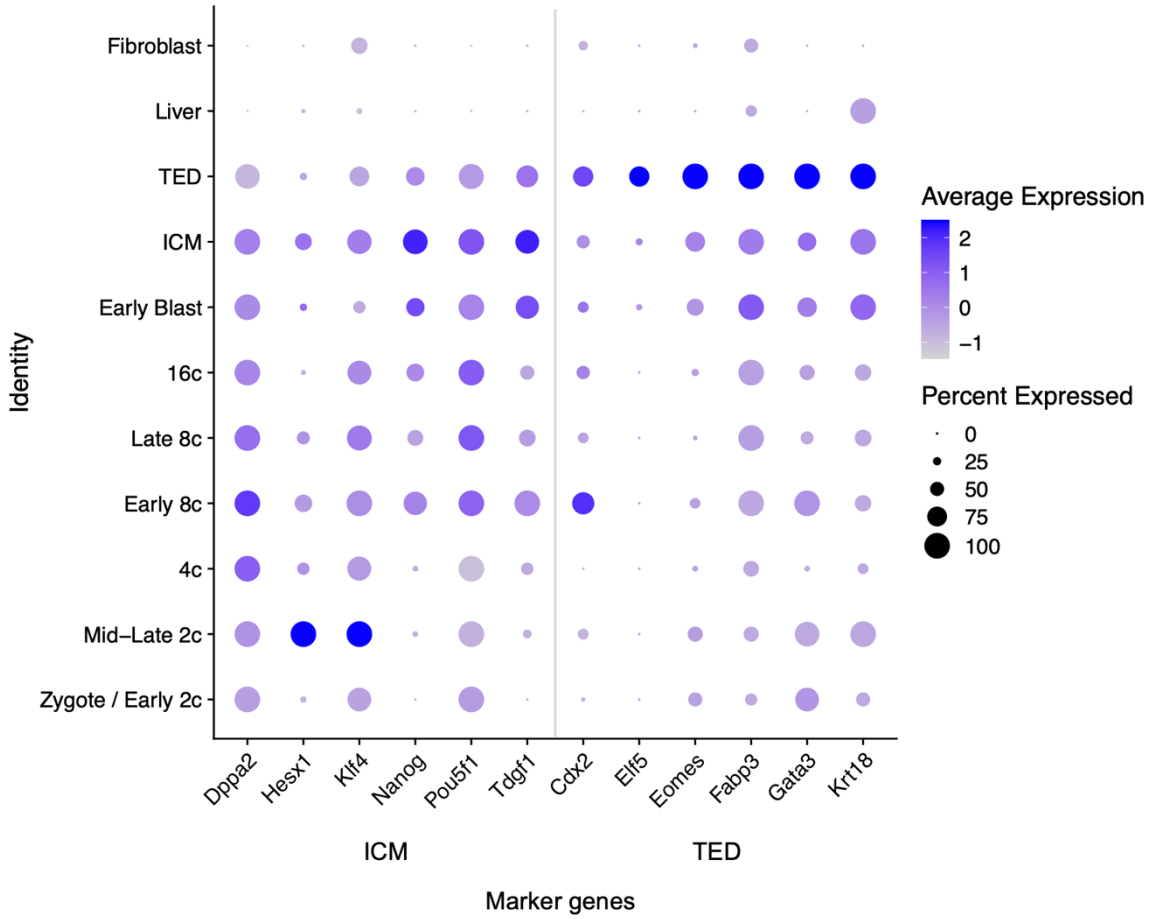


Figure S6.15 Marker genes for each blastocyst lineage in the mouse preimplantation embryo. Six representative marker genes are shown for each cluster; further genes were used for cluster identification. Point size = percent of cells expressing gene. Point colour = average expression level in each cell; darker colours = higher expression. Abbreviations: c = cell; ICM = inner cell mass; TED = trophectoderm.

Electronic supplementary material

ESM6.1 Supplementary Tables S6.1 – S6.11.

Chapter 7: General Discussion

7.1 Synopsis

Current explanations for the existence of transcription factors that function during embryonic development but display rapid evolution are in large part unsatisfactory. In broad terms, the aim of this thesis has been to use the Eutherian Totipotent Cell Homeobox (ETCHbox) group to push forward our understanding of (i) the functions such genes have been recruited to perform and (ii) the processes underlying their unusual evolutionary trajectories. In Chapter 2, I confirmed that the ETCHbox group does indeed exhibit rapid evolution both in terms of protein-coding sequences and copy number variation, and found that positive selection and extensive gene conversion have contributed to sequence evolution[525]. Chapter 3 consists of transcriptomic analysis following the ectopic expression of bovine ETCHbox genes, from which four main conclusions emerge: first, bovine ETCHbox proteins ARGFX and LEUTX are involved in the structural formation of the blastocyst; second, the gene sets regulated by bovine ARGFX and LEUTX have a large overlap, suggesting functional redundancy; third, this redundancy facilitates the survival of individuals homozygous for a pseudogenising mutation in ARGFX; and finally, the function of ARGFX has changed dramatically between humans and cattle[598]. Given this final result, Chapter 4 asked whether similar changes have occurred between more closely related species, and found evidence for continued rapid sequence divergence but only small modifications to ETCHbox protein function within primates. In Chapter 5, I showed that several ETCHbox genes downregulate a multiplicity of transposable element (TE)-repressing genes. I found preliminary evidence that, in cell culture, this caused an increase in the rate of retrotransposition occurring, but no detectable increase in TE transcription. Finally, the single-cell transcriptomic approach of Chapter 6 established that ETCHbox genes are consistently activated at the major wave of embryonic genome activation (EGA) in three

species; suggested that in humans this is mediated by DUXA; and uncovered strong specificity of human *ARGFX* to the embryonic lineages that go on to form the embryo proper. Overall, this work represents an advance to our understanding of the evolutionary dynamics of homeobox genes expressed during embryonic development. In this final chapter, I shall not repeat the arguments of previous discussion sections and will instead focus on tying together and exploring the general themes that emerge from this work.

7.2 Functional redundancy as an explanation for rapid evolutionary change

It is an inescapable conclusion from this thesis and previous work that ETCHbox genes are rapidly-evolving. Indeed, the results of my study in primates show that they did not simply undergo change during mammalian diversification followed by relative stasis, but rapid evolution has persisted even between closely related species. In the introductory chapter, I proposed six possible explanations for rapid evolution. Results throughout this thesis provide further evidence if it was necessary that ETCHbox genes are not evolving rapidly by genetic drift due to the lack of a functional role. This is epitomised by findings in Chapter 2 that no sampled species possesses fewer than two ETCHbox genes, and that positive selection has been a key factor in their evolution. In addition, my analysis of single-cell RNA-sequencing data in Chapter 6 shows definitively that ETCHbox genes are not active in the trophectoderm and therefore are not rapidly-evolving as part of a role in placentation, and data from Chapter 5 suggest that they are not in an arms race with TEs.

Instead, several key results in this thesis support the idea that mutational robustness through the mechanism of functional redundancy has resulted in relaxed selection and subsequently an accelerated rate of evolution. Royall[171] previously argued that distributed robustness is the cause of mutational robustness within the ETCHbox group,

but since this theory posits that perturbations to a network are compensated by non-redundant parts[124], in my opinion it is not consistent with the abundant data supporting functional overlaps. The strongest evidence emerges from Chapter 3, in which I found not only a large overlap of the gene sets regulated by bovine ARGFX and LEUTX, suggesting their transcription factor function is partially redundant, but also that a pseudogenising mutation of ARGFX which eliminates the homeodomain is non-lethal. I propose that this is because LEUTX can compensate for the absence of a functional ARGFX. The relaxed selection and ubiquitous nature of gene loss within the ETCHbox genes detected in Chapter 2 also supports this idea. In addition, the lack of sequence divergence and highly similar expression profiles of mouse *Obox* (*TPRX2*) genes (Chapter 6) strongly imply the existence of redundancy in this system.

Two recent papers support this contention. First, *TPRX1/2/L* triple knockdown embryos exhibit developmental delays and the failure to activate many EGA genes, while double and single knockdowns result in progressively less severe effects[94]. This suggests that paralogous *TPRX* genes compensate for gene loss in single and double knockdown embryos. Second, redundancy can be extended outside the ETCHbox group itself: the mouse *TPRX2* gene *Obox4* exhibits partial redundancy with another PRD-class homeobox gene, *Dux*, binding over 50% of the gene loci bound by Dux and compensating for Dux in *Dux* knockdowns[119]. Taken together, there is strong evidence for redundancy between ETCHbox proteins and even with homeobox genes of other groups. This supports a model in which the genes' position on the most dynamic autosome in the genome (chromosome 19 in human) generates variation such as duplications and losses, and redundancy shields this variation from selection, permitting fast evolutionary change and especially gene loss.

7.3 The lability of mammalian early embryonic development

The search for an explanation for the ETCHbox genes' rapid evolution inevitably leads us to look to other gene families for answers. In doing so, it becomes clear that rapid evolution is not a unique property of the ETCHbox group but is mirrored by several other preimplantation embryo-expressed families, including other PRD-class homeobox genes. *RHOX* and *DUX* are two striking examples that have fundamental and indispensable roles in eutherian preimplantation development[60–62,69,609]. Indeed, these homeobox genes reflect a more general remarkable divergence in the molecular mechanisms underlying preimplantation development between mammalian species (reviewed previously[726,727]). A relatively minor example from the work in this thesis is the strong specificity of ARGFX to the inner cell mass in human embryos which is not replicated in either mouse or pig.

The question of why different eutherians build their embryos so differently has increasingly troubled me throughout the progress of the work that now comprises this thesis. At first, I instinctively reached for the "if it ain't broke, don't fix it" maxim, especially given the critical importance of early development: if we assume that preimplantation development was effective in the last eutherian common ancestor, why have the molecular mechanisms underlying it been modified so extensively? Given the irrefutable evidence that this tenet does not apply here[335,726,727], I now wonder whether a more general explanation for why preimplantation development is so variable might be necessary before we fully understand the evolution of ETCHbox and other unusual homeobox gene families. It seems implausible that such divergence could be adaptive and driven by positive selection; it does not appear as if one lineage or another is particularly 'better' or 'worse' at building an early embryo and, indeed, there is evidence that mammalian development is surprisingly error-prone across several divergent species[307–309,538]. Is this then, an example of developmental system drift (DSD)[728]; the

idea that changes in the molecular underpinnings of conserved features of development are ubiquitous and occur through chance as opposed to positive selection? It may be the case that selection is particularly relaxed at this early stage, and therefore that the lability arises from the lack of selection against changes. Of course, the extraordinary conservation of many homeobox genes provides evidence to the contrary, and suggests there may be something special about ETCHbox and similar families that makes them fast-evolving.

The hourglass model provides another lens through which to view this argument. It posits that early and late stages of embryonic development are labile while the middle part (the phylotypic stage) is highly constrained and similar across species within a phylum, possibly because conserved genes like Hox are pattering the embryo and laying out the body plan at this point, leading even the smallest perturbations to be deleterious[729,730] (Figure 7.1). Initially proposed based on embryonic morphology, two seminal 2010 papers in the same issue of Nature extended this to a molecular level in *Drosophila* and vertebrates[731,732]. By this argument, it is a lack of constraint at the early stage that permits variability, rather than positive selection for adaptation actively driving it. At the level of the gene, an alteration of molecular function in early development may be tolerated by evolution as long as the function is not completely lost, permitting rapid sequence change. Indeed, there are several examples that show the lack of an effect on adult phenotype of dramatic divergence in early development[733] which could be used to support this reasoning. It is possible that the special feature of ETCHbox and Dux genes is simply the timing of their expression to the very early embryo.

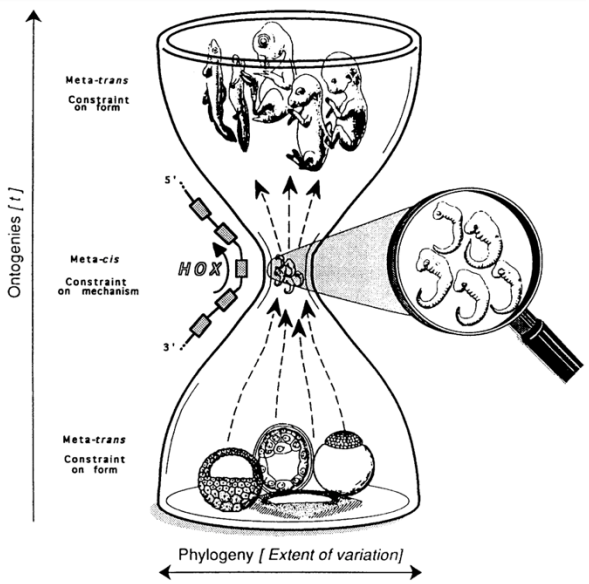


Figure 7.1 The hourglass model of embryonic development. Classic diagram from Duboule (1994)[729] depicting the hourglass model for development in vertebrates. While there is variation at the start and towards the end of development, embryo morphology and gene expression converge around the early somites/tailbud stages, the time at which Hox genes are active.

Much consideration has been given to the factors causing the convergence of the phylotypic stage (e.g.[734]) but relatively less attention has been paid to what permits the extensive early divergence. The aforementioned DSD is one possibility. Gene duplication followed by subfunctionalisation is a known mechanism of DSD[735]: the overall function performed by the gene(s) pre- and post-duplication is unchanged but the molecular mechanisms underlying it (i.e. which genes are performing it) have been altered. This is consistent with our current knowledge surrounding ETCHbox genes: we have a mechanistic explanation for their high rates of gene duplication through their position on a highly dynamic chromosome with the highest rate of duplication in the genome[120,123], but we require an evolutionary explanation for the frequency with which duplicates are retained in the genome. One can imagine a situation in which the general lability of preimplantation development permits the retention of duplicates that are initially redundant, and occur frequently in ETCHbox genes due to their chromosomal location; over time, subfunctionalisation occurs, leading to modifications of the gene regulatory

networks (GRNs) underlying preimplantation development without effects on the embryonic phenotype.

Ultimately, we find ourselves asking whether relaxed selection in early development is sufficient to explain the observed diversity of the GRNs in the preimplantation embryo and the remarkable rate of evolution of gene families like ETCHbox. A recent study that detected increased positive selection acting on enhancers in early development versus those at the phylotypic stage[736] suggests not, and that adaptation may also play a role. The extensive positive selection I found acting on ETCHbox protein-coding sequences also strongly refutes the idea that they are evolving due to relaxed selection alone, and evidence of positive selection acting on the fast-evolving PRD-class *RHOX* genes supports this[69,72–74].

One adaptive hypothesis for early divergence is that species-specific ecological niches result in selection for divergent reproductive strategies which manifest in early development[733]. For instance, one potential driver of early divergence is selection for accelerated development imposed by environmental pressures. Long-germ versus short-germ development in insects and direct versus indirect development in sea urchins, for example, both result in large decreases in developmental duration with minimal effect on the adult phenotype[733,737–740] but significant changes in the molecular mechanisms underlying development[733,741,742]. This provides an adaptive hypothesis to explain divergence in early development without invoking DSD and is an intriguing concept when we come to think about eutherians. Gestation time varies dramatically across the clade (e.g. 20 days in mice versus 280 days in humans); rodents, for example, have been selected for rapid reproduction and shorter generation times. Even within primates, the gestation period varies from 60 to 289 days[743]. Importantly, this variation translates to differences in the duration of preimplantation development, which takes around 4.5 days in mice compared to 7 in human, 9 in cattle and 40 in

horse[283,285,326,336,744], and it is therefore possible that selection imposed by ecological conditions and acting on life history traits like generation time causes an acceleration (or deceleration) in preimplantation development which requires alterations to GRNs.

The overall picture remains unclear, but while relaxed selection is unquestionably important, it seems likely that at least some divergence in the molecular mechanisms underpinning preimplantation development is adaptive. In the case of ETCHbox genes, the observed positive selection acting on sites of functional importance is a key indicator that this is indeed a factor. In addition, functional redundancy has undeniably been another important contributor; it is possible that redundancy between genes expressed in very early development is a more general explanation for the lability at this stage. Evidence that this is not restricted to ETCHbox genes comes from another PRD-like group, *CPHX*, of which the two human copies have been shown to have highly overlapping gene sets[75]. To close this section, I conclude that one single factor is insufficient to explain the observed dynamism, and a combination of DSD, functional redundancy and positive selection has shaped ETCHbox repertoires since their arrival at the base of the eutherians

7.4 The multifunctional nature of the ETCHbox group

7.4.1 ETCHbox proteins and the eutherian blastocyst

Having for the majority of this discussion focussed on the factors contributing to ETCHbox genes' puzzling evolutionary properties, I now briefly consider their functional role in the preimplantation embryo. In addition to a previously identified function in the upregulation of EGA genes[75,104], I identify roles in blastocyst formation, cellular potency and possibly the regulation of TEs. One thing that becomes clear from this body of work is that we cannot pin one specific function on ETCHbox proteins. To talk about the 'function' of the ETCHbox group is without doubt an oversimplification and we must

instead consider them as multifunctional components of early embryo GRNs. Indeed, it may suggest that they are at the core of these networks, initiating various cascades that result in these described roles, as opposed to peripheral components. The finding that TPRX1 is at the centre of the 8-cell-like cell GRN lends support to this contention[101–103].

Data from multiple sections of this thesis suggest that one of these roles is in blastocyst formation. The presence of an inner cell mass (ICM) in the blastocyst is a eutherian novelty[341]. Along with several other innovations, such as embryonic compaction, it adds extra complexity to eutherian preimplantation development that is not present in their sister group, the metatherians[127,341]. Previous work suggested that ETCHbox genes evolved by a mechanism denoted duplication-degeneration-complementation+ (DDC+)[92], taking on the embryonic role of the ancestral CRX protein while also gaining further functions. I propose that these additional functions involve coordinating the development of the more complex eutherian blastocyst. Two lines of evidence support this. First, bovine ectopic expression experiments showed that both ARGFX and LEUTX upregulate blastocyst-specific genes with functions that suggest they are involved in the regulation of blastocyst structure. Second, I found that ARGFX is highly specific to the novel eutherian ICM, and that it upregulates key ICM genes like *BAMBI*. This demonstrates the recruitment of an ETCHbox protein to the development of a lineage-specific trait, and suggests that the reason multiple genes (ETCHbox) are required in eutherians for a function ably performed by only CRX in metatherians is the more complex nature of the eutherian blastocyst.

7.4.2 ETCHbox proteins and cellular potency

The work presented in Chapter 3 established that across humans, mice and cattle, ETCHbox proteins downregulate critically important pluripotency genes including several of the OKSM factors used for reprogramming somatic cells into induced pluripotent stem cells[148,149]. The downregulation of key pluripotency factors at first may appear

counterintuitive, but I propose that it is indeed consistent with expectations based on recently published data. As described in Chapter 1 Section 1.3.4, a small percentage of cells in human pluripotent embryonic stem cell (ESC) cultures (which resemble cells from the epiblast), termed 8-cell-like cells (8CLCs), are thought to be totipotent and exhibit properties highly similar to 8-cell blastomeres[101–103]. *LEUTX* and *TPRX1* are not expressed in the standard pluripotent ESCs but are strongly upregulated in 8CLCs, and indeed *TPRX1* knockout prevents the acquisition of the 8-cell-like state[101–103]. The transition from an ESC state to 8CLC state involves steep downregulation of pluripotency factors including OKSM, suggesting their presence precludes the acquisition of a totipotent state. This provides strong evidence that several ETCHbox genes are integrated into the totipotency network rather than the pluripotency network and explains their downregulation of pluripotency genes. In addition, it bears mentioning that pluripotency-related genes are conspicuous in their absence from the downregulated gene set of human ARGFX. This is consistent with the protracted expression of *ARGFX* which persists into the pluripotent epiblast.

7.4.3 ETCHbox proteins and transposable elements

My discussion in Chapter 5 is extensive, so I will keep this brief, but the link between ETCHbox function and TE dynamics bears highlighting. I found that several human and bovine ETCHbox proteins downregulate myriad TE-repressive genes and subsequently aimed to test the biological importance of this by quantifying the effect of ETCHbox expression on retrotransposition rate and transposon expression levels. However, the use of cell culture models for these assays leaves open the question of whether the results reflect the events occurring in the preimplantation embryo. Despite this, the establishment of this link between ETCHbox proteins and TEs lays the foundation for further investigation. In particular, my retrotransposition assay tested just one protein (TPRX1) from one species (human) against one TE (LINE-1), but there is scope for a significant expansion of this assay to further characterise this effect. Indeed, I optimised

conditions for an Alu assay, but consumable costs prevented a full assay from being completed. Follow-up work should aim to test the effects of human ARGFX and LEUTX, and include both LINE-1 and Alu treatments.

7.5 Perspectives and future work

The nature of my curiosity and biological interests, combined with the academic environment in which I have worked, led me to take a strongly evolution-centric approach to understanding the peculiar nature of ETCHbox genes. From this has emerged, in my opinion, a more comprehensive understanding of the characteristics of their fast evolution. I have answered key questions regarding, for instance, (i) whether the designation as fast-evolving genes endures now a broader selection of high-quality genomes is available; (ii) whether such rapid evolution has continued between closely-related species; and (iii) whether observed sequence changes have resulted in changes to protein function. However, while my results lend significant support to genetic functional redundancy as an important factor permitting the rapid evolution we observe, I conclude above that a complete understanding of these factors has yet to be realised. In my opinion, while significant progress has been made toward answering this question, our incomplete understanding of the genes' functional role in embryonic development is obstructing further advancements. As such, I think a change in the methodological approach is required, refocusing on the molecular function of ETCHbox proteins. Only once we understand the selected role these proteins perform during the very earliest stages of development will we be able to fully elucidate the drivers of their rapid evolution.

I argue that the recent paper by Zou et al. (2022)[94] is a perfect example of this, as their in vivo *TPRX* morpholino knockdown experiments, while aiming to understand the functional role of *TPRX* genes, also provide strong evidence for redundancy. CRISPR/Cas9-mediated knockouts are now possible in human[745], non-human primate[746]

and ungulate[747,748] embryos, and knockouts of ETCHbox genes both alone and in combination should be prioritised to test the genes' necessity for embryonic development. Moreover, either chromatin immunoprecipitation (ChIP)-sequencing or cleavage under targets and release using nuclease (CUT&RUN)-sequencing to establish binding sites of ETCHbox transcription factors would focus the list of differentially expressed genes from ectopic expression experiments to those which are direct targets.

In addition to this, to facilitate a more complete understanding of the position ETCHbox proteins take up in the GRNs active in the preimplantation embryo, and to pinpoint their role in EGA, it is necessary to identify the factors which regulate the expression of the genes themselves. Prior to this thesis, minimal attention has been paid to this question, but my GRN reconstruction approach using single-cell RNA-sequencing data identified a highly promising candidate in DUXA. It is tantalising that the ETCHbox proteins TPRX1 and DUXA were found to be two of the three key regulators of the 8CLC GRN[101], suggesting a critical and cooperative role in GRNs at the 8-cell stage of development. Future work should aim to test the association I identified by using ChIP-seq to search for evidence of DUXA binding at ETCHbox loci, and given the ubiquity of reciprocal interactions during preimplantation development the reverse experiment would also be informative.

In Chapters 3 and 4 I ectopically expressed bovine and then primate ETCHbox genes in fibroblast cell culture. My motivation for this choice in both cases was to facilitate comparisons with previous work in which fibroblasts were also used. However, a potential criticism that could be levelled at this approach is that the genomic environment of a fibroblast (for example, chromatin state) varies significantly from that of the preimplantation embryo in which the genes are natively expressed, and therefore that transcriptional activity of ETCHbox proteins in this setting may not reflect their role in the embryo. I argue that previous work[90,96] and my experiments show conclusively that

expression of ETCHbox genes in fibroblasts results in the activation of embryonic genes. Further, induced pluripotent stem cell reprogramming experiments show unequivocally that expression of embryonic transcription factors in somatic cells can result in spectacular changes to a cell's properties and even the transition to an embryonic-like state[148,149]. However, it is undoubtedly true that ESCs or, even more so, 8CLCs are a closer facsimile of the preimplantation embryo, and I recommend their use for future experiments where practical. In particular, an ectopic expression approach was pursued precisely because no cell lines existed with native ETCHbox expression to be used for knockdown or knockout experiments. The 2022 discovery of 8CLCs, which express ETCHbox genes to a high level, offers a timely opportunity for such experiments to probe the effects of the absence of ETCHbox proteins in an environment in which they are natively expressed but without the tribulations of in vivo experiments.

7.6 Concluding remarks

The fundamental question at the heart of this thesis is why several groups of embryonically-expressed genes have cast off the tight evolutionary constraints that curtail the divergence of many developmental gene families and undergone extraordinary changes to their sequences and copy number within relatively short evolutionary timescales. To tackle this problem, I have taken the eutherian-specific ETCHbox group as a model to understand the forces dictating the evolution of such genes. This work combines comparative genomics and transcriptomics with an ectopic expression approach to expand our understanding of the characteristics of ETCHbox genes' rapid evolution while also offering insights into their functional role in the preimplantation embryo. I suggest that ETCHbox proteins are multifunctional regulators of several key processes during preimplantation development, including the structural formation of the blastocyst; ICM development; the activity of transposable elements; and transitions in cell potency. I propose mutational robustness via functional redundancy as

an important factor permitting rapid evolutionary change, but my evidence also suggests that positive selection for changes to transcription factor targets has led to divergence of protein-coding sequences. I am convinced that fast-evolving genes will continue their emergence as important contributors to embryonic development and morphological diversity, and the work in this thesis provides the foundations for a more complete understanding of the mechanisms underlying this.

References

1. Lewis EB. A gene complex controlling segmentation in Drosophila. *Nature*. 1978. 276(5688).
2. Izpisúa-Belmonte JC, Falkenstein H, Dollé P, Renucci A, Duboule D. Murine genes related to the Drosophila AbdB homeotic gene are sequentially expressed during development of the posterior part of the body. *EMBO J*. 1991. 10(8).
3. Gaunt SJ, Sharpe PT, Duboule D. Spatially restricted domains of homeo-gene transcripts in mouse embryos: Relation to a segmented body plan. *Development*. 1988. 104(Supplement).
4. Duboule D. The vertebrate limb: a model system to study the Hox/HOM gene network during development and evolution. *BioEssays*. 1992. 14(6).
5. Dollé P, Izpisúa-Belmonte JC, Falkenstein H, Renucci A, Duboule D. Coordinate expression of the murine Hox-5 complex homoeobox-containing genes during limb pattern formation. *Nature*. 1989. 342(6251).
6. Peterson KJ, Lyons JB, Nowak KS, Takacs CM, Wargo MJ, McPeek MA. Estimating metazoan divergence times with a molecular clock. *Proc Natl Acad Sci U S A*. 2004. 101(17).
7. Duboule D, Dollé P. The structural and functional organization of the murine HOX gene family resembles that of Drosophila homeotic genes. *EMBO J*. 1989. 8(5).
8. Graham A, Papalopulu N, Krumlauf R. The murine and Drosophila homeobox gene complexes have common features of organization and expression. *Cell*. 1989. 57(3).
9. Akam M. Hox and HOM: Homologous gene clusters in insects and vertebrates. *Cell*. 1989. 57(3).
10. McGinnis W, Levine MS, Hafen E, Kuroiwa A, Gehring WJ. A conserved DNA sequence in homoeotic genes of the Drosophila Antennapedia and bithorax complexes. *Nature*. 1984. 308(5958).
11. Scott MP, Weiner AJ. Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila. *Proc Natl Acad Sci U S A*. 1984. 81(13).
12. Derelle R, Lopez P, Le Guyader H, Manuel M. Homeodomain proteins belong to the ancestral molecular toolkit of eukaryotes. *Evol Dev*. 2007. 9(3).
13. de Mendoza A, Sebé-Pedrós A, Šestak MS, et al. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages.

Proc Natl Acad Sci U S A. 2013. 110(50).

14. Holland PWH. Evolution of homeobox genes. *Wiley Interdiscip Rev Dev Biol*. 2013. 2(1).
15. Zhong YF, Holland PWH. The dynamics of vertebrate homeobox gene evolution: Gain and loss of genes in mouse and human lineages. *BMC Evol Biol*. 2011. 11:169.
16. De Robertis EM. The homeobox in cell differentiation and evolution. In: Duboule D, ed. *Guidebook to the Homeobox Genes*. Oxford: Oxford University Press. 1994. pp.13-23.
17. Niehrs C, Keller R, Cho KWY, De Robertis EM. The homeobox gene goosecoid controls cell migration in Xenopus embryos. *Cell*. 1993. 72(4).
18. Del Bene F, Wittbrodt J. Cell cycle control by homeobox genes in development and disease. *Semin Cell Dev Biol*. 2005. 16(3).
19. Taniguchi Y. Hox transcription factors: Modulators of cell-cell and cell-extracellular matrix adhesion. *Biomed Res Int*. 2014. 2014:591374.
20. Gellon G, McGinnis W. Shaping animal body plans in development and evolution by modulation of Hox expression patterns. *BioEssays*. 1998. 20(2).
21. Holland PWH. Did homeobox gene duplications contribute to the Cambrian explosion? *Zool Lett*. 2015. 1(1).
22. Strumpf D, Mao CA, Yamanaka Y, et al. Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst. *Development*. 2005. 132(9).
23. Nichols J, Zevnik B, Anastassiadis K, et al. Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell*. 1998. 95(3).
24. Mitsui K, Tokuzawa Y, Itoh H, et al. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell*. 2003. 113(5).
25. Brooke NM, Garcia-Fernàndez J, Holland PWH. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature*. 1998. 392(6679).
26. Wagner GP, Amemiya C, Ruddle F. Hox cluster duplications and the opportunity for evolutionary novelties. *Proc Natl Acad Sci U S A*. 2003. 100(25).
27. Duverger O, Morasso MI. Role of homeobox genes in the patterning, specification, and differentiation of ectodermal appendages in mammals. *J Cell Physiol*. 2008. 216(2).
28. Davidson EH, Erwin DH. Gene regulatory networks and the evolution of animal body plans. *Science*. 2006. 311(5762).
29. Carroll SB. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic

Theory of Morphological Evolution. *Cell*. 2008. 134(1).

30. Carroll SB. Endless Forms: The Evolution of Gene Regulation and Morphological Diversity. *Cell*. 2000. 101(6).
31. Manak JR, Scott MP. A class act: Conservation of homeodomain protein functions. *Development*. 1994. 120(Supplement).
32. Burke AC, Nelson CE, Morgan BA, Tabin C. Hox genes and the evolution of vertebrate axial morphology. *Development*. 1995. 121(2).
33. Guerreiro I, Nunes A, Woltering JM, et al. Role of a polymorphism in a Hox/Pax-responsive enhancer in the evolution of the vertebrate spine. *Proc Natl Acad Sci U S A*. 2013. 110(26).
34. Brunetti CR, Selegue JE, Monteiro A, French V, Brakefield PM, Carroll SB. The generation and diversification of butterfly eyespot color patterns. *Curr Biol*. 2001. 11(20).
35. Shapiro MD, Marks ME, Peichel CL, et al. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*. 2004. 428(6984).
36. Chan YF, Marks ME, Jones FC, et al. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a pitxl enhancer. *Science*. 2010. 327(5963).
37. Holland PWH, Marlétaz F, Maeso I, Dunwell TL, Paps J. New genes from old: Asymmetric divergence of gene duplicates and the evolution of development. *Philos Trans R Soc B Biol Sci*. 2017. 372(1713).
38. Conant GC, Wagner A. Asymmetric sequence divergence of duplicate genes. *Genome Res*. 2003. 13(9).
39. Chai CL, Zhang Z, Huang FF, et al. A genomewide survey of homeobox genes and identification of novel structure of the Hox cluster in the silkworm, Bombyx mori. *Insect Biochem Mol Biol*. 2008. 38(12).
40. Ferguson L, Marlétaz F, Carter JM, et al. Ancient Expansion of the Hox Cluster in Lepidoptera Generated Four Homeobox Genes Implicated in Extra-Embryonic Tissue Formation. *PLoS Genet*. 2014. 10(10).
41. Mulhair PO, Crowley L, Boyes DH, et al. Diversity, duplication and genomic organisation of homeobox genes in Lepidoptera. *Genome Res*. 2023. 33(1).
42. Mulhair PO, Holland PWH. Evolution of the insect Hox gene cluster: Comparative analysis across 243 species. *Semin Cell Dev Biol*. 2022. S1084-9521(22).
43. Stauber M, Jäckle H, Schmidt-Ott U. The anterior determinant bicoid of Drosophila is a derived Hox class 3 gene. *Proc Natl Acad Sci U S A*. 1999. 96(7).

44. Falciani F, Hausdorf B, Schröder R, et al. Class 3 Hox genes in insects and the origin of zen. *Proc Natl Acad Sci U S A*. 1996. 93(16).
45. Driever W, Nüsslein-Volhard C. The bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner. *Cell*. 1988. 54(1).
46. Lambert JD. Developmental Patterns in Spiralian Embryos. *Curr Biol*. 2010. 20(2).
47. Morino Y, Hashimoto N, Wada H. Expansion of TALE homeobox genes and the evolution of spiralian development. *Nat Ecol Evol*. 2017. 1(12).
48. Morino Y. Dynamic evolutionary history of spiralian-specific TALE homeobox genes in mollusks. *Dev Growth Differ*. 2022. 64(4).
49. Barton-Owen TB, Szabó R, Somorjai IML, Ferrier DEK. A revised spiralian homeobox gene classification incorporating new polychaete transcriptomes reveals a diverse TALE class and a divergent hox gene. *Genome Biol Evol*. 2018. 10(9).
50. Phuangphong S, Tsunoda J, Wada H, Morino Y. Duplication of spiralian-specific TALE genes and evolution of the blastomere specification mechanism in the bivalve lineage. *Evodevo*. 2021. 12(1).
51. Henry JQ. Spiralian model systems. *Int J Dev Biol*. 2014. 58(6-8).
52. Dunwell TL, Holland PWH. A sister of NANOG regulates genes expressed in pre-implantation human development. *Open Biol*. 2017. 7(4).
53. Zhang YE, Long M. New genes contribute to genetic and phenotypic novelties in human evolution. *Curr Opin Genet Dev*. 2014. 29.
54. Villanueva-Cañas JL, Ruiz-Orera J, Agea MI, Gallo M, Andreu D, Albà MM. New genes and functional innovation in mammals. *Genome Biol Evol*. 2017. 9(7).
55. Clapp J, Mitchell LM, Bolland DJ, et al. Evolutionary conservation of a coding function for D4Z4, the tandem DNA repeat mutated in facioscapulohumeral muscular dystrophy. *Am J Hum Genet*. 2007. 81(2).
56. Leidenroth A, Hewitt JE. A family history of DUX4: phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. *BMC Evol Biol*. 2010. 10:364.
57. Macfarlan TS, Gifford WD, Driscoll S, et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*. 2012. 487(7405).
58. Peaston AE, Evsikov A V., Graber JH, et al. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev Cell*. 2004. 7(4)
59. Rodriguez-Terrones D, Torres-Padilla ME. Nimble and Ready to Mingle: Transposon Outbursts of Early Development. *Trends Genet*. 2018. 34(10).
60. De Iaco A, Planet E, Coluccio A, Verp S, Duc J, Trono D. DUX-family

transcription factors regulate zygotic genome activation in placental mammals. *Nat Genet*. 2017. 49(6).

61. Hendrickson PG, Doráis JA, Grow EJ, et al. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat Genet*. 2017. 49(6).
62. Eidahl JO, Giesige CR, Domire JS, et al. Mouse Dux is myotoxic and shares partial functional homology with its human paralog DUX4. *Hum Mol Genet*. 2016. 25(20).
63. Yoshihara M, Kirjanov I, Nykänen S, et al. Transient DUX4 expression in human embryonic stem cells induces blastomere-like expression program that is marked by SLC34A2. *Stem Cell Reports*. 2022. 17(7).
64. Vuoristo S, Bhagat S, Hydén-Granskog C, et al. DUX4 is a multifunctional factor priming human embryonic genome activation. *iScience*. 2022. 25(4).
65. Hermant C, Torres-Padilla ME. TFs for TEs: The transcription factor repertoire of mammalian transposable elements. *Genes Dev*. 2021. 35(1).
66. Whiddon JL, Langford AT, Wong CJ, Zhong JW, Tapscott SJ. Conservation and innovation in the DUX4-family gene network. *Nat Genet*. 2017. 49(6).
67. Leidenroth A, Clapp J, Mitchell LM, et al. Evolution of DUX gene macrosatellites in placental mammals. *Chromosoma*. 2012. 121(5).
68. MacLean JA, Chen MA, Wayne CM, et al. Rhox: A new homeobox gene cluster. *Cell*. 2005. 120(3).
69. Niu AL, Wang YQ, Zhang H, et al. Rapid evolution and copy number variation of primate RHOXF2, an X-linked homeobox gene involved in male reproduction and possibly brain function. *BMC Evolutionary Biology*. 2011. 11:298.
70. Song HW, Anderson RA, Bayne RA, et al. The RHOX homeobox gene cluster is selectively expressed in human oocytes and male germ cells. *Hum Reprod*. 2013. 28(6).
71. Wayne CM, MacLean JA, Cornwall G, Wilkinson MF. Two novel human X-linked homeobox genes, hPEPP1 and hPEPP2, selectively expressed in the testis. *Gene*. 2002. 301(1-2).
72. MacLean JA, Wilkinson MF. The Rhox genes. *Reproduction*. 2010. 140(2).
73. Sutton KA, Wilkinson MF. Rapid evolution of a homeodomain: Evidence for positive selection. *J Mol Evol*. 1997. 45(6).
74. Maiti S, Doskow J, Sutton K, et al. The Pem homeobox gene: Rapid evolution of the homeodomain, X chromosomal localization, and expression in reproductive tissue. *Genomics*. 1996. 34(3).
75. Madissoon E, Jouhilahti EM, Vesterlund L, et al. Characterization and target

genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. *Sci Rep*. 2016. 6:28995.

76. Li H, Tsai M-S, Chen C-Y, et al. A novel maternally transcribed homeobox gene, Eso-1, is preferentially expressed in oocytes and regulated by cytoplasmic polyadenylation. *Mol Reprod Dev*. 2006. 73(7).
77. Wilming LG, Boychenko V, Harrow JL. Comprehensive comparative homeobox gene annotation in human and mouse. *Database*. 2015. 2015:bav091.
78. Holland PWH, Booth HAF, Bruford EA. Classification and nomenclature of all human homeobox genes. *BMC Biol*. 2007. 5:47.
79. Furukawa T, Morrow EM, Cepko CL. Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. *Cell*. 1997. 91(4).
80. Samson M, Emerson MM, Cepko CL. Robust marking of photoreceptor cells and pinealocytes with several reporters under control of the Crx gene. *Dev Dyn*. 2009. 238(12).
81. Furukawa T, Morrow EM, Li T, Davis FC, Cepko CL. Retinopathy and attenuated circadian entrainment in Crx-deficient mice. *Nat Genet*. 1999. 23(4).
82. Sohocki MM, Sullivan LS, Mintz-Hittner HA, et al. A Range of Clinical Phenotypes Associated with Mutations in CRX, a Photoreceptor Transcription-Factor Gene. *Am J Hum Genet*. 1998. 63(5).
83. Freund CL, Gregory-Evans CY, Furukawa T, et al. Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor. *Cell*. 1997. 91(4).
84. Chen S, Wang QL, Nie Z, et al. Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. *Neuron*. 1997. 19(5).
85. Plouhinec JL, Sauka-Spengler T, Germot A, et al. 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*. 2003. 20(4).
86. Vignali R, Colombetti S, Lupo G, et al. Xotx5b, a new member of the Otx gene family, may be involved in anterior and eye development in Xenopus laevis. *Mech Dev*. 2000. 96(1).
87. Acampora D, Gulisano M, Simeone A. Genetic and molecular roles of Otx homeodomain proteins in head development. *Gene*. 2000. 246(1).
88. Boroviak T, Stirparo GG, Dietmann S, et al. Single cell transcriptome analysis of

human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. *Development*. 2018. 145(21).

89. Bürglin TR. Homeodomain subtypes and functional diversity. *Subcell Biochem*. 2011. 52.
90. Maeso I, Dunwell TL, Wyatt CDR, et al. Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals. *BMC Biol*. 2016. 14:45.
91. Booth HAF, Holland PWH. Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. *Gene*. 2007. 387(1-2)
92. Royall AH, Frankenberg S, Pask AJ, Holland PWH. Of eyes and embryos: subfunctionalization of the CRX homeobox gene in mammalian evolution. *Proc R Soc B Biol Sci*. 2019. 286(1907).
93. Li G, Holland PW. The origin and evolution of ARGFX homeobox loci in mammalian radiation. *BMC Evol Biol*. 2010. 10:182.
94. Zou Z, Zhang C, Wang Q, et al. Translatome and transcriptome co-profiling reveals a role of TPRXs in human zygotic genome activation. *Science*. 2022. 378(6615).
95. Katayama S, Ranga V, Jouhilahti EM, et al. Phylogenetic and mutational analyses of human LEUTX, a homeobox gene implicated in embryogenesis. *Sci Rep*. 2018. 8(1).
96. Royall AH, Maeso I, Dunwell TL, Holland PWH. Mouse Obox and Crxos modulate preimplantation transcriptional profiles revealing similarity between paralogous mouse and human homeobox genes. *Evodevo*. 2018. 9:2.
97. Vanin EF. Processed pseudogenes: characteristics and evolution. *Annu Rev Genet*. 1985. 19.
98. Maestre J, Tchénio T, Dhellin O, Heidmann T. mRNA retroposition in human cells: processed pseudogene formation. *EMBO J*. 1995. 14(24).
99. Dunwell TL, Holland PWH. Diversity of human and mouse homeobox gene expression in development and adult tissues. *BMC Dev Biol*. 2016. 16(1).
100. Töhönen V, Katayama S, Vesterlund L, et al. Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. *Nat Commun*. 2015. 6:8207.
101. Mazid MA, Ward C, Luo Z, et al. Rolling back of human pluripotent stem cells to an 8-cell embryo-like stage. *Nature*. 2022. 605(7909).
102. Taubenschmid-Stowers J, Rostovskaya M, Santos F, et al. 8C-like cells capture the human zygotic genome activation program in vitro. *Cell Stem Cell*. 2022. 29(3).

103. Yu X, Liang S, Chen M, et al. Recapitulating early human development with 8C-like cells. *Cell Rep*. 2022. 39(12).
104. Jouhilahti E-M, Madissoon E, Vesterlund L, et al. The human PRD-like homeobox gene LEUTX has a central role in embryo genome activation. *Development*. 2016. 143(19).
105. Gehring WJ, Affolter M, Bürglin T. Homeodomain proteins. *Annu Rev Biochem*. 1994. 63(1).
106. Bürglin TR, Affolter M. Homeodomain proteins: an update. *Chromosoma*. 2016. 125(3).
107. Bruun JA, Thomassen EIS, Kristiansen K, et al. The third helix of the homeodomain of paired class homeodomain proteins acts as a recognition helix both for DNA and protein interactions. *Nucleic Acids Res*. 2005. 33(8).
108. Wilson D, Sheng G, Lecuit T, Dostatni N, Desplan C. Cooperative dimerization of paired class homeo domains on DNA. *Genes Dev*. 1993. 7(11).
109. Hanes SD, Brent R. DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9. *Cell*. 1989. 57(7).
110. Schier AF, Gehring WJ. Direct homeodomain-DNA interaction in the autoregulation of the fushi tarazu gene. *Nature*. 1992. 356(6372).
111. Ménézo YJR, Hérubel F. Mouse and bovine models for human IVF. *Reprod Biomed Online*. 2002. 4(2).
112. Rajkovic A, Yan C, Yan W, Klysik M, Matzuk MM. Obox, a family of homeobox genes preferentially expressed in germ cells. *Genomics*. 2002. 79(5).
113. Cheng WC, Hsiu MHL, Yeh YJ, Li H. Mice lacking the Obox6 homeobox gene undergo normal early embryonic development and are fertile. *Dev Dyn*. 2007. 236(9).
114. Saito K, Abe H, Nakazawa M, et al. Cloning of complementary DNAs encoding structurally related homeoproteins from preimplantation mouse embryos: Their involvement in the differentiation of embryonic stem cells. *Biol Reprod*. 2010. 82(4).
115. Cui W, Dai X, Marcho C, et al. Towards Functional Annotation of the Preimplantation Transcriptome: An RNAi Screen in Mammalian Embryos. *Sci Rep*. 2016. 6:37396.
116. Chen Z, Zhang Y. Loss of DUX causes minor defects in zygotic genome activation and is compatible with mouse development. *Nat Genet*. 2019. 51(6).
117. Bosnakovski D, Gearhart MD, Ho Choi S, Kyba M. Dux facilitates post-implantation development, but is not essential for zygotic genome activation. *Biol Reprod*. 2021. 104(1).

118. De Iaco A, Verp S, Offner S, Grun D, Trono D. DUX is a non-essential synchronizer of zygotic genome activation. *Development*. 2020. 147(2).
119. Guo Y, Kitano T, Murano K, et al. Obox4 secures zygotic genome activation upon loss of Dux. *bioRxiv*. 2022. 2022.07.04.498763.
120. Grimwood J, Gordon LA, Olsen A, et al. The DNA sequence and biology of human chromosome 19. *Nature*. 2004. 428(6982).
121. Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA. Chromatin architecture of the human genome: Gene-rich domains are enriched in open chromatin fibers. *Cell*. 2004. 118(5).
122. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001. 409(6822).
123. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. Genetics: A fine-scale map of recombination rates and hotspots across the human genome. *Science*. 2005. 310(5746).
124. Wagner A. Distributed robustness versus redundancy as causes of mutational robustness. *BioEssays*. 2005. 27(2).
125. Wagner A. Genetic redundancy caused by gene duplications and its evolution in networks of transcriptional regulators. *Biol Cybern*. 1996. 74(6).
126. Kafri R, Springer M, Pilpel Y. Genetic Redundancy: New Tricks for Old Genes. *Cell*. 2009. 136(3).
127. Renfree MB. Review: Marsupials: Placental Mammals with a Difference. *Placenta*. 2010. 31(Suppl).
128. Freyer C, Renfree MB. The mammalian yolk sac placenta. *J Exp Zool Part B Mol Dev Evol*. 2009. 312(6).
129. Göke J, Lu X, Chan YS, et al. Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell*. 2015. 16(2).
130. Grow EJ, Flynn RA, Chavez SL, et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature*. 2015. 522(7555).
131. Yandım C, Karakülah G. Expression dynamics of repetitive DNA in early human embryonic development. *BMC Genomics*. 2019. 20(1).
132. Tan K, Kim ME, Song HW, et al. The RHOX gene cluster suppresses germline LINE1 transposition. *Proc Natl Acad Sci U S A*. 2021. 118(23).
133. Patil S, Totey S. Developmental failure of hybrid embryos generated by in vitro fertilization of water buffalo (Bubalus bubalis) oocyte with bovine spermatozoa. *Mol Reprod Dev*. 2003. 64(3).
134. Condic ML. Totipotency: What it is and what it is not. *Stem Cells Dev*. 2014.

23(8).

135. Morgani SM, Brickman JM. The molecular underpinnings of totipotency. *Philos Trans R Soc B Biol Sci*. 2014. 369(1657).
136. Suwińska A, Czołowska R, Ozdzeński W, Tarkowski AK. 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. *Dev Biol*. 2008. 322(1).
137. Riveiro AR, Brickman JM. From pluripotency to totipotency: an experimentalist's guide to cellular potency. *Development*. 2020. 147(16).
138. Reya T, Morrison SJ, Clarke MF, Weissman IL. Stem cells, cancer, and cancer stem cells. *Nature*. 2001. 414(6859).
139. Martello G, Smith A. The Nature of Embryonic Stem Cells. *Annu Rev Cell Dev Biol*. 2014. 30(1).
140. Evans MJ, Kaufman MH. Establishment in culture of pluripotential cells from mouse embryos. *Nature*. 1981. 292(5819).
141. Martin GR. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc Natl Acad Sci U S A*. 1981. 78(12).
142. Thomson JA. Embryonic stem cell lines derived from human blastocysts. *Science*. 1998. 282(5391).
143. Reubinoff BE, Pera MF, Fong CY, Trounson A, Bongso A. Embryonic stem cell lines from human blastocysts: Somatic differentiation in vitro. *Nat Biotechnol*. 2000. 18(4).
144. Bradley A, Evans M, Kaufman MH, Robertson E. Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. *Nature*. 1984. 309(5965).
145. Wray J, Kalkan T, Smith AG. The ground state of pluripotency. *Biochem Soc Trans*. 2010. 38(4).
146. Ying QL, Wray J, Nichols J, et al. The ground state of embryonic stem cell self-renewal. *Nature*. 2008. 453(7194).
147. Nichols J, Jones K, Phillips JM, et al. Validated germline-competent embryonic stem cell lines from nonobese diabetic mice. *Nat Med*. 2009. 15(7).
148. Takahashi K, Yamanaka S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*. 2006. 126(4).
149. Takahashi K, Tanabe K, Ohnuki M, et al. Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell*. 2007. 131(5).
150. Wernig M, Meissner A, Foreman R, et al. In vitro reprogramming of fibroblasts

into a pluripotent ES-cell-like state. *Nature*. 2007. 448(7151).

151. Okita K, Ichisaka T, Yamanaka S. Generation of germline-competent induced pluripotent stem cells. *Nature*. 2007. 448(7151).
152. Takahashi K, Yamanaka S. A decade of transcription factor-mediated reprogramming to pluripotency. *Nat Rev Mol Cell Biol*. 2016. 17(3).
153. Han J, Yuan P, Yang H, et al. Tbx3 improves the germ-line competency of induced pluripotent stem cells. *Nature*. 2010. 463(7284).
154. Schöler HR, Dressler GR, Balling R, Rohdewohld H, Gruss P. Oct-4: a germline-specific transcription factor mapping to the mouse t-complex. *EMBO J*. 1990. 9(7).
155. Okamoto K, Okazawa H, Okuda A, Sakai M, Muramatsu M, Hamada H. A novel octamer binding transcription factor is differentially expressed in mouse embryonic cells. *Cell*. 1990. 60(3).
156. Niwa H, Miyazaki JI, Smith AG. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat Genet*. 2000. 24(4).
157. Pardo M, Lang B, Yu L, et al. An Expanded Oct4 Interaction Network: Implications for Stem Cell Biology, Development, and Disease. *Cell Stem Cell*. 2010. 6(4).
158. van den Berg DLC, Snoek T, Mullin NP, et al. An Oct4-Centered Protein Interaction Network in Embryonic Stem Cells. *Cell Stem Cell*. 2010. 6(4).
159. Yuan H, Corbi N, Basilico C, Dailey L. Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. *Genes Dev*. 1995. 9(21).
160. Young RA. Control of the embryonic stem cell state. *Cell*. 2011. 144(6).
161. Silva J, Smith A. Capturing Pluripotency. *Cell*. 2008. 132(4).
162. Niwa H. How is pluripotency determined and maintained? *Development*. 2007. 134(4).
163. Chambers I, Smith A. Self-renewal of teratocarcinoma and embryonic stem cells. *Oncogene*. 2004. 23(43).
164. Nichols J, Smith A. Pluripotency in the embryo and in culture. *Cold Spring Harb Perspect Biol*. 2012. 4(8).
165. Ivanova N, Dobrin R, Lu R, et al. Dissecting self-renewal in stem cells with RNA interference. *Nature*. 2006. 442(7102).
166. Niwa H, Ogawa K, Shimosato D, Adachi K. A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells. *Nature*. 2009. 460(7251).
167. Zhang X, Zhang J, Wang T, Esteban MA, Pei D. Esrrb activates Oct4

transcription and sustains self-renewal and pluripotency in embryonic stem cells. *J Biol Chem*. 2008. 283(51).

168. Matoba R, Niwa H, Masui S, et al. Dissecting Oct3/4-regulated gene networks in embryonic stem cells by expression profiling. *PLoS One*. 2006. 1(1).
169. Schiebinger G, Shu J, Tabaka M, et al. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*. 2019. 176(4).
170. Wu L, Wu Y, Peng B, et al. Oocyte-Specific Homeobox 1, Obox1, Facilitates Reprogramming by Promoting Mesenchymal-to-Epithelial Transition and Mitigating Cell Hyperproliferation. *Stem Cell Reports*. 2017. 9(5).
171. Royall AH. Eutherian-specific homeobox genes in the preimplantation embryo: function, evolution, and applications. *PhD Thesis: University of Oxford*. 2019.
172. Kim HM, Ahn HJ, Lee HS, et al. Obox4 regulates the expression of histone family genes and promotes differentiation of mouse embryonic stem cells. *FEBS Lett*. 2010. 584(3).
173. Saito R, Yamasaki T, Nagai Y, et al. CrxOS maintains the self-renewal capacity of murine embryonic stem cells. *Biochem Biophys Res Commun*. 2009. 390(4).
174. Sato S, Nakazawa M, Kihara Y, et al. 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. 120(5).
175. Iha M, Watanabe M, Kihara Y, et al. 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. 143(4).
176. Morgani SM, Canham MA, Nichols J, et al. Totipotent Embryonic Stem Cells Arise in Ground-State Culture Conditions. *Cell Rep*. 2013. 3(6).
177. Canham MA, Sharov AA, Ko MSH, Brickman JM. Functional heterogeneity of embryonic stem cells revealed through translational amplification of an early endodermal transcript. *PLoS Biol*. 2010. 8(5).
178. Chambers I, Silva J, Colby D, et al. Nanog safeguards pluripotency and mediates germline development. *Nature*. 2007. 450(7173).
179. Toyooka Y, Shimosato D, Murakami K, Takahashi K, Niwa H. Identification and characterization of subpopulations in undifferentiated ES cell culture. *Development*. 2008. 135(5).
180. Singh AM, Hamazaki T, Hankowski KE, Terada N. A Heterogeneous Expression Pattern for Nanog in Embryonic Stem Cells. *Stem Cells*. 2007. 25(10).
181. Rodriguez-Terrones D, Gaume X, Ishiuchi T, et al. A molecular roadmap for the

emergence of early-embryonic-like cells in culture. *Nat Genet*. 2018. 50(1).

182. Li WZ, Wang ZW, Chen LL, et al. Hesx1 enhances pluripotency by working downstream of multiple pluripotency-associated signaling pathways. *Biochem Biophys Res Commun*. 2015. 464(3).
183. Serebrowsky AS. Genes scute and achaete in Drosophila melanogaster and a hypothesis of gene divergency. *Comp Rend Acad Sci URSS*. 1938. 14.
184. Bridges CB. The Bar “gene” a duplication. *Science*. 1936. 83(2148).
185. Lewis EB. Pseudoallelism and gene evolution. *Cold Spring Harb Symp Quant Biol*. 1951. 16.
186. Taylor JS, Raes J. Duplication and divergence: The evolution of new genes and old ideas. *Annu Rev Genet*. 2004. 38.
187. Ohno S. *Evolution by Gene Duplication*. Berlin: Springer-Verlag Berlin Heidelberg. 1970.
188. Holland PWH. Gene duplication: Past, present and future. *Semin Cell Dev Biol*. 1999. 10(5).
189. Conant GC, Wolfe KH. Turning a hobby into a job: How duplicated genes find new functions. *Nat Rev Genet*. 2008. 9(12).
190. Zhang J. Evolution by gene duplication: An update. *Trends Ecol Evol*. 2003. 18(6).
191. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science*. 2000. 290(5494).
192. Van Oss SB, Carvunis AR. De novo gene birth. *PLoS Genet*. 2019. 15(5).
193. Carvunis AR, Rolland T, Wapinski I, et al. Proto-genes and de novo gene birth. *Nature*. 2012. 487(7407).
194. Long M, Betrán E, Thornton K, Wang W. The origin of new genes: Glimpses from the young and old. *Nat Rev Genet*. 2003. 4(11).
195. Irwin NAT, Pittis AA, Richards TA, Keeling PJ. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat Microbiol*. 2022. 7(2).
196. Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet*. 2008. 9(8).
197. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. *Genome Res*. 2010. 20(10).
198. Hahn MW. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered*. 2009. 100(5).
199. Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. Gene duplication as a major force in evolution. *J Genet*. 2013. 92(1).

200. Ohta T. *Evolution and Variation of Multigene Families*. Berlin/New York: Springer-Verlag. 1980.
201. Ohta T. An extension of a model for the evolution of multigene families by unequal crossing over. *Genetics*. 1979. 91(3).
202. Smith GP. Unequal crossover and the evolution of multigene families. *Symp Quant Biol*. 1973. 38.
203. Pâques F, Haber JE. Multiple Pathways of Recombination Induced by Double-Strand Breaks in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev*. 1999. 63(2).
204. Brosius J. Retroposons - Seeds of evolution. *Science*. 1991. 251(4995).
205. Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet*. 2000. 24(4).
206. Kaessmann H, Vinckenbosch N, Long M. RNA-based gene duplication: Mechanistic and evolutionary insights. *Nat Rev Genet*. 2009. 10(1).
207. Bailey JA, Eichler EE. Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nat Rev Genet*. 2006. 7(7).
208. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res*. 2001. 11(6).
209. She X, Liu G, Ventura M, et al. A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. *Genome Res*. 2006. 16(5).
210. Samonte RV, Eichler EE. Segmental duplications and the evolution of the primate genome. *Nat Rev Genet*. 2002. 3(1).
211. Otto SP, Whitton J. Polyploid incidence and evolution. *Annu Rev Genet*. 2000. 34.
212. Vinckenbosch N, Dupanloup I, Kaessmann H. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A*. 2006. 103(9).
213. Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol*. 2005. 3(11).
214. Zhang J, Zhang YP, Rosenberg HF. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat Genet*. 2002. 30(4).
215. Moriyama Y, Koshiba-Takeuchi K. Significance of whole-genome duplications on the emergence of evolutionary novelties. *Brief Funct Genomics*. 2018. 17(5).
216. Freeling M, Thomas BC. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res*. 2006.

16(7).

217. Dehal P, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*. 2005. 3(10).
218. Ohno S, Wolf U, Atkin NB. Evolution from fish to mammals by gene duplication. *Hereditas*. 1968. 59(1).
219. Putnam NH, Butts T, Ferrier DEK, et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature*. 2008. 453(7198).
220. Simakov O, Marlétaz F, Yue JX, et al. Deeply conserved synteny resolves early events in vertebrate evolution. *Nat Ecol Evol*. 2020. 4(6).
221. Shimeld SM, Holland PWH. Vertebrate innovations. *Proc Natl Acad Sci U S A*. 2000. 97(9).
222. Holland PWH, Garcia-Fernàndez J. Hox genes and chordate evolution. *Dev Biol*. 1996. 173(2).
223. Sharman AC, Holland PWH. Conservation, duplication, and divergence of developmental genes during chordate evolution. *Netherlands J Zool*. 1996. 46(1-2).
224. Sauka-Spengler T, Meulemans D, Jones M, Bronner-Fraser M. Ancient Evolutionary Origin of the Neural Crest Gene Regulatory Network. *Dev Cell*. 2007. 13(3).
225. Martinez-Morales JR, Henrich T, Ramialison M, Wittbrodt J. New genes in the evolution of the neural crest differentiation program. *Genome Biol*. 2007. 8(3).
226. Braasch I, Volff JN, Schartl M. The endothelin system: Evolution of vertebrate-specific ligand-receptor interactions by three rounds of genome duplication. *Mol Biol Evol*. 2009. 26(4).
227. Holland LZ, Short S. Gene duplication, co-option and recruitment during the origin of the vertebrate brain from the invertebrate chordate brain. *Brain Behav Evol*. 2008. 72(2).
228. Innan H, Kondrashov F. The evolution of gene duplications: Classifying and distinguishing between models. *Nat Rev Genet*. 2010. 11(2).
229. Ohno S. So much “junk” DNA in our genome. *Brookhaven Symp Biol*. 1972. 23.
230. Kimura M. *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press. 1983.
231. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*. 1999. 151(4).
232. Des Marais DL, Rausher MD. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*. 2008. 454(7205).

233. Hughes AL. The evolution of functionally novel proteins after gene duplication. *Proc R Soc B Biol Sci*. 1994. 256(1346).
234. Kondrashov FA, Koonin E V. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet*. 2004. 20(7).
235. Kondrashov FA, Rogozin IB, Wolf YI, Koonin E V. Selection in the evolution of gene duplications. *Genome Biol*. 2002. 3(2).
236. Haldane JBS. The part played by recurrent mutation in evolution. *Am Nat*. 1933. 67(708).
237. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. Role of duplicate genes in genetic robustness against null mutations. *Nature*. 2003. 421(6918).
238. Clark AG. Invasion and maintenance of a gene duplication. *Proc Natl Acad Sci U S A*. 1994. 91(8).
239. Katju V, Lynch M. On the formation of novel genes by duplication in the caenorhabditis elegans genome. *Mol Biol Evol*. 2006. 23(5).
240. Lynch M, Katju V. The altered evolutionary trajectories of gene duplicates. *Trends Genet*. 2004. 20(11).
241. Proulx SR, Phillips PC. Allelic divergence precedes and promotes gene duplication. *Evolution*. 2006. 60(5).
242. Spofford JB. Heterosis and the Evolution of Duplications. *Am Nat*. 1969. 103(932).
243. Papp B, Pál C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. *Nature*. 2003. 424(6945).
244. Liang H, Plazonic KR, Chen J, Li WH, Fernández A. Protein under-wrapping causes dosage sensitivity and decreases gene duplicability. *PLoS Genet*. 2008. 4(1).
245. Chen JM, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. Gene conversion: Mechanisms, evolution and human disease. *Nat Rev Genet*. 2007. 8(10)
246. Schildkraut E, Miller CA, Nickoloff JA. Gene conversion and deletion frequencies during double-strand break repair in human cells are controlled by the distance between direct repeats. *Nucleic Acids Res*. 2005. 33(5).
247. Sugino RP, Innan H. Estimating the time to the whole-genome duplication and the duration of concerted evolution via gene conversion in yeast. *Genetics*. 2005. 171(1).
248. Sugino RP, Innan H. Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. *Trends Genet*. 2006. 22(12).
249. Walsh JB. Sequence-dependent gene conversion: can duplicated genes diverge

fast enough to escape conversion? *Genetics*. 1987. 117(3).

250. Teshima KM, Innan H. Neofunctionalization of duplicated genes under the pressure of gene conversion. *Genetics*. 2008. 178(3).
251. Fawcett JA, Innan H. Neutral and non-neutral evolution of duplicated genes with gene conversion. *Genes (Basel)*. 2011. 2(1).
252. Albalat R, Cañestro C. Evolution by gene loss. *Nat Rev Genet*. 2016. 17(7).
253. Putnam NH, Srivastava M, Hellsten U, et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*. 2007. 317(5834).
254. Kortschak RD, Samuel G, Saint R, Miller DJ. EST Analysis of the Cnidarian Acropora millepora Reveals Extensive Gene Loss and Rapid Sequence Divergence in the Model Invertebrates. *Curr Biol*. 2003. 13(24).
255. Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. Identification and analysis of unitary pseudogenes: Historic and contemporary gene losses in humans and other primates. *Genome Biol*. 2010. 11(3).
256. Tsai IJ, Zarowiecki M, Holroyd N, et al. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature*. 2013. 496(7443).
257. Annunziata R, Arnone MI. A dynamic regulatory network explains ParaHox gene control of gut patterning in the sea urchin. *Development*. 2014.
258. Zhong Y, Herrera-Úbeda C, Garcia-Fernàndez J, Li G, Holland PWH. Mutation of amphioxus Pdx and Cdx demonstrates conserved roles for ParaHox genes in gut, anus and tail patterning. *BMC Biol*. 2020. 18(1).
259. Hottes AK, Freddolino PL, Khare A, Donnell ZN, Liu JC, Tavazoie S. Bacterial Adaptation through Loss of Function. *PLoS Genet*. 2013. 9(7).
260. Sokurenko EV, Hasty DL, Dykhuizen DE. Pathoadaptive mutations: Gene loss and variation in bacterial pathogens. *Trends Microbiol*. 1999. 7(5).
261. Jain N, Li L, Hsueh YP, et al. Loss of allergen 1 confers a hypervirulent phenotype that resembles mucoid switch variants of Cryptococcus neoformans. *Infect Immun*. 2009. 77(1).
262. Sharma V, Hecker N, Roscito JG, Foerster L, Langer BE, Hiller M. A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nat Commun*. 2018. 9(1).
263. Novembre J, Galvani AP, Slatkin M. The geographic spread of the CCR5 Δ32 HIV-resistance allele. *PLoS Biol*. 2005. 3(11).
264. Dean M, Carrington M, Winkler C, et al. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. *Science*. 1996. 273(5283).

265. Armstrong DM, Jones JK, Wilson DE. Encyclopedia Britannica: Mammal.
266. Kemp TS. *Mammals: A Very Short Introduction*. Oxford: Oxford University Press. 2017.
267. Vaughan TA, Ryan JM, Czaplewski NJ. *Mammalogy*. 4th ed. Fort Worth: Saunders College Pub. 2000.
268. Kaas JH. The origin and evolution of neocortex: From early mammals to modern humans. *Prog Brain Res*. 2019. 250.
269. O'Leary MA, Bloch JI, Flynn JJ, et al. The placental mammal ancestor and the Post-K-Pg radiation of placentals. *Science.* 2013. 339(6120).
270. Kemp TS. *The Origin and Evolution of Mammals*. Oxford: Oxford University Press. 2005.
271. Pough FH, Janis CM, Hesier JB. *Vertebrate Life*. 9th ed. Boston: Pearson. 2013.
272. McGowen MR, Erez O, Romero R, Wildman DE. The evolution of embryo implantation. *Int J Dev Biol*. 2014. 58(2-4).
273. Mossman HW. *Vertebrate Fetal Membranes*. New Brunswick, NJ: Rutgers University Press. 1987.
274. Renfree MB, Suzuki S, Kaneko-Ishino T. The origin and evolution of genomic imprinting and viviparity in mammals. *Philos Trans R Soc B Biol Sci*. 2013. 368(1609).
275. Williamson TE, Brusatte SL, Wilson GP. The origin and early evolution of metatherian mammals: The Cretaceous record. *Zookeys*. 2014. (465).
276. Foley NM, Springer MS, Teeling EC. Mammal madness: Is the mammal tree of life not yet resolved? *Philos Trans R Soc B Biol Sci*. 2016. 371(1699).
277. Teeling EC, Hedges SB. Making the impossible possible: Rooting the tree of placental mammals. *Mol Biol Evol*. 2013. 30(9).
278. McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res*. 2012. 22(4).
279. Song S, Liu L, Edwards SV, Wu S. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc Natl Acad Sci U S A*. 2012. 109(37).
280. Tarver JE, dos Reis M, Mirarab S, et al. The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biol Evol*. 2016. 8(2).
281. Morgan CC, Foster PG, Webb AE, Pisani D, McInerney JO, O'Connell MJ. Heterogeneous Models Place the Root of the Placental Mammal Phylogeny. *Mol*

Biol Evol. 2013. 30(9).

282. Lu F, Zhang Y. Cell totipotency: Molecular features, induction, and maintenance. *Natl Sci Rev*. 2015. 2(2).
283. Wamaitha SE, Niakan KK. Human pre-gastrulation development. *Current Topics in Developmental Biology*. 2018. 128.
284. Aiken CEM, Swoboda PPL, Skepper JN, Johnson MH. The direct measurement of embryogenic volume and nucleo-cytoplasmic ratio during mouse pre-implantation development. *Reproduction*. 2004. 128(5).
285. Cockburn K, Rossant J. Making the blastocyst: Lessons from the mouse. *J Clin Invest*. 2010. 120(4).
286. Guo H, Zhu P, Yan L, et al. The DNA methylation landscape of human early embryos. *Nature*. 2014. 511(7511).
287. Smith ZD, Chan MM, Humm KC, et al. DNA methylation dynamics of the human preimplantation embryo. *Nature*. 2014. 511(7511).
288. Ross PJ, Sampaio R V. Epigenetic remodeling in preimplantation embryos: Cows are not big mice. *Anim Reprod*. 2018. 15(3).
289. Jukam D, Shariati SAM, Skotheim JM. Zygotic Genome Activation in Vertebrates. *Dev Cell*. 2017. 42(4).
290. Bachvarova R. Gene expression during oogenesis and oocyte development in mammals. *Dev Biol*. 1985. 1.
291. Braude P, Bolton V, Moore S. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. *Nature*. 1988. 332(6163).
292. Vassena R, Boué S, González-Roca E, et al. Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. *Development*. 2011. 138(17).
293. Taylor DM, Ray PF, Ao A, Winston RML, Handyside AH. Paternal transcripts for glucose-6-phosphate dehydrogenase and adenosine deaminase are first detectable in the human preimplantation embryo at the three- to four-cell stage. *Mol Reprod Dev*. 1997. 48(4).
294. Asami M, Lam BYH, Ma MK, et al. Human embryonic genome activation initiates at the one-cell stage. *Cell Stem Cell*. 2022. 29(2).
295. Vastenhouw NL, Cao WX, Lipshitz HD. The maternal-to-zygotic transition revisited. *Development*. 2019. 146(11).
296. Eckersley-Maslin MA, Alda-Catalinas C, Reik W. Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nat Rev Mol Cell Biol*. 2018. 19(7).

297. Firmin J, Maître JL. Morphogenesis of the human preimplantation embryo: bringing mechanics to the clinics. *Semin Cell Dev Biol*. 2021. 120.
298. Edwards RG, Purdy JM, Steptoe PC, Walters DE. The growth of human preimplantation embryos in vitro. *Am J Obstet Gynecol*. 1981. 141(4).
299. Nikas G, Ao A, Winston RML, Handyside AH. Compaction and surface polarity in the human embryo in vitro. *Biol Reprod*. 1996. 55(1).
300. Johnson MH, Ziomek CA. Induction of polarity in mouse 8-cell blastomeres: Specificity, geometry, and stability. *J Cell Biol*. 1981. 91(1).
301. Iwata K, Yumoto K, Sugishima M, et al. Analysis of compaction initiation in human embryos by using time-lapse cinematography. *J Assist Reprod Genet*. 2014. 31(4).
302. Niakan KK, Han J, Pedersen RA, Simon C, Pera RAR. Human pre-implantation embryo development. *Development*. 2012. 139(5).
303. Gardner RL, Rossant J. Investigation of the fate of 4-5 day post-coitum mouse inner cell mass cells by blastocyst injection. *Development*. 1979. 52(1).
304. Gardner RL. Investigation of cell lineage and differentiation in the extraembryonic endoderm of the mouse embryo. *Development*. 1982. 68(1).
305. Gardner RL, Gardner RL, Lawrence PA. Clonal analysis of early mammalian development. *Philos Trans R Soc London B, Biol Sci*. 1985. 312(1153).
306. Posfai E, Rovic I, Jurisicova A. The mammalian embryo's first agenda: Making trophectoderm. *Int J Dev Biol*. 2019. 63(3-5).
307. Kim S-M, Kim J-S. A Review of Mechanisms of Implantation. *Dev Reprod*. 2017. 21(4).
308. Gardner DK, Lane M, Stevens J, Schlenker T, Schoolcraft WB. Blastocyst score affects implantation and pregnancy outcome: Towards a single blastocyst transfer. *Fertil Steril*. 2000. 73(6).
309. Wilcox AJ, Baird DD, Weinberg CR. Time of Implantation of the Conceptus and Loss of Pregnancy. *N Engl J Med*. 1999. 340(23).
310. Russ AP, Wattler S, Colledge WH, et al. Eomesodermin is required for mouse trophoblast development and mesoderm formation. *Nature*. 2000. 404(6773).
311. Latos PA, Sienerth AR, Murray A, et al. Elf5-centered transcription factor hub controls trophoblast stem cell self-renewal and differentiation through stoichiometry sensitive shifts in target gene networks. *Genes Dev*. 2015. 29(23).
312. Wicklow E, Blij S, Frum T, et al. HIPPO Pathway Members Restrict SOX2 to the Inner Cell Mass Where It Promotes ICM Fates in the Mouse Blastocyst. *PLoS Genet*. 2014. 10(10).
313. Palmieri SL, Peter W, Hess H, Schöler HR. Oct-4 Transcription Factor Is

Differentially Expressed in the Mouse Embryo during Establishment of the First Two Extraembryonic Cell Lineages Involved in Implantation. *Dev Biol*. 1994. 166(1).

314. Chambers I, Colby D, Robertson M, et al. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell*. 2003. 113(5).
315. Avilion AA, Nicolis SK, Pevny LH, Perez L, Vivian N, Lovell-Badge R. Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev*. 2003. 17(1).
316. Niwa H, Toyooka Y, Shimosato D, et al. Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. *Cell*. 2005. 123(5).
317. Fujikura J, Yamato E, Yonemura S, et al. Differentiation of embryonic stem cells is induced by GATA factors. *Genes Dev*. 2002. 16(7).
318. Koutsourakis M, Langeveld A, Patient R, Beddington R, Grosveld F. The transcription factor GATA6 is essential for early extraembryonic development. *Development*. 1999. 126(4).
319. Bessonnard S, De Mot L, Gonze D, et al. Gata6, Nanog and Erk signaling control cell fate in the inner cell mass through a tristable regulatory network. *Development*. 2014. 141(19).
320. Shimosato D, Shiki M, Niwa H. Extra-embryonic endoderm cells derived from ES cells induced by GATA Factors acquire the character of XEN cells. *BMC Dev Biol*. 2007. 7.
321. Chazaud C, Yamanaka Y, Pawson T, Rossant J. Early Lineage Segregation between Epiblast and Primitive Endoderm in Mouse Blastocysts through the Grb2-MAPK Pathway. *Dev Cell*. 2006. 10(5).
322. Meilhac SM, Adams RJ, Morris SA, Danckaert A, Le Garrec JF, Zernicka-Goetz M. Active cell movements coupled to positional induction are involved in lineage segregation in the mouse blastocyst. *Dev Biol*. 2009. 331(2).
323. Plusa B, Piliszek A, Frankenberg S, Artus J, Hadjantonakis AK. Distinct sequential cell behaviours direct primitive endoderm formation in the mouse blastocyst. *Development*. 2008. 135(18).
324. Xenopoulos P, Kang M, Puliafito A, Di Talia S, Hadjantonakis AK. Heterogeneities in Nanog expression drive stable commitment to pluripotency in the mouse blastocyst. *Cell Rep*. 2015. 10(9).
325. Gerri C, McCarthy A, Alanis-Lobato G, et al. Initiation of a conserved trophectoderm program in human, cow and mouse embryos. *Nature*. 2020. 587(7834).

326. Akizawa H, Saito S, Kohri N, et al. Deciphering two rounds of cell lineage segregations during bovine preimplantation development. *FASEB J*. 2021. 35(10).
327. Chazaud C, Yamanaka Y. Lineage specification in the mouse preimplantation embryo. *Development*. 2016. 143(7).
328. Artus J, Piliszek A, Hadjantonakis AK. The primitive endoderm lineage of the mouse blastocyst: Sequential transcription factor activation and regulation of differentiation by Sox17. *Dev Biol*. 2011. 350(2).
329. Aoki F, Worrad DM, Schultz RM. Regulation of transcriptional activity during the first and second cell cycles in the preimplantation mouse embryo. *Dev Biol*. 1997. 181(2).
330. Hamatani T, Carter MG, Sharov AA, Ko MSH. Dynamics of global gene expression changes during mouse preimplantation development. *Dev Cell*. 2004. 6(1).
331. Flach G, Johnson MH, Braude PR, Taylor RA, Bolton VN. The transition from maternal to embryonic control in the 2-cell mouse embryo. *EMBO J*. 1982. 1(6).
332. Wang QT, Piotrowska K, Ciemerych MA, et al. A genome-wide study of gene activity reveals developmental signaling pathways in the preimplantation mouse embryo. *Dev Cell*. 2004. 6(1).
333. Hertig AT, Rock J, Adams EC, Menkin MC. Thirty-four fertilized human ova, good, bad and indifferent, recovered from 210 women of known fertility; a study of biologic wastage in early human pregnancy. *Pediatrics*. 1959. 23(1 Part 2).
334. Brinster RL. A Method for in vitro cultivation of mouse ova from two-cell to blastocyst. *Exp Cell Res*. 1963. 32(1).
335. Xie D, Chen CC, Ptaszek LM, et al. Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. *Genome Res*. 2010. 20(6).
336. Niakan KK, Eggan K. Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Dev Biol*. 2013. 375(1).
337. Berg DK, Smith CS, Pearton DJ, et al. Trophectoderm Lineage Determination in Cattle. *Dev Cell*. 2011. 20(2).
338. Kohri N, Akizawa H, Iisaka S, et al. Trophectoderm regeneration to support full-term development in the inner cell mass isolated from bovine blastocyst. *J Biol Chem*. 2019. 294(50).
339. Blakeley P, Fogarty NME, Del Valle I, et al. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development*. 2015. 142(18).

340. Singh M, Widmann TJ, Bansal V, et al. The selection arena in early human blastocysts resolves the pluripotent inner cell mass. *bioRxiv*. 2019. 318329.
341. Frankenberg SR, de Barros FRO, Rossant J, Renfree MB. The mammalian blastocyst. *WIREs Dev Biol*. 2016. 5(2):210-232.
342. Johnson MH, Selwood L. Nomenclature of early development in mammals. *Reprod Fertil Dev*. 1996. 8(4).
343. Frankenberg S, Shaw G, Freyer C, Pask AJ, Renfree MB. Early cell lineage specification in a marsupial: A case for diverse mechanisms among mammals. *Development*. 2013. 140(5).
344. Bourque G, Burns KH, Gehring M, et al. Ten things you should know about transposable elements. *Genome Biol*. 2018. 19(1).
345. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. *Nature*. 1980. 284(5757).
346. Orgel LE, Crick FHC. Selfish DNA: The ultimate parasite. *Nature*. 1980. 284(5757).
347. Hurst GDD, Werren JH. The role of selfish genetic elements in eukaryotic evolution. *Nat Rev Genet*. 2001. 2(8).
348. Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002. 420(6915).
349. Boeke JD, Garfinkel DJ, Styles CA, Fink GR. Ty elements transpose through an RNA intermediate. *Cell*. 1985. 40(3).
350. Ray DA, Feschotte C, Pagan HJT, et al. Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. *Genome Res*. 2008. 18(5).
351. Ray DA, Pagan HJT, Thompson ML, Stevens RD. Bats with hATs: Evidence for recent DNA transposon activity in genus Myotis. *Mol Biol Evol*. 2007. 24(3).
352. Mitra R, Li X, Kapusta A, et al. Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. *Proc Natl Acad Sci U S A*. 2013. 110(1).
353. Piégu B, Bire S, Arensburger P, Bigot Y. A survey of transposable element classification systems - A call for a fundamental update to meet the challenge of their diversity and complexity. *Mol Phylogenet Evol*. 2015. 86.
354. Brown PO, Bowerman B, Varmus HE, Bishop JM. Correct integration of retroviral DNA in vitro. *Cell*. 1987. 49(3).
355. Luan DD, Korman MH, Jakubczak JL, Eickbush TH. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. *Cell*. 1993. 72(4).
356. Mager DL, Stoye JP. Mammalian endogenous retroviruses. *Microbiol Spectr*.

2015. 3(1).

357. Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. *Microbiol Spectr*. 2015. 3(2).
358. Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proc Natl Acad Sci U S A*. 2016. 113(16).
359. Mills RE, Bennett EA, Iskow RC, Devine SE. Which transposable elements are active in the human genome? *Trends Genet*. 2007. 23(4).
360. Maksakova IA, Romanish MT, Gagnier L, Dunn CA, Van De Lagemaat LN, Mager DL. Retroviral elements and their hosts: Insertional mutagenesis in the mouse germ line. *PLoS Genet*. 2006. 2(1).
361. Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet*. 2003. 35(1).
362. Raiz J, Damert A, Chira S, et al. The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res*. 2012. 40(4).
363. Dewannieux M, Heidmann T, Yaniv M. L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells. *J Mol Biol*. 2005. 349(2).
364. Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum Mol Genet*. 2011. 20(17).
365. Evgen'ev MB, Arkhipova IR. Penelope-like elements - A new class of retroelements: Distribution, function and possible evolutionary significance. *Cytogenet Genome Res*. 2005. 110(1-4).
366. Beck CR, Collier P, Macfarlane C, et al. LINE-1 retrotransposition activity in human genomes. *Cell*. 2010. 141(7).
367. Brouha B, Schustak J, Badge RM, et al. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A*. 2003. 100(9).
368. Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA. SVA elements: A hominid-specific retroposon family. *J Mol Biol*. 2005. 354(4).
369. Hancks DC, Kazazian HH. SVA retrotransposons: Evolution and genetic instability. *Semin Cancer Biol*. 2010. 20(4).
370. Batzer MA, Deininger PL. Alu repeats and human genomic diversity. *Nat Rev Genet*. 2002. 3(5).
371. Bennett EA, Keller H, Mills RE, et al. Active Alu retrotransposons in the human

genome. *Genome Res*. 2008. 18(12).

372. Chénais B. Transposable Elements and Human Diseases: Mechanisms and Implication in the Response to Environmental Pollutants. *Int J Mol Sci*. 2022. 23(5).
373. Kazazian HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*. 1988. 332(6160).
374. Hancks DC, Kazazian HH. Roles for retrotransposon insertions in human disease. *Mob DNA*. 2016. 7:9
375. Goodier JL. Restricting retrotransposons: A review. *Mob DNA*. 2016. 7:16.
376. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: From conflicts to benefits. *Nat Rev Genet*. 2017. 18(2).
377. Lu X, Sachs F, Ramsay LA, et al. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat Struct Mol Biol*. 2014. 21(4).
378. Percharde M, Lin CJ, Yin Y, et al. A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell*. 2018. 174(2).
379. Gerdes P, Richardson SR, Mager DL, Faulkner GJ. Transposable elements in the mammalian embryo: Pioneers surviving through stealth and service. *Genome Biol*. 2016. 17:100.
380. Wang J, Xie G, Singh M, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*. 2014. 516(7531).
381. Loewer S, Cabili MN, Guttman M, et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet*. 2010. 42(12).
382. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol*. 2012. 13(11).
383. Piriyapongsa J, Jordan IK. A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS One*. 2007. 2(2).
384. Piriyapongsa J, Mariño-Ramírez L, Jordan IK. Origin and evolution of human microRNAs from transposable elements. *Genetics*. 2007. 176(2).
385. Qin S, Jin P, Zhou X, Chen L, Ma F. The role of transposable elements in the origin and evolution of microRNAs in human. *PLoS One*. 2015. 10(6).
386. Dupressoir A, Lavialle C, Heidmann T. From ancestral infectious retroviruses to bona fide cellular genes: Role of the captured syncytins in placentation. *Placenta*. 2012. 33(9).
387. Cornelis G, Vernochet C, Carradec Q, et al. Retroviral envelope gene captures

and syncytin exaptation for placentation in marsupials. *Proc Natl Acad Sci U S A*. 2015. 112(5).

388. Fadloun A, Le Gras S, Jost B, et al. Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. *Nat Struct Mol Biol*. 2013. 20(3).
389. Liu L, Leng L, Liu C, et al. An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. *Nat Commun*. 2019. 10(1).
390. Ishiuchi T, Enriquez-Gasca R, Mizutani E, et al. Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. *Nat Struct Mol Biol*. 2015. 22(9).
391. Jachowicz JW, Bing X, Pontabry J, Bošković A, Rando OJ, Torres-Padilla ME. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat Genet*. 2017. 49(10).
392. Kohlrausch FB, Berteli TS, Wang F, Navarro PA, Keefe DL. Control of LINE-1 Expression Maintains Genome Integrity in Germline and Early Embryo Development. *Reprod Sci*. 2022. 29(2).
393. Beraldi R, Pittoggi C, Sciamanna I, Mattei E, Spadafora C. Expression of LINE-1 retroposons is essential for murine preimplantation development. *Mol Reprod Dev*. 2006. 73(3).
394. Ohnuki M, Tanabe K, Sutou K, et al. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proc Natl Acad Sci U S A*. 2014. 111(34).
395. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009. 10(1).
396. Velculescu VE, Zhang L, Zhou W, et al. Characterization of the yeast transcriptome. *Cell*. 1997. 88(2).
397. Piétu G, Mariage-Samson R, Fayein NA, et al. The genexpress IMAGE knowledge base of the human brain transcriptome: A prototype integrated resource for functional and computational genomics. *Genome Res*. 1999. 9(2).
398. Clark TA, Sugnet CW, Ares M. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science.* 2002. 296(5569).
399. Wang B, Kumar V, Olson A, Ware D. Reviving the transcriptome studies: An insight into the emergence of single-molecule transcriptome sequencing. *Front Genet*. 2019. 10:384.
400. Emrich SJ, Barbazuk WB, Li L, Schnable PS. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res*. 2007. 17(1).

401. Lister R, O'Malley RC, Tonti-Filippini J, et al. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*. 2008. 133(3).
402. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced applications of RNA sequencing and challenges. *Bioinform Biol Insights*. 2015. 9.
403. Sheng Q, Vickers K, Zhao S, et al. Multi-perspective quality control of Illumina RNA sequencing data analysis. *Brief Funct Genomics*. 2017. 16(4).
404. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010.
405. Liao Y, Shi W. Read trimming is not required for mapping and quantification of RNA-seq reads at the gene level. *NAR Genom Bioinform*. 2020. 2(3).
406. Williams CR, Baccarella A, Parrish JZ, Kim CC. Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics*. 2016. 17(1)
407. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016. 34(5).
408. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017. 14(4).
409. Srivastava A, Malik L, Sarkar H, et al. Alignment and mapping methodology influence transcript abundance estimation. *Genome Biol*. 2020. 21(1).
410. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011. 29(7).
411. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014. 15(12).
412. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2009. 26(1).
413. Wang B, Tseng E, Regulski M, et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat Commun*. 2016. 7:11708.
414. Garalde DR, Snell EA, Jachimowicz D, et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods*. 2018. 15(3).
415. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nat Rev Genet*. 2019. 20(11).
416. Fu GK, Hu J, Wang PH, Fodor SPA. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc Natl Acad Sci U S A*. 2011. 108(22).
417. Kivioja T, Vähärautio A, Karlsson K, et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods*. 2012. 9(1).

418. Moor AE, Itzkovitz S. Spatial transcriptomics: paving the way for tissue-level systems biology. *Curr Opin Biotechnol*. 2017. 46.
419. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. *Front Genet*. 2019. 10:317.
420. Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*. 2015. 161(5).
421. Zheng GXY, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017. 8:14049.
422. 10x Genomics. LIT015_Transcriptional Profiling of 1.3 Million Brain Cells with the Chromium™ Single Cell 3' Solution. 2017.
423. Regev A, Teichmann S, Lander E, et al. Science Forum: The Human Cell Atlas. *Elife*. 2017. 6:e27041.
424. Duboule D. *Guidebook to the Homeobox Genes.* Oxford: Sambrook & Tooze Publication at Oxford University Press. 1994.
425. Mallo M, Wellik DM, Deschamps J. Hox genes and regional patterning of the vertebrate body plan. *Dev Biol*. 2010. 344(1).
426. Wellik DM. Hox patterning of the vertebrate axial skeleton. *Dev Dyn*. 2007. 236(9).
427. Duboule D. The rise and fall of Hox gene clusters. *Development*. 2007. 134(14).
428. Maeda RK, Karch F. The Bithorax Complex of Drosophila; An Exceptional Hox Cluster. *Curr Top Dev Biol*. 2009. 88.
429. Gruss P, Walther C. Pax in development. *Cell*. 1992. 69(5).
430. Dahl E, Koseki H, Balling R. Pax genes and organogenesis. *BioEssays*. 1997. 19(9).
431. Blake JA, Ziman MR. Pax genes: regulators of lineage specification and progenitor cell maintenance. *Development*. 2014. 141(4).
432. Herr W, Sturm RA, Clerc RG, et al. The POU domain: a large conserved region in the mammalian pit-1, oct-1, oct-2, and Caenorhabditis elegans unc-86 gene products. *Genes Dev*. 1988. 2(12A).
433. Phillips K, Luisi B. The virtuoso of versatility: POU proteins that flex to fit. *J Mol Biol*. 2000. 302(5).
434. Hobert O, Westphal H. Functions of LIM-homeobox genes. *Trends Genet*. 2000. 16(2).
435. Sheng HZ, Zhadanov AB, Mosinger B, et al. Specification of pituitary cell lineages by the LIM homeobox gene Lhx3. *Science.* 1996. 272(5264).
436. Costello I, Nowotschin S, Sun X, et al. Lhx1 functions together with Otx2, Foxa2, and Ldb1 to govern anterior mesendoderm, node, and midline development.

Genes Dev. 2015. 29(20).

437. Mukherjee K, Bürglin TR. Comprehensive analysis of animal TALE homeobox genes: New conserved motifs and cases of accelerated evolution. *J Mol Evol*. 2007. 65(2).
438. Bürglin TR, Cassata G. Loss and gain of domains during evolution of cut superclass homeobox genes. *Int J Dev Biol*. 2002. 46(1).
439. Paps J, Xu F, Zhang G, Holland PWH. Reinforcing the egg-timer: Recruitment of novel Lophotrochozoa homeobox genes to early and late development in the Pacific oyster. *Genome Biol Evol*. 2015. 7(3).
440. Booth HAF, Holland PWH. Eleven daughters of NANOG. *Genomics*. 2004. 84(2).
441. Jiang Z, Sun J, Dong H, et al. Transcriptional profiles of bovine in vivo pre-implantation development. *BMC Genomics*. 2014. 15(1).
442. Graf A, Krebs S, Zakhartchenko V, Schwalb B, Blum H, Wolf E. Fine mapping of genome activation in bovine embryos by RNA sequencing. *Proc Natl Acad Sci U S A*. 2014. 111(11).
443. Liu X, Wang Y, Gao Y, et al. H3K9 demethylase KDM4E is an epigenetic regulator for bovine embryonic development and a defective factor for nuclear reprogramming. *Development*. 2018. 145(4).
444. Bernardo AS, Jouneau A, Marks H, et al. Mammalian embryo comparison identifies novel pluripotency genes associated with the naïve or primed state. *Biol Open*. 2018. 7(8).
445. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013. 29(1).
446. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015. 33(3).
447. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990. 215(3).
448. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res*. 1997. 25(17).
449. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011. 29(1).
450. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comput Biol*. 2000. 7(1-2).
451. Ronquist F, Teslenko M, van der Mark P, et al. MrBayes 3.2: Efficient Bayesian

Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst Biol*. 2012. 61(3).

452. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*. 2001. 17(8).
453. Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res*. 2000. 28(21).
454. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011. 7:539
455. Gouy M, Guindon S, Gascuel O. Sea view version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol*. 2010. 27(2).
456. Letunic I, Bork P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res*. 2019. 47(W1).
457. Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol*. 2013. 30(8).
458. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: A computational tool for the study of gene family evolution. *Bioinformatics*. 2006. 22(10).
459. Durand D, Halldórsson BV, Vernot B. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol.* 2006. 13(2).
460. Chen K, Durand D, Farach-Colton M. NOTUNG: A program for dating gene duplications and optimizing gene family trees. *J Comput Biol.* 2000. 7(3-4).
461. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol*. 2017. 34(7).
462. Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res*. 2005. 15(8).
463. Hahn MW, Demuth JP, Han SG. Accelerated rate of gene gain and loss in primates. *Genetics*. 2007. 177(3).
464. Mansai SP, Innan H. The power of the methods for detecting interlocus gene conversion. *Genetics*. 2010. 184(2).
465. Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. R package version 2.57.0.
466. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2021.

467. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 1970. 48(3).
468. Warnes GR, Bolker B, Bonebakker L, et al. Package “gplots”: Various R programming tools for plotting data. R Package version 3.0.3. 2020.
469. Brown SDJ, Collins RA, Boyer S, et al. Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Mol Ecol Resour*. 2012. 12(3).
470. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*. 1980. 16(2).
471. Kosakovsky Pond SL, Frost SDW, Muse SV. HyPhy: Hypothesis testing using phylogenies. *Bioinformatics*. 2005. 21(5).
472. Kosakovsky Pond SL, Poon AFY, Velazquez R, et al. HyPhy 2.5 - A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol Biol Evol*. 2020. 37(1).
473. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. GARD: A genetic algorithm for recombination detection. *Bioinformatics*. 2006. 22(24).
474. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol*. 2006. 23(10).
475. Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. Datamonkey 2.0: A modern web application for characterizing selective and other evolutionary processes. *Mol Biol Evol*. 2018. 35(3).
476. Hurvich CM, Tsai CL. Regression and time series model selection in small samples. *Biometrika*. 1989. 76(2).
477. Akaike H. A New Look at the Statistical Model Identification. *IEEE Trans Automat Contr*. 1974. 19.
478. Sugiura N. Further Analysis of the Data by Akaike’s Information Criterion and the Finite Corrections. *Commun Stat - Theory Methods*. 1978. 7(1).
479. Ritz C, Spiess AN. qpcR: An R package for sigmoidal model selection in quantitative real-time polymerase chain reaction analysis. *Bioinformatics*. 2008. 24(13).
480. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer. 2002.
481. Wagenmakers EJ, Farrell S. AIC model selection using Akaike weights. *Psychon Bull Rev*. 2004. 11(1).
482. Sand A, Holt MK, Johansen J, Brodal GS, Mailund T, Pedersen CNS. TqDist: A

library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics*. 2014. 30(14).

483. Smith MR. Quartet: Comparison of Phylogenetic Trees Using Quartet and Bipartition Measures. *Comprehensive R Archive Network*. 2019.
484. Estabrook GF, McMorris FR, Meacham CA. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Syst Zool*. 1985. 34(2).
485. Smith MR. Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. *Biol Lett*. 2019. 15(2).
486. Sawyer S. Statistical tests for detecting gene conversion. *Mol Biol Evol*. 1989. 6(5).
487. Tajima F. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics*. 1993. 135(2).
488. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001.
489. Murrell B, Weaver S, Smith MD, et al. Gene-wide identification of episodic selection. *Mol Biol Evol*. 2015. 32(5).
490. Suyama M, Torrents D, Bork P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 2006. 34(Web Server Issue).
491. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007. 24(8).
492. Yang Z. Paml: A program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*. 1997. 13(5).
493. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet*. 2012. 8(7).
494. Lu A, Guindon S. Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. *Mol Biol Evol*. 2014. 31(2).
495. Miyazono KI, Zhi Y, Takamura Y, et al. Cooperative DNA-binding and sequence-recognition mechanism of aristaless and clawless. *EMBO J*. 2010. 29(9).
496. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000. 28(1).
497. Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. 1993. 234(3).
498. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera - A visualization

system for exploratory research and analysis. *J Comput Chem*. 2004. 25(13).

499. Wertheim JO, Murrell B, Smith MD, Pond SLK, Scheffler K. RELAX: Detecting relaxed selection in a phylogenetic framework. *Mol Biol Evol*. 2015. 32(3).
500. Shriner D, Nickle DC, Jensen MA, Mullins JI. Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet Res*. 2003. 81(2).
501. Anisimova M, Nielsen R, Yang Z. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*. 2003. 164(3).
502. Ekker SC, Jackson DG, Von Kassler DP, Sun BI, Young KE, Beachy PA. The degree of variation in DNA sequence recognition among four Drosophila homeotic proteins. *EMBO J*. 1994. 13(15).
503. Damante G, Pellizzari L, Esposito G, et al. A molecular code dictates sequence-specific DNA recognition by homeodomains. *EMBO J*. 1996. 15(18).
504. Furukubo-Tokunaga K, Müller M, Affolter M, Pick L, Kloter U, Gehring WJ. In vivo analysis of the helix-turn-helix motif of the fushi tarazu homeo domain of Drosophila melanogaster. *Genes Dev*. 1992. 6(6)
505. Simon KJ, Grueneberg DA, Gilman M. Protein and DNA contact surfaces that mediate the selective action of the Phox1 homeodomain at the c-fos serum response element. *Mol Cell Biol*. 1997. 17(11).
506. Wilson DS, Guenther B, Desplan C, Kuriyan J. High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. *Cell*. 1995. 82(5).
507. Plaza S, Prince F, Adachi Y, Punzo C, Cribbs DL, Gehring WJ. Cross-regulatory protein-protein interactions between Hox and Pax transcription factors. *Proc Natl Acad Sci U S A*. 2008. 105(36).
508. Altamirano-Torres C, Salinas-Hernández JE, Cárdenas-Chávez DL, Rodríguez-Padilla C, Reséndez-Pérez D. Transcription factor TFIIEβ interacts with two exposed positions in helix 2 of the Antennapedia homeodomain to control homeotic function in Drosophila. *PLoS One*. 2018. 13(10).
509. Zaffran S, Frasch M. The homeodomain of Tinman mediates homo- and heterodimerization of NK proteins. *Biochem Biophys Res Commun*. 2005. 334(2).
510. Zhang J. Genetic Redundancies and Their Evolutionary Maintenance. *Adv Exp Med Biol*. 2012. 751.
511. McNulty CL, Peres JN, Bardine N, van den Akker WMR, Durston AJ. Knockdown of the complete Hox paralogous group 1 leads to dramatic hindbrain

and neural crest defects. *Development*. 2005. 132(12).

512. Tvrdik P, Capecchi MR. Reversal of Hox1 Gene Subfunctionalization in the Mouse. *Dev Cell*. 2006. 11(2).
513. Ruff JS, Saffarini RB, Ramoz LL, et al. Fitness assays reveal incomplete functional redundancy of the hoxa1 and hoxb1 paralogs of mice. *Genetics*. 2015. 201(2).
514. Nowak MA, Boerlijst MC, Cooke J, Smith JM. Evolution of genetic redundancy. *Nature*. 1997. 388(6638).
515. Vavouri T, Semple JI, Lehner B. Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. *Trends Genet*. 2008. 24(10).
516. Castresana J. Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. *Nucleic Acids Res*. 2002. 30(8).
517. Lercher MJ, Blumenthal T, Hurst LD. Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. *Genome Res*. 2003. 13(2).
518. Zimmer EA, Martin SL, Beverley SM, Kan YW, Wilson AC. Rapid duplication and loss of genes coding for the α chains of hemoglobin. *Proc Natl Acad Sci U S A*. 1980. 77(4).
519. Arnheim N. Concerted evolution of multigene families. In: Nei M, Koehn RK, eds. *Evolution of Genes and Proteins*. Sunderland, MA: Sinauer. 1983. pp38-61.
520. Korunes KL, Noor MAF. Gene conversion and linkage: effects on genome evolution and speciation. *Mol Ecol*. 2017. 26(1).
521. Innan H. A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. *Proc Natl Acad Sci U S A*. 2003. 100(15).
522. Winderickx J, Battisti L, Hibiya Y, Motulsky AG, Deeb SS. Haplotype diversity in the human red and green opsin genes: Evidence for frequent sequence exchange in exon 3. *Hum Mol Genet*. 1993. 2(9).
523. Mano S, Innan H. The evolutionary rate of duplicated genes under concerted evolution. *Genetics*. 2008. 180(1).
524. Mori Y, Sakuraoka M, Suzuki T, et al. Exogenous TPRX1 homeoprotein modulates the gene expression of lineage-specific transcription factors in human embryonal carcinoma cells. *Biotechnol Biotechnol Equip*. 2018. 32(3).
525. Lewin TD, Royall AH, Holland PWH. Dynamic Molecular Evolution of Mammalian Homeobox Genes: Duplication, Loss, Divergence and Gene Conversion Sculpt PRD-Class Repertoires. *J Mol Evol*. 2021. 89(6).
526. Ritchie H, Roser M. Meat and Dairy Production *Our World in Data*. 2017.

527. FAO. Production: Crops and Livestock Products. *FAOSTAT*. 2019. Available online at: https://www.fao.org/faostat/en/#data [Date accessed: 17/12/2021].
528. FAO. Global agriculture towards 2050. *High Level Expert Forum - How to Feed the World in 2050.* Rome, Italy. 2009. Available online at: https://www.fao.org/fileadmin/templates/wsfs/docs/Issues_papers/HLEF2050_Global_Agriculture.pdf [Date accessed: 17/12/2021]
529. Delgado CL. Rising Consumption of Meat and Milk in Developing Countries Has Created a New Food Revolution. *J Nutr*. 2003. 133(11 Suppl 2).
530. Lucy MC. Reproductive Loss in High-Producing Dairy Cattle: Where Will It End? *J Dairy Sci*. 2001. 84(6).
531. Oltenacu PA, Algers B. Selection for increased production and the welfare of dairy cows: Are new breeding goals needed? *Ambio*. 2005. 34(4-5).
532. Berry DP, Wall E, Pryce JE. Genetics and genomics of reproductive performance in dairy and beef cattle. *Animal*. 2014. 8(SUPPL. 1).
533. Sheldon IM, Dobson H. Reproductive challenges facing the cattle industry at the beginning of the 21st century. *Reprod Suppl*. 2003. 61.
534. Ahlman T, Berglund B, Rydhmer L, Strandberg E. Culling reasons in organic and conventional dairy herds and genotype by environment interaction for longevity. *J Dairy Sci*. 2011. 94(3).
535. Ansari-Lari M, Mohebbi-Fani M, Rowshan-Ghasrodashti A. Causes of culling in dairy cows and its relation to age at culling and interval from calving in Shiraz, Southern Iran. *Vet Res Forum.* 2012. 3(4).
536. Rilanto T, Reimus K, Orro T, Emanuelson U, Viltrop A, Mõtus K. Culling reasons and risk factors in Estonian dairy cows. *BMC Vet Res*. 2020. 16(1).
537. Chiumia D, Chagunda MGG, MacRae AI, Roberts DJ. Predisposing factors for involuntary culling in Holstein-Friesian dairy cows. *J Dairy Res*. 2013. 80(1).
538. Diskin MG, Morris DG. Embryonic and Early Foetal Losses in Cattle and Other Ruminants. *Reprod Domest Anim*. 2008. 43(Suppl 2).
539. Perkel KJ, Tscherner A, Merrill C, Lamarre J, Madan P. The ART of selecting the best embryo: A review of early embryonic mortality and bovine embryo viability assessment methods. *Mol Reprod Dev*. 2015. 82(11).
540. Galan A, Diaz-Gimeno P, Poo ME, et al. Defining the Genomic Signature of Totipotency and Pluripotency during Early Human Development. *PLoS One*. 2013. 8(4).
541. Chen N, Fu W, Zhao J, et al. BGVD: An Integrated Database for Bovine Sequencing Variations and Selective Signatures. *Genomics Proteomics Bioinformatics*. 2020. 18(2).

542. Watson AJ, Barcroft LC. Regulation of blastocyst formation. *Front Biosci*. 2001. 6(1).
543. Hasley A, Chavez S, Danilchik M, Wühr M, Pelegri F. Vertebrate embryonic cleavage pattern determination. *Adv Exp Med Biol.* 2017. 953.
544. Damsky C, Sutherland A, Fisher S. Extracellular matrix 5: Adhesive interactions in early mammalian embryogenesis, implantation, and placentation. *FASEB J*. 1993. 7(14).
545. Lehtonen E, Ordónez G, Reima I. Cytoskeleton in preimplantation mouse development. *Cell Differ*. 1988. 24(3).
546. Albertini DF, Overstrom EW, Ebert KM. Changes in the organization of the actin cytoskeleton during preimplantation development of the pig embryo. *Biol Reprod*. 1987. 37(2).
547. Lim HYG, Plachta N. Cytoskeletal control of early mammalian development. *Nat Rev Mol Cell Biol*. 2021. 22(8).
548. Enders AC, Lantz KC, Schlafke S. The Morula-Blastocyst Transition in Two Old World Primates: The Baboon and Rhesus Monkey. *J Med Primatol*. 1990. 19(8).
549. Koyama H, Suzuki H, Yang X, Jiang S, Foote RH. Analysis of polarity of bovine and rabbit embryos by scanning electron microscopy. *Biol Reprod*. 1994. 50(1).
550. Vestweber D, Gossler A, Boller K, Kemler R. Expression and distribution of cell adhesion molecule uvomorulin in mouse preimplantation embryos. *Dev Biol*. 1987. 124(2).
551. Eckert JJ, Fleming TP. Tight junction biogenesis during early development. *Biochim Biophys Acta*. 2008. 1778(3).
552. Hardy K, Warner A, Winston RML, Becker DL. Expression of intercellular junctions during preimplantation development of the human embryo. *Mol Hum Reprod*. 1996. 2(8).
553. Fleming TP. Cell adhesion in the preimplantation mammalian embryo and its role in trophectoderm differentiation and blastocyst morphogenesis. *Front Biosci*. 2001. 6(1).
554. Fleming TP, Ghassemifar MR, Sheth B. Junctional complexes in the early mammalian embryo. *Semin Reprod Med*. 2000. 18(2).
555. Aberkane A, Essahib W, Spits C, et al. Expression of adhesion and extracellular matrix genes in human blastocysts upon attachment in a 2D co-culture system. *Mol Hum Reprod*. 2018. 24(7).
556. He B, Zhang H, Wang J, et al. Blastocyst activation engenders transcriptome reprogram affecting X-chromosome reactivation and inflammatory trigger of implantation. *Proc Natl Acad Sci U S A*. 2019. 116(33).

557. Renfree MB, Shaw G. Embryo-endometrial interactions during early development after embryonic diapause in the marsupial tammar wallaby. *Int J Dev Biol*. 2014. 58(2-4).
558. Renfree MB, Lewis AMD. Cleavage in vivo and in vitro in the marsupial Macropus eugenii. *Reprod Fertil Dev*. 1996. 8(4).
559. Hu K. On Mammalian Totipotency: What Is the Molecular Underpinning for the Totipotency of Zygote? *Stem Cells Dev*. 2019. 28(14).
560. Indraccolo S, Pfeffer U, Minuzzo S, et al. Identification of Genes Selectively Regulated by IFNs in Endothelial Cells. *J Immunol*. 2007. 178(2).
561. Kuepper MK, Bütow M, Herrmann O, et al. Stem cell persistence in CML is mediated by extrinsically activated JAK1-STAT3 signaling. *Leukemia*. 2019. 33(8).
562. Tirumurugaan KG, Pawar RM, Raj GD, Thangavelu A, Hammond JA, Parida S. RNAseq reveals the contribution of interferon stimulated genes to the increased host defense and decreased PPR viral replication in cattle. *Viruses*. 2020. 12(4).
563. Vladimer GI, Górna MW, Superti-Furga G. IFITs: Emerging roles as key anti-viral proteins. *Front Immunol*. 2014. 5:94.
564. DeFilippis VR, Alvarado D, Sali T, Rothenburg S, Früh K. Human Cytomegalovirus Induces the Interferon Response via the DNA Sensor ZBP1. *J Virol*. 2010. 84(1).
565. Hernandez JM, Elahi A, Clark W, et al. The Tumor Suppressive Effects of HPP1 Are Mediated Through JAK-STAT-Interferon Signaling Pathways. *DNA Cell Biol*. 2015. 34(8).
566. Harwardt T, Lukas S, Zenger M, et al. Human Cytomegalovirus Immediate-Early 1 Protein Rewires Upstream STAT3 to Downstream STAT1 Signaling Switching an IL6-Type to an IFNγ-Like Response. *PLoS Pathog*. 2016. 12(7).
567. Hu Y, Wang J, Yang B, et al. Guanylate Binding Protein 4 Negatively Regulates Virus-Induced Type I IFN and Antiviral Response by Targeting IFN Regulatory Factor 7. *J Immunol*. 2011. 187(12).
568. Schneider WM, Chevillotte MD, Rice CM. Interferon-stimulated genes: A complex web of host defenses. *Annu Rev Immunol*. 2014. 32.
569. DeDiego ML, Nogales A, Martinez-Sobrido L, Topham DJ. Interferon-induced protein 44 interacts with cellular fk506-binding protein 5, negatively regulates host antiviral responses, and supports virus replication. *mBio*. 2019. 10(4).
570. Seo JY, Yaneva R, Cresswell P. Viperin: A multifunctional, interferon-inducible protein that regulates virus replication. *Cell Host Microbe*. 2011. 10(6).
571. Kane M, Yadav SS, Bitzegeio J, et al. MX2 is an interferon-induced inhibitor of

HIV-1 infection. *Nature*. 2013. 502(7472).

572. Sun Y, Qiao L, Xia HHX, et al. Regulation of XAF1 expression in human colon cancer cell by interferon β: Activation by the transcription regulator STAT1. *Cancer Lett*. 2008. 260(1-2).
573. Zhou X, Michal JJ, Zhang L, et al. Interferon induced IFIT family genes in host antiviral defense. *Int J Biol Sci*. 2013. 9(2).
574. Haan C, Kreis S, Margue C, Behrmann I. Jaks and cytokine receptors-An intimate relationship. *Biochem Pharmacol*. 2006. 72(11).
575. Bousoik E, Montazeri Aliabadi H. "Do We Know Jack" About JAK? A Closer Look at JAK/STAT Signaling Pathway. *Front Oncol*. 2018. 8:287.
576. Krebs DL, Hilton DJ. SOCS Proteins: Negative Regulators of Cytokine Signaling. *Stem Cells*. 2001. 19(5).
577. Matsumoto A, Masuhara M, Mitsui K, et al. CIS, a cytokine inducible SH2 protein, is a target of the JAK-STAT5 pathway and modulates STAT5 activation. *Blood*. 1997. 89(9).
578. Stirparo GG, Kurowski A, Yanagida A, et al. OCT4 induces embryonic pluripotency via STAT3 signaling and metabolic mechanisms. *Proc Natl Acad Sci U S A*. 2021. 118(3).
579. Yang J, Van Oosten AL, Theunissen TW, Guo G, Silva JCR, Smith A. Stat3 activation is limiting for reprogramming to ground state pluripotency. *Cell Stem Cell*. 2010. 7(3).
580. Van Oosten AL, Costa Y, Smith A, Silva JCR. JAK/STAT3 signalling is sufficient and dominant over antagonistic cues for the establishment of naive pluripotency. *Nat Commun*. 2012. 3:817.
581. Meng F, Forrester-Gauntlett B, Turner P, Henderson H, Oback B. Signal inhibition reveals JAK/STAT3 pathway as critical for bovine inner cell mass development. *Biol Reprod*. 2015. 93(6).
582. Niwa H, Burdon T, Chambers I, Smith A. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev*. 1998. 12(13).
583. Nichols J, Chambers I, Taga T, Smith A. Physiological rationale for responsiveness of mouse embryonic stem cells to gp130 cytokines. *Development*. 2001. 128(12).
584. Do DV, Ueda J, Messerschmidt DM, et al. A genetic and developmental pathway from STAT3 to the OCT4-NANOG circuit is essential for maintenance of ICM lineages in vivo. *Genes Dev*. 2013. 27(12).
585. Vangipuram M, Ting D, Kim S, Diaz R, Schüle B. Skin punch biopsy explant culture for derivation of primary human fibroblasts. *J Vis Exp*. 2013. (77).

586. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016. 32(19).
587. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014. 30(15).
588. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Research*. 2016. 4:1521.
589. Blighe K, Rana S, Lewis M. EnhancedVolcano version 1.16.0: Publication-ready volcano plots with enhanced colouring and labeling. *R Package*. 2022.
590. Hulsen T, de Vlieg J, Alkema W. BioVenn - A web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics*. 2008. 9:488.
591. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. *Nat Genet*. 2000. 25(1).
592. Thomas PD, Campbell MJ, Kejariwal A, et al. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res*. 2003. 13(9).
593. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res*. 2013. 41(D1).
594. Mi H, Ebert D, Muruganujan A, et al. PANTHER version 16: A revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res*. 2021. 49(D1).
595. Altenhoff AM, Train C-M, Gilbert KJ, et al. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res*. 2021. 49(D1).
596. Kumar L, Futschik ME. Mfuzz: A software package for soft clustering of microarray data. *Bioinformation*. 2007. 2(1).
597. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B*. 1995. 57(1).
598. Lewin TD, Fouladi-Nashta AA, Holland PWH. PRD-Class Homeobox Genes in Bovine Early Embryos: Function, Evolution, and Overlapping Roles. *Mol Biol Evol*. 2022. 39(5).
599. Wilkinson RD, Steiper ME, Soligo C, Martin RD, Yang Z, Tavaré S. Dating Primate Divergences through an Integrated Analysis of Palaeontological and Molecular Data. *Syst Biol*. 2011. 60(1).
600. Pozzi L, Hodgson JA, Burrell AS, Sterner KN, Raaum RL, Disotell TR. Primate

phylogenetic relationships and divergence dates inferred from complete mitochondrial genomes. *Mol Phylogenet Evol*. 2014. 75.

601. dos Reis M, Gunnell GF, Barba-Montoya J, Wilkins A, Yang Z, Yoder AD. Using Phylogenomic Data to Explore the Effects of Relaxed Clocks and Calibration Strategies on Divergence Time Estimation: Primates as a Test Case. *Syst Biol*. 2018. 67(4).
602. Clarke ND, Kissinger CR, Desjarlais J, Gilliland GL, Pabo CO. Structural studies of the engrailed homeodomain. *Protein Sci*. 1994. 3(10).
603. Piskacek S, Gregor M, Nemethova M, Grabner M, Kovarik P, Piskacek M. Nine-amino-acid transactivation domain: Establishment and prediction utilities. *Genomics*. 2007. 89(6).
604. Wang J-R, Huang W-L, Tsai M-J, Hsu K-T, Huang H-L, Ho S-Y. ESA-UbiSite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives. *Bioinformatics*. 2017. 33(5).
605. Yan L, Yang M, Guo H, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol*. 2013. 20(9).
606. Dunwell TL, Paps J, Holland PWH. Novel and divergent genes in the evolution of placental mammals. *Proc R Soc B Biol Sci*. 2017. 284(1864).
607. Bosshard HR, Marti DN, Jelesarov I. Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings. *J Mol Recognit*. 2004. 17(1).
608. Donald JE, Kulp DW, DeGrado WF. Salt bridges: geometrically specific, designable interactions. *Proteins Struct Funct Bioinforma*. 2011. 79(3).
609. MacLean JA, Chen MA, Wayne CM, et al. Rhox: A new homeobox gene cluster. *Cell*. 2005. 120(3).
610. Nakamura T, Okamoto I, Sasaki K, et al. A developmental coordinate of pluripotency among mice, monkeys and humans. *Nature*. 2016. 537(7618).
611. Wang X, Liu D, He D, et al. Transcriptome analyses of rhesus monkey preimplantation embryos reveal a reduced capacity for DNA double-strand break repair in primate oocytes and early embryos. *Genome Res*. 2017. 27(4).
612. Hu Y, Huang K, Zeng Q, et al. Single-cell analysis of nonhuman primate preimplantation development in comparison to humans and mice. *Dev Dyn*. 2021. 250(7).
613. Carter AM, Enders AC. Comparative aspects of trophoblast development and placentation. *Reprod Biol Endocrinol*. 2004. 2(1).
614. Siriwardena D, Boroviak TE. Evolutionary divergence of embryo implantation in

primates. *Philos Trans R Soc B Biol Sci*. 2022. 377(1865).

615. Hurles M. Gene duplication: the genomic trade in spare parts. *PLoS Biol*. 2004. 2(7).
616. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol*. 2015. 32(1).
617. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol*. 2018. 35(2).
618. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017. 14(6).
619. Kumar S, Suleski M, Craig JM, et al. TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol Biol Evol*. 2022. 39(8).
620. Zhong YF, Butts T, Holland PWH. HomeoDB: A database of homeobox gene diversity. *Evol Dev*. 2008. 10(5).
621. Zhong YF, Holland PWH. HomeoDB2: Functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evol Dev*. 2011. 13(6).
622. Kosakovsky Pond SL, Frost SDW. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. *Mol Biol Evol*. 2005. 22(5).
623. Chitwood JL, Burruel VR, Halstead MM, Meyers SA, Ross PJ. Transcriptome profiling of individual rhesus macaque oocytes and preimplantation embryos. *Biol Reprod*. 2017. 97(3).
624. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. 2016. 32(18).
625. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*. 2019. 35(12).
626. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L-P, Mi H. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci*. 2022. 31(1).
627. Kopera HC, Larson PA, Moldovan JB, Richardson SR, Liu Y, Moran JV. LINE-1 cultured cell retrotransposition assay. *Methods Mol Biol*. 2016. 1400.
628. Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, et al. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. *Science*. 2009. 324(5926).

629. Narita N, Nishio H, Kitoh Y, et al. Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. *J Clin Invest*. 1993. 91(5).
630. Ooi SKT, O'Donnell AH, Bestor TH. Mammalian cytosine methylation at a glance. *J Cell Sci*. 2009. 122(16).
631. Bestor TH, Bourc'his D. Transposon silencing and imprint establishment in mammalian germ cells. *Cold Spring Harb Symp Quant Biol*. 2004. 69.
632. Aravin AA, Bourc'his D. Small RNA guides for de novo DNA methylation in mammalian germ cells. *Genes Dev*. 2008. 22(8).
633. Aravin AA, Sachidanandam R, Bourc'his D, et al. A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice. *Mol Cell*. 2008. 31(6).
634. Yang N, Kazazian HH. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat Struct Mol Biol*. 2006. 13(9).
635. Jacobs FMJ, Greenberg D, Nguyen N, et al. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*. 2014. 516(7530).
636. Ecco G, Imbeault M, Trono D. KRAB zinc finger proteins. *Development*. 2017. 144(15).
637. Imbeault M, Helleboid PY, Trono D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*. 2017. 543(7646).
638. Rowe HM, Jakobsson J, Mesnard D, et al. KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature*. 2010. 463(7278).
639. Frietze S, O'Geen H, Blahnik KR, Jin VX, Farnham PJ. ZNF274 recruits the histone methyltransferase SETDB1 to the 3' ends of ZNF genes. *PLoS One*. 2010. 5(12).
640. Belancio VP, Hedges DJ, Deininger P. LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res*. 2006. 34(5).
641. Fu Y, Zhou Z, Wang H, et al. IFITM1 suppresses expression of human endogenous retroviruses in human embryonic stem cells. *FEBS Open Bio*. 2017. 7(8).
642. Hu S, Li J, Xu F, et al. SAMHD1 Inhibits LINE-1 Retrotransposition by Promoting Stress Granule Formation. *PLoS Genet*. 2015. 11(7).
643. Herrmann A, Wittmann S, Thomas D, et al. The SAMHD1-mediated block of LINE-1 retroelements is regulated by phosphorylation. *Mob DNA*. 2018. 9:11.
644. Sanz-Ramos M, Stoye JP. Capsid-binding retrovirus restriction factors:

Discovery, restriction specificity and implications for the development of novel therapeutics. *J Gen Virol*. 2013. 94(12).

645. Volkmann B, Wittmann S, Lagisquet J, et al. Human TRIM5α senses and restricts LINE-1 elements. *Proc Natl Acad Sci U S A*. 2020. 117(30).
646. Goodier JL, Cheung LE, Kazazian HH. MOV10 RNA Helicase Is a Potent Inhibitor of Retrotransposition in Cells. *PLoS Genet*. 2012. 8(10).
647. Zheng YH, Jeang KT, Tokunaga K. Host restriction factors in retroviral infection: Promises in virus-host interaction. *Retrovirology*. 2012. 9:112.
648. Arjan-Odedra S, Swanson CM, Sherer NM, Wolinsky SM, Malim MH. Endogenous MOV10 inhibits the retrotransposition of endogenous retroelements but not the replication of exogenous retroviruses. *Retrovirology*. 2012. 9:53.
649. Stenglein MD, Harris RS. APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. *J Biol Chem*. 2006. 281(25).
650. Hulme AE, Bogerd HP, Cullen BR, Moran JV. Selective inhibition of Alu retrotransposition by APOBEC3G. *Gene*. 2007. 390(1-2).
651. Stavrou S, Ross SR. APOBEC3 Proteins in Viral Immunity. *J Immunol*. 2015. 195(10).
652. Peddigari S, Li PWL, Rabe JL, Martin SL. HnRNPL and nucleolin bind LINE-1 RNA and function as host factors to modulate retrotransposition. *Nucleic Acids Res*. 2013. 41(1).
653. Perepelitsa-Belancio V, Deininger P. RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nat Genet*. 2003. 35(4).
654. Daujat S, Weiss T, Mohn F, et al. H3K64 trimethylation marks heterochromatin and is dynamically remodeled during developmental reprogramming. *Nat Struct Mol Biol*. 2009. 16(7).
655. Xu Q, Xie W. Epigenome in Early Mammalian Development: Inheritance, Reprogramming and Establishment. *Trends Cell Biol*. 2018. 28(3).
656. Platt II RN, Vandewege MW, Ray DA. Mammalian transposable elements and their impacts on genome evolution. *Chromosom Res*. 2018. 26(1-2).
657. Lanciano S, Cristofari G. Measuring and interpreting transposable element expression. *Nat Rev Genet*. 2020. 21(12).
658. Marinov GK, Wang J, Handler D, et al. Pitfalls of Mapping High-Throughput Sequencing Data to Repetitive Sequences: Piwi's Genomic Targets Still Not Identified. *Dev Cell*. 2015. 32(6).
659. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing:

computational challenges and solutions. *Nat Rev Genet*. 2012. 13(1).

660. Schwarz R, Koch P, Wilbrandt J, Hoffmann S. Locus-specific expression analysis of transposable elements. *Brief Bioinform*. 2022. 23(1).
661. Jeong HH, Yalamanchili HK, Guo C, Shulman JM, Liu Z. An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. *Pac Symp Biocomput*. 2018. 23.
662. Corsinotti A, Kapopoulou A, Gubelmann C, et al. Global and Stage Specific Patterns of Krüppel-Associated-Box Zinc Finger Protein Gene Expression in Murine Early Embryonic Cells. *PLoS One*. 2013. 8(2).
663. Petropoulos S, Panula SP, Plaza Reyes A, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell*. 2016. 165(4).
664. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018. 36(5).
665. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*. 2015. 33(5).
666. Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH. High frequency retrotransposition in cultured mammalian cells. *Cell*. 1996. 87(5).
667. Richardson SR, Narvaiza I, Planegger RA, Weitzman MD, Moran JV. APOBEC3A deaminates transiently exposed single-strand DNA during LINE-1 retrotransposition. *Elife*. 2014. 3:e2008.
668. Smit A, Hubley R, Grenn P. RepeatMasker Open-4.0. 2015.
669. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA*. 2021. 12(1).
670. Goodier JL, Pereira GC, Cheung LE, Rose RJ, Kazazian HH. The Broad-Spectrum Antiviral Protein ZAP Restricts Human Retrotransposition. *PLoS Genet*. 2015. 11(5).
671. Frassinelli L, Orecchini E, Al-Wardat S, et al. The RNA editing enzyme ADAR2 restricts L1 mobility. *RNA Biol*. 2021. 18(S1).
672. Orecchini E, Doria M, Antonioni A, et al. ADAR1 restricts LINE-1 retrotransposition. *Nucleic Acids Res*. 2017. 45(1).
673. Stetson DB, Ko JS, Heidmann T, Medzhitov R. Trex1 Prevents Cell-Intrinsic Initiation of Autoimmunity. *Cell*. 2008. 134(4).
674. Zhao K, Du J, Han X, et al. Modulation of LINE-1 and Alu/SVA Retrotransposition by Aicardi-Goutières Syndrome-Related SAMHD1. *Cell Rep*.

2013. 4(6).

675. Valente L, Nishikura K. ADAR Gene Family and A-to-I RNA Editing: Diverse Roles in Posttranscriptional Gene Regulation. *Prog Nucleic Acid Res*. 2005. 79.
676. Kimura M. On the evolutionary adjustment of spontaneous mutation rates. *Genet Res*. 1967. 9(1).
677. Nilsson M, Snoad N. Optimal mutation rates in dynamic environments. *Bull Math Biol*. 2002. 64.
678. Loh E, Salk JJ, Loeb LA. Optimization of DNA polymerase mutation rates during bacterial evolution. *Proc Natl Acad Sci U S A*. 2010. 107(3).
679. Feschotte C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet*. 2008. 9(5).
680. Bourque G, Leong B, Vega VB, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res*. 2008. 18(11).
681. Sundaram V, Cheng Y, Ma Z, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res*. 2014. 24(12).
682. Esnault C, Heidmann O, Delebecque F, et al. APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses. *Nature*. 2005. 433(7024).
683. Khatua AK, Taylor HE, Hildreth JEK, Popik W. Inhibition of LINE-1 and Alu retrotransposition by exosomes encapsidating APOBEC3G and APOBEC3F. *Virology*. 2010. 400(1).
684. Ying Y, Yang X, Zhao K, et al. The Krüppel-associated box repressor domain induces reversible and irreversible regulation of endogenous mouse genes by mediating different chromatin states. *Nucleic Acids Res*. 2015. 43(3).
685. Oleksiewicz U, Gładych M, Raman AT, et al. TRIM28 and interacting KRAB-ZNFs control self-renewal of human pluripotent stem cells through epigenetic repression of pro-differentiation genes. *Stem cell reports*. 2017. 9(6).
686. Quenneville S, Turelli P, Bojkowska K, et al. The KRAB-ZFP/KAP1 system contributes to the early embryonic establishment of site-specific DNA methylation patterns maintained during development. *Cell Rep*. 2012. 2(4).
687. Wiznerowicz M, Jakobsson J, Szulc J, et al. The Krüppel-associated box repressor domain can trigger de novo promoter methylation during mouse early embryogenesis. *J Biol Chem*. 2007. 282(47).
688. Yang WR, Ardeljan D, Pacyna CN, Payer LM, Burns KH. SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Res*. 2019. 47(5).
689. Bejerano G, Lowe CB, Ahituv N, et al. A distal enhancer and an ultraconserved

exon are derived from a novel retroposon. *Nature*. 2006. 441(7089).

690. Sasaki T, Nishihara H, Hirakawa M, et al. Possible involvement of SINEs in mammalian-specific brain formation. *Proc Natl Acad Sci U S A*. 2008. 105(11).
691. Santangelo AM, de Souza FSJ, Franchini LF, Bumaschny VF, Low MJ, Rubinstein M. Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. *PLoS Genet*. 2007. 3(10).
692. Lam DD, de Souza FSJ, Nasif S, et al. Partially redundant enhancers cooperatively maintain Mammalian pomc expression above a critical functional threshold. *PLoS Genet*. 2015. 11(2).
693. Pedersen RA, Wu K, Bałakier H. Origin of the inner cell mass in mouse embryos: Cell lineage analysis by microinjection. *Dev Biol*. 1986. 117(2).
694. Cao S, Han J, Wu J, et al. Specific gene-regulation networks during the pre-implantation development of the pig embryo as revealed by deep sequencing. *BMC Genomics*. 2014. 15(1).
695. Qian X, Jin L, Lloyd R V. In Situ Hybridization: Basic Approaches and Recent Development. *J Histotechnol*. 2004. 27(1):53-67.
696. Gall JG, Pardue ML. Formation and detection of rna-dna hybrid molecules in cytological preparations. *Proc Natl Acad Sci U S A*. 1969. 63(2).
697. Silverman AP, Kool ET. Oligonucleotide Probes for RNA-Targeted Fluorescence In Situ Hybridization. *Adv Clin Chem*. 2007. 43.
698. Lavagi I, Krebs S, Simmet K, et al. Single-cell RNA sequencing reveals developmental heterogeneity of blastomeres during major genome activation in bovine embryos. *Sci Rep*. 2018. 8(1).
699. Ramos-Ibeas P, Sang F, Zhu Q, et al. Pluripotency and X chromosome dynamics revealed in pig pre-gastrulating embryos by single cell analysis. *Nat Commun*. 2019. 10(1).
700. Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science.* 2014. 343(6167).
701. Kong Q, Yang X, Zhang H, et al. Lineage specification and pluripotency revealed by transcriptome analysis from oocyte to blastocyst in pig. *FASEB J*. 2020. 34(1).
702. Mercatelli D, Scalambra L, Triboli L, Ray F, Giorgi FM. Gene regulatory network inference resources: A practical overview. *Biochim Biophys Acta Gene Regul Mech*. 2020. 1863(6).
703. Ohta T. Role of gene duplication in evolution. *Genome*. 1989. 31(1).

704. Gu Z, Nicolae D, Lu HHS, Li WH. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet*. 2002. 18(12).
705. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019. 177(7).
706. Ilicic T, Kim JK, Kolodziejczyk AA, et al. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol*. 2016. 17(1).
707. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Softw*. 2018. 3(29).
708. Meistermann D, Bruneau A, Loubersac S, et al. Integrated pseudotime analysis of human pre-implantation embryo single-cell transcriptomes reveals the dynamics of lineage specification. *Cell Stem Cell*. 2021. 28(9).
709. Hermitte S, Chazaud C. Primitive endoderm differentiation: From specification to epithelium formation. *Philos Trans R Soc B Biol Sci*. 2014. 369(1657).
710. Stirparo GG, Boroviak T, Guo G, Nichols J, Smith A, Bertone P. Integrated analysis of single-cell embryo data yields a unified transcriptome signature for the human pre-implantation epiblast. *Development*. 2018. 145(3).
711. Hayakawa K, Ohgane J, Tanaka S, Yagi S, Shiota K. Oocyte-specific linker histone H1foo is an epigenomic modulator that decondenses chromatin and impairs pluripotency. *Epigenetics*. 2012. 7(9).
712. Zhang X, Lan Y, Xu J, et al. CellMarker: A manually curated resource of cell markers in human and mouse. *Nucleic Acids Res*. 2019. 47(D1).
713. Cao J, Spielmann M, Qiu X, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*. 2019. 566(7745).
714. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. 2014. 32(4).
715. Aibar S, González-Blas CB, Moerman T, et al. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods*. 2017. 14(11).
716. Moerman T, Aibar Santos S, Bravo González-Blas C, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*. 2019. 35(12).
717. Aibar S, Aerts S. RcisTarget: Identify transcription factor binding motifs enriched on a gene list. R package. 2016.
718. Van Der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008. 9(11).
719. Fan X, Tang D, Liao Y, et al. Single-cell RNA-seq analysis of mouse preimplantation embryos by third-generation sequencing. *PLOS Biol*. 2021.

18(12).

720. Maddox-Hyttel P, Dinnyés A, Laurincik J, et al. Gene expression during pre- and peri-implantation embryonic development in pigs. *Reprod Suppl*. 2001. 58.
721. Zhi M, Zhang J, Tang Q, et al. Generation and characterization of stable pig pregastrulation epiblast stem cell lines. *Cell Res*. 2022. 32(4).
722. Latham KE, Schultz RM. Embryonic genome activation. *Front Biosci*. 2001. 6(3).
723. Jiang S, Williams K, Kong X, et al. Single-nucleus RNA-seq identifies divergent populations of FSHD2 myotube nuclei. *PLOS Genet*. 2020. 16(5).
724. Tanaka Y, Kawazu M, Yasuda T, et al. Transcriptional activities of DUX4 fusions in B-cell acute lymphoblastic leukemia. *Haematologica*. 2018. 103(11).
725. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010. 28(5).
726. de Paepe C, Krivega M, Cauffman G, Geens M, van de Velde H. Totipotency and lineage segregation in the human embryo. *Mol Hum Reprod*. 2014. 20(7).
727. Kuijk E, Geijsen N, Cuppen E. Pluripotency in the light of the developmental hourglass. *Biol Rev*. 2015. 90(2).
728. True JR, Haag ES. Developmental system drift and flexibility in evolutionary trajectories. *Evol Dev*. 2001. 3(2).
729. Duboule D. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Development*. 1994. 1994(Supplement).
730. Raff RA. *The shape of life: genes, development, and the evolution of animal form*. Chicago: University of Chicago Press. 1996.
731. Kalinka AT, Varga KM, Gerrard DT, et al. Gene expression divergence recapitulates the developmental hourglass model. *Nature*. 2010. 468(7325).
732. Domazet-Lošo T, Tautz D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*. 2010. 468(7325).
733. Kalinka AT, Tomancak P. The evolution of early animal embryos: conservation or divergence? *Trends Ecol Evol*. 2012. 27(7).
734. Duboule D. The (unusual) heuristic value of Hox gene clusters; a matter of time? *Dev Biol*. 2022. 484.
735. Haag ES, True JR. Developmental system drift. In: Nuño de la Rosa L, Müller G, Eds. *Evolutionary Developmental Biology: A Reference Guide.* Springer. 2021. pp99-110.
736. Liu J, Viales RR, Khoueiry P, et al. The hourglass model of evolutionary conservation during embryogenesis extends to developmental enhancers with

signatures of positive selection. *Genome Res*. 2021. 31(9).

737. Raff RA, Smith MS. Axis formation and the rapid evolutionary transformation of larval form. *Curr Top Dev Biol*. 2009. 86.
738. Wray GA, Raff RA. Rapid evolution of gastrulation mechanisms in a sea urchin with lecithotrophic larvae. *Evolution*. 1991. 45(8).
739. Liu PZ, Kaufman TC. Short and long germ segmentation: unanswered questions in the evolution of a developmental mode. *Evol Dev*. 2005. 7(6).
740. Peel AD. The evolution of developmental gene networks: lessons from comparative studies on holometabolous insects. *Philos Trans R Soc B Biol Sci*. 2008. 363(1496).
741. Jaeger J. The gap gene network. *Cell Mol Life Sci*. 2011. 68(2).
742. Lynch JA, Roth S. The evolution of dorsal–ventral patterning mechanisms in insects. *Genes Dev*. 2011. 25(2).
743. Ardito G. Check-list of the data on the gestation length of primates. *J Hum Evol*. 1976. 5(2).
744. Allen WR, Wilsher S. A review of implantation and early placentation in the mare. *Placenta*. 2009. 30(12).
745. Fogarty NME, McCarthy A, Snijders KE, et al. Genome editing reveals a role for OCT4 in human embryogenesis. *Nature*. 2017. 550(7674).
746. Niu Y, Shen B, Cui Y, et al. Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos. *Cell*. 2014. 156(4).
747. Crispo M, Mulet AP, Tesson L, et al. Efficient generation of myostatin knock-out sheep using CRISPR/Cas9 technology and microinjection into zygotes. *PLoS One*. 2015. 10(8).
748. Wang X, Yu H, Lei A, et al. Generation of gene-modified goats targeting MSTN and FGF5 via zygote injection of CRISPR/Cas9 system. *Sci Rep*. 2015. 5(1).

Appendices

Appendix A: List of abbreviations

This appendix contains non-S.I. abbreviations used in this thesis. Gene names, protein names and species names are excluded.

λ – rate of gene gain
μ – rate of gene loss
2CLC – two-cell-like cells
2R – two rounds
8CLC – eight-cell-like cells
AIC_C – small sample Akaike information criterion
BFF – bovine foetal fibroblast
BLAST – Basic Local Alignment Search Tool
BUSTED – Branch-site Unrestricted Statistical Test for Episodic Diversification
bp – base pairs
CAFE – Computational Analysis of gene Family Evolution
ChIP-seq – chromatin immunoprecipitation sequencing
CMV – cytomegalovirus
CUT&RUN-seq – cleavage under targets and release using nuclease sequencing
DAPI – 4′,6-diamidino-2-phenylindole
DDC – duplication degeneration complementation
DE – differentially expressed
DGE – differential gene expression
DNA – deoxyribonucleic acid
DSD – developmental system drift
E – embryonic day
EAC – escape from adaptive conflict
EC – equivalence class
EGA – embryonic genome activation
EPI – epiblast
EpiSC – epiblast stem cell
ERV – endogenous retrovirus
ESC – embryonic stem cell
EST – expressed sequence tag
ETCHbox – Eutherian Totipotent Cell Homeobox

FC – fold change

FDR – false discovery rate

FEL – Fixed Effects Likelihood

GARD – Genetic Algorithm for Recombination Detection

GEO – Gene Expression Omnibus

GO – gene ontology

GRN – gene regulatory network

h – hours

HD – homeodomain

HDF – human dermal fibroblast

HERV – human endogenous retrovirus

HGT – horizontal gene transfer

ICM – inner cell mass

IGV – Integrative Genomics Viewer

iPSC – induced pluripotent stem cell

ISG – interferon-stimulated gene

K2P – Kimura 2-parameter

kb – kilobase

K-Pg – Cretaceous-Paleogene

KRAB – Krüppel-associated box

KZNF – KRAB zinc-finger

LFC – log fold change

LINE – long interspersed nuclear element

LRT – likelihood ratio test

LTR – long terminal repeat

MEME – Mixed Effects Model of Evolution

MERV – murine endogenous retrovirus

ML – maximum likelihood

NCBI – National Centre for Biotechnology Information

NGS – next generation sequencing

OKSM – OCT4, KLF4, SOX2, MYC

ONT – Oxford Nanopore Technology

ORF – open reading frame

PacBio – Pacific Biosciences

PAML – Phylogenetic Analysis by Maximum Likelihood

PBS – phosphate-buffered saline

PCR – polymerase chain reaction

PE – primitive endoderm

PFA - paraformaldehyde

piRNA – piwi-interacting RNA

PIWI – P-element induced wimpy testes

PLE – Penelope-like elements

QC – quality control

RNA – ribonucleic acid

RNA-seq – RNA-sequencing

rpm – revolutions per minute

SAGE – serial analysis of gene expression

SC – stem cell

SCENIC – Single-Cell rEgulatory Network Inference and Clustering

scRNA-seq – single cell RNA-sequencing

SINE – short interspersed nuclear element

siRNA – small interfering RNA

SPILE – spiralian TALE

SRA – Sequence Read Archive

SV40 – simian virus 40

SVA – SINE-VNTR-Alu

TBE – tris-borate-EDTA

TE – transposable element

TED – trophectoderm (non-standard abbreviation to distinguish from TE, transposable element)

TF – transcription factor

TPM – transcripts per million

t-SNE – t-distributed stochastic neighbour embedding

UMAP – uniform manifold approximation and projection

UMI – unique molecular identifier

UTR – untranslated region

WGD – whole genome duplication

$\mathbf{w_i}$ – Akaike weight

WT – wild type

ZGA – zygotic genome activation

ZP – zona pellucida

Appendix B: Additional files

This appendix lists the electronic supplementary data associated with this thesis.

ESM2.1 Supplementary Tables S2.1 – S2.13
ESM3.1 Supplementary Tables S3.1 – S3.10
ESM4.1 Supplementary Tables S4.1 – S4.8
ESM5.1 Supplementary Tables S5.1 – S5.9
ESM6.1 Supplementary Tables S6.1 – S6.11

Appendix C: Resulting publications

Two of the chapters comprising this thesis have been published in peer-reviewed journals at the time of thesis submission. This appendix contains PDFs of these two publications.

Chapter 2

Lewin TD, Royall AH, Holland PWH (2021). Dynamic molecular evolution of mammalian homeobox genes: duplication, loss, divergence and gene conversion sculpt PRD-class repertoires. *J Mol Evol*. 89(6):396-414.

Chapter 3

Lewin TD, Fouladi-Nashta AA, Holland PWH (2022). PRD-class homeobox genes in bovine early embryos: function, evolution, and overlapping roles. *Mol Biol Evol.* 39(5):msac098.

Journal of Molecular Evolution (2021) 89:396–414
https://doi.org/10.1007/s00239-021-10012-6

ORIGINAL ARTICLE

Dynamic Molecular Evolution of Mammalian Homeobox Genes: Duplication, Loss, Divergence and Gene Conversion Sculpt PRD Class Repertoires

Thomas D. Lewin[1] · **Amy H. Royall**[1] · **Peter W. H. Holland**[1]

Received: 5 March 2021 / Accepted: 11 May 2021 / Published online: 7 June 2021
© The Author(s) 2021

Abstract
The majority of homeobox genes are highly conserved across animals, but the eutherian-specific ETCHbox genes, embryonically expressed and highly divergent duplicates of *CRX*, are a notable exception. Here we compare the ETCHbox genes of 34 mammalian species, uncovering dynamic patterns of gene loss and tandem duplication, including the presence of a large tandem array of *LEUTX* loci in the genome of the European rabbit (*Oryctolagus cuniculus*). Despite extensive gene gain and loss, all sampled species possess at least two ETCHbox genes, suggesting their collective role is indispensable. We find evidence for positive selection and show that *TPRX1* and *TPRX2* have been the subject of repeated gene conversion across the Boreoeutheria, homogenising their sequences and preventing divergence, especially in the homeobox region. Together, these results are consistent with a model where mammalian ETCHbox genes are dynamic in evolution due to functional overlap, yet have collective indispensable roles.

Keywords Etchbox · Genome evolution · Homeodomain · Positive selection · Tandem duplication

Introduction

Homeobox genes encode a diverse set of transcription factors found across the Eukaryota, each of which has a characteristic DNA-binding homeodomain of around 60 amino acids (Duboule 1994; Derelle et al. 2007; Holland et al. 2007). Many homeobox genes play critical roles in early embryo patterning and cell fate specification (Wellik 2007; Mallo et al. 2010; Holland 2013) and, as fundamental components of developmental gene regulatory networks, are generally highly conserved over large phylogenetic distances (Bürglin and Affolter 2016). Indeed, most research on homeobox genes has focused on highly conserved examples, such as *HOX* (e.g. Burke et al. 1995; Duboule 2007; Maeda and Karch 2009; Mallo et al. 2010), *PAX* (e.g. Gruss and Walther 1992; Dahl et al. 1997; Blake and Ziman 2014), *POU* (e.g. Herr et al. 1988; Phillips and Luisi 2000) and *LIM* (Sheng et al. 1996; Hobert and Westphal 2000; Costello et al. 2015) genes.

In contrast, there are a smaller number of fast-evolving, taxon-specific homeobox genes found in some animals, including genes expressed during nematode (Bürglin and Cassata 2002; Mukherjee and Bürglin 2007), lepidopteran (Chai et al. 2008; Ferguson et al. 2014), spiralian (Paps et al. 2015; Morino et al. 2017) and mammalian (Maeso et al. 2016) embryonic development. We consider these genes to be fast-evolving on the basis of extensive amino acid change over relatively short timescales, following their origin by gene duplication. In some cases, the amino acid divergence from the deduced parental gene is so great as to cloud insights into their origins, unless additional information such as chromosomal location is also used.

Within mammals, the clearest examples of fast-evolving homeobox genes are *NANOGNB*, a member of the ANTP class (Dunwell and Holland 2017), and several loci classified within the PRD class, although they lack a PAIRED box. These genes include *CPHX1* and *CPHX2*, the *RHOX* genes, the double homeobox genes *DUXA* and *DUXB*, and the Eutherian Totipotent Cell Homeobox (ETCHbox) genes (MacLean et al. 2005; Töhönen et al. 2015; Madissoon et al. 2016; Maeso et al. 2016). Six paralogous groups

Handling editor: **David Liberles**

✉ Peter W. H. Holland
peter.holland@zoo.ox.ac.uk

[1] Department of Zoology, University of Oxford, 11a Mansfield Road, Oxford OX1 3SZ, UK

make up the ETCHbox genes—*Arginine-Fifty Homeobox* (*ARGFX*), *Divergent Paired-Related Homeobox* (*DPRX*), *Leucine-Twenty Homeobox* (*LEUTX*), *Parent of ARGFX* (*PARGFX*), *Tetra-Peptide Repeat Homeobox 1* (*TPRX1*) *and Tetra-Peptide Repeat Homeobox 2* (*TPRX2*)—all derived by duplication and extensive sequence divergence from the OTX-family member *Cone-rod homeobox* (*CRX*) (Booth and Holland 2007; Maeso et al. 2016). These genes are a synapomorphy of the Eutheria. They are absent from monotremes and marsupials, having arisen in the lineage leading to the eutherians, after which they underwent rapid asymmetric evolution and diverged extensively from *CRX* (Maeso et al. 2016).

The ETCHbox genes are notable for their remarkably specific temporal expression patterns. Though there are slight variations, human and cow ETCHbox genes are expressed between the 4-cell stage and early blastocyst of the preimplantation embryo and then never expressed again (Maeso et al. 2016). Despite being extensively duplicated, mouse ETCHbox genes are also expressed in the preimplantation embryo (Rajkovic et al. 2002; Cheng et al. 2007; Saito et al. 2010; Maeso et al. 2016; Royall et al. 2018). Furthermore, data from ectopic expression experiments in human fibroblasts and human embryonic stem cells suggest that ETCHbox genes regulate preimplantation embryo-expressed genes and that *LEUTX* has a role in embryonic genome activation (Jouhilahti et al. 2016; Maeso et al. 2016).

Maeso et al. (2016) published the most extensive characterisation of ETCHbox gene complements to date, comparing nine eutherian species. These data suggested that the genes are more dynamic than typical homeobox genes, with a high prevalence of gene duplication and loss, contrasting with the genes' conserved and highly specific expression pattern. However, this analysis was limited by sparse phylogenetic coverage and the low-quality genome assemblies available at the time. Katayama et al. (2018) performed a deeper analysis of *LEUTX* evolution, but this study was restricted to the one gene and without analysis of gene loss. Recent improvements to long-read DNA sequencing mean that many additional mammalian genome sequences are now available, assembled to high contiguity and accuracy. These data give a timely opportunity to describe the number and organisation of ETCHbox genes across the eutherian phylogeny, which is a necessary step towards understanding the reasons underpinning their unusual pattern of evolution.

The causes and consequences of ETCHbox genes' dynamic evolution are yet to be elucidated. It has been proposed that their dynamism may be driven by a possible role in the evolution of reproductive traits in mammals, such as placentation, which is highly variable between eutherians (Maeso et al. 2016), or due to selection in some lineages for shorter gestation times (Katayama et al. 2018). Alternatively, the genes' dynamic evolution may be a consequence of partial functional redundancy, which would cause relaxed selection on loss-of-function mutations, or distributed robustness, where the perturbation of one part of a system (loss of a gene) is compensated by non-redundant parts (other genes) (Wagner 2005; Royall et al. 2018). Finally, the genes may lack an important function, and, therefore, their high rates of pseudogenisation and loss would be a consequence of relaxed selection.

Here we analyse publicly available genome sequences to produce a dataset of ETCHbox repertoires for 34 mammals. We focus particularly on assemblies made with long-read DNA sequencing technology and species chosen to optimise phylogenetic coverage, allowing us to deduce with confidence the underlying patterns and pathways of gene duplication and loss. We uncover large arrays of tandem ETCHbox duplicates across multiple species and show that the ETCHbox genes have been the subject of positive selection and concerted evolution.

Materials and Methods

Comparative Genomics

Genome sequences for 32 eutherian species were downloaded from the NCBI webpage, focussing on taxa with high contiguity genome assemblies (Online Resource Table S1); this includes re-analysis of species previously analysed using lower quality genome data (Maeso et al. 2016). When possible, genomes sequenced using long-read technologies were utilised, as such data facilitates assessment of tandem gene duplication and gene loss. To include species from every order of the Boreoeutheria (Laurasiatheria and Euarchontoglires), three taxa were included despite lacking long-read assemblies: *Galeopterus variegatus* (Sunda flying lemur, Dermoptera), *Condylura cristata* (star-nosed mole, Eulipotyphla) and *Manis javanica* (Sunda pangolin, Pholidota). High-quality human and mouse genome assemblies were analysed by Maeso et al. (2016) and Royall et al. (2018), respectively, and are used but not recharacterised here, giving a total dataset of 34 species.

Homo sapiens (human; Maeso et al. 2016) and *Bos taurus* (cattle; this work) ETCHbox gene structures were verified using transcriptome data. Briefly, for *B. taurus*, RNA-seq reads (Online Resource Table S2; Graf et al. 2014; Jiang et al. 2014; Bernardo et al. 2018; Liu et al. 2018) were obtained from the NCBI Sequence Read Archive (SRA), aligned to the *B. taurus* reference genome ARS-UCD1.2 using STAR version 2.7.0 (Dobin et al. 2013) and assembled into transcripts using StringTie version 1.3.6 (Pertea et al. 2015). Genes of interest were identified in each transcriptome using the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990, 1997), and intron/exon

boundaries refined by inspection of raw reads using the Integrative Genomics Viewer (IGV) (Robinson et al. 2011).

For species lacking appropriate transcriptome data, ETCHbox genes were identified and annotated using sequence similarity searches of genome assemblies (blastn, blastp, megablast and tblastn; Altschul et al. 1990, 1997; Zhang et al. 2000). Gene identities were assigned using a combination of reciprocal BLAST, neighbouring genes and phylogenetic analysis (MrBayes; Huelsenbeck and Ronquist 2001; Ronquist et al. 2012). Intron/exon boundaries were refined manually using (a) retrogene sequences, (b) comparison to human and cow sequences validated by transcriptome data, and (c) mammalian consensus splice sites (Burset et al. 2000). The first coding exon of ETCHbox genes is very short and highly variable, and was therefore not always identified. Genes were considered probable pseudogenes when there were stop codons, splice site mutations or frameshifts upstream of (or within) the homeobox. Genes with frameshifts or premature stop codons immediately downstream of the homeobox are of unknown functional status. If no gene was identified by BLAST and the expected syntenic region surrounding the gene was split over two or more scaffolds we do not conclude certain gene loss.

For phylogenetic analysis, amino acid sequence alignments were made using Clustal Omega in Seaview version 4.7 (Gouy et al. 2010; Sievers et al. 2011) and phylogenies inferred using MrBayes version 3.2.7a (Huelsenbeck and Ronquist 2001; Ronquist et al. 2012) and rendered using iTOL (Letunic and Bork 2019).

Estimating Gene Gain and Loss

Two methods were used to assess gene gain and loss. First, ETCHbox genes were grouped into gene families and the stochastic birth and death model in CAFE (De Bie et al. 2006; Han et al. 2013) used to calculate maximum likelihood values of λ and μ (rates of gain or loss, respectively, per gene per million years) and estimate gene numbers at internal nodes. Second, the event-inference parsimony algorithm in Notung version 2.9 (Chen et al. 2000; Durand et al. 2006) was used. Gene trees were generated for each ETCHbox gene as above and Notung run with a duplication-loss model and default parameters (weights: duplications = 1.5, co-divergences = 0.0, losses = 1.0) to reconcile gene and species trees and estimate the timing and minimum weighted number of independent duplication and loss events. To prevent weakly supported branches causing overestimation of gene turnover, gene trees were rearranged using the 'Rearrange' function, allowing branches with posterior probabilities < 95% to be reconfigured to minimise duplications and losses. The species tree used was generated using TimeTree (Kumar et al. 2017).

CAFE was also used to test for an acceleration in the rate of gene duplication of each ETCHbox gene compared to other homeobox genes present in mammals using the Monte Carlo sampling procedure described previously (Hahn et al. 2005, 2007). The Viterbi assignment method (De Bie et al. 2006) was used to establish which branches contributed to such accelerations. For the purpose of gene duplication analyses, Cetartiodactyla *TPRX3* genes were assigned as *TPRX2* duplicates, as by Maeso et al. (2016).

Detecting Gene Conversion

We defined *TPRX1* and *TPRX2* using neighbouring genes and orientation, not sequence: *TPRX1* is upstream of *CRX* and in inverse orientation, *TPRX2* is downstream of *CRX* on the same strand. To test for interlocus gene conversion between *TPRX1* and *TPRX2*, four methods were used, following the guidelines of Mansai and Innan (2010). First, the expected *TPRX* duplication history was compared to Bayesian gene trees to search for phylogenetic incompatibilities. Protein sequence alignments of TPRX1 and TPRX2 were trimmed using Gblocks version 0.91b (Talavera and Castresana 2007), converted to codon alignments using PAL2NAL (Suyama et al. 2006) and compared using the phylogenetic methods outlined above. Second, sequence similarity was assessed by running Biostrings version 2.57.1 (Pagès et al. 2020) in R version 4.0.0 'Arbor Day' (R Core Team 2020) to compute all versus all Needleman-Wunsch (Needleman and Wunsch 1970) global pairwise alignments. Percent nucleotide identities were calculated and plotted using gplots version 3.0.3 (Warnes et al. 2020). To understand whether sequence similarity is constant across the length of the genes, a sliding window analysis was performed using Spider version 1.5.0 (Brown et al. 2012), measuring Kimura 2-parameter (K2P) distance (Kimura 1980) between the *TPRX1* and *TPRX2* genes of a given species in 50 base pair (bp) overlapping windows with increments of 1 bp. Only species with at least one putatively functional copy of both *TPRX1* and *TPRX2* were used.

Third, we tested for incompatibilities between phylogenies built using different partitions of the genes. The HyPhy (Kosakovsky Pond et al. 2005, 2020) programme GARD (Kosakovsky Pond et al. 2006a, b) was run using Datamonkey (Weaver et al. 2018) with default parameters on codon alignments of all *TPRX1* and *TPRX2* genes (GARD was also run on alignments of *Oryctolagus cuniculus* and *Microcebus murinus LEUTX* tandem duplicates and *Peromyscus leucopus TPRX* and *LEUTX* genes). GARD uses an aggressive population-based hill climber to search multiple sequence alignments for phylogenetic incongruity and identify putative gene conversion and recombination breakpoints. The AIC_C (small sample Akaike Information Criterion) was used to select the model with the best fit to the data, with

Akaike weights (w_i) calculated using the R package qpcR version 1.4.1 and used to assist model selection (Akaike 1974; Sugiura 1978; Hurvich and Tsai 1989; Burnham and Anderson 2002; Wagenmakers and Farrell 2004; Ritz and Spiess 2008). The alignment was split into two partitions based on the breakpoint identified by GARD, and Bayesian phylogenies of each partition built as above. To measure tree dissimilarity, the tqDist algorithm (Sand et al. 2014) was implemented in the R package Quartet version 1.2.0 (Smith 2020) to calculate quartet distance (Estabrook et al. 1985) and quartet divergence (Smith 2019). Finally, GENECONV version 1.81a (Sawyer 1989) was used to identify putative gene conversion events by searching for fragments of sequences with sufficient nucleotide similarity to suggest gene conversion. GENECONV was run with default parameters apart from /lp (implements pairwise comparisons), /w123 (creates reproducible results by initiating at the same random seed number) and -gscale = 2 (allows mismatches in the conversion tracts with a penalty of 2). GENECONV returns *p* values based on 10,000 permutations for fragments found with global (corrected for multiple comparisons) and pairwise sequence comparisons (corrected for alignment length but not multiple sequence comparisons). Given a significance threshold of 0.05, it is expected that if there was no gene conversion in the dataset, then 69 of the 1378 pairwise comparisons would produce false positives. We identified 613 events, suggesting that the majority are not false positives. Furthermore, as a negative control, GENECONV was run with option -Randomize_sites; this permutes sites once and therefore removes any gene conversion signal. This identified just seven gene conversion events, again suggesting that the events detected above are not false positives. Events identified in the negative control analysis were discarded from the results. Only putatively functional genes were included in the gene conversion analysis, and *Mus musculus* and *Peromyscus leucopus Obox* (*TPRX2*) genes were also omitted because they show extreme lineage-specific sequence divergence and their inclusion may disrupt analysis.

Tests for Accelerated Divergence and Positive Selection

To test for changes in the rate of homeodomain sequence evolution, MEGA X (Kumar et al. 2018; Stecher et al. 2020) was used to undertake Tajima's relative rate test (Tajima 1993) ($\alpha = 0.05$). Each ETCHbox homeodomain was compared to its conspecific CRX protein, using a marsupial CRX sequence (*Monodelphis domestica*) as an outgroup. Where there are lineage-specific duplications, only one duplicate was used. The Benjamini-Yekutieli (Benjamini and Yekutieli 2001) false discovery rate method (false discovery rate = 0.05) was used to correct for multiple testing as it does not require independence of tests.

Episodic positive selection in ETCHbox genes was detected using the HyPhy (Kosakovsky Pond et al. 2005, 2020) Branch-Site Unrestricted Statistical Test for Episodic Diversification (BUSTED) (Murrell et al. 2015) via Datamonkey (Weaver et al. 2018) with default parameters using codon alignments generated with PAL2NAL (Suyama et al. 2006) and phylogenies reflecting known species relationships. To test for pervasive positive selection, pamlX (Xu and Yang 2013) was used to run CODEML (Model = 0, NSsites = 0, 1, 2, 7, 8) in Phylogenetic Analysis by Maximum Likelihood (PAML) version 4.8 (Yang 1997, 2007). Sites with a gap in more than 50% of sequences were removed, and CODEML run with the option cleanData = 0. Likelihood ratio tests (LRTs) were used to compare model 2 (M2, positive selection model) to model 1 (M1, nearly neutral model) and model 8 (M8, beta and ω model—positive selection) to model 7 (M7, beta model—no positive selection).

The HyPhy Mixed Effects Model of Evolution (MEME) (Murrell et al. 2012), which uses mixed-effects branch-site models, was used to detect specific codon sites evolving under episodic positive selection. MEME is preferred to the branch-site mode of CODEML because it does not require a priori specification of branches to be tested but retains good statistical power (Lu and Guindon 2014). Sites with a gap in more than 50% of sequences were removed from this analysis. Position of residues in relation to homeodomain structure was deduced by comparative structural modelling to the PRD-class homeodomain of *Drosophila melanogaster* Aristaless (Al) in complex with DNA (RCSB Protein Data Bank entry 3LNQ; Berman et al. 2000; Miyazono et al. 2010) using Modeller (Šali and Blundell 1993) implemented in UCSF Chimera 1.15 (Pettersen et al. 2004).

RELAX (Wertheim et al. 2015) was run with default parameters using codon alignments of ETCHbox and *CRX* homeoboxes to test for relaxed selection in each ETCHbox gene versus a reference group of six *CRX* genes (*Canis lupus familiaris*, *Condylura cristata*, *Equus caballus*, *Homo sapiens*, *Mus musculus* and *Ovis aries*).

Mus musculus and *Peromyscus leucopus Obox* (*TPRX2*) genes and genes with frameshifts or early stop codons downstream of the homeodomain were omitted from the selection analysis. Furthermore, the phylogenetic incongruity caused by gene conversion could lead to inaccurate results when testing for selection (Anisimova et al. 2003; Shriner et al. 2003; Kosakovsky Pond et al. 2006b). To account for gene conversion in the *TPRX* genes, we used the gene conversion breakpoint identified by GARD (Kosakovsky Pond et al. 2006a, b) to partition the alignment into two sections. MrBayes (Huelsenbeck and Ronquist 2001; Ronquist et al. 2012) was used as above to calculate gene trees for each

partition. The above methods were then performed separately for each of the two partitions.

Results

Identification of ETCHbox Genes in Eutherian Genomes

We first characterised the ETCHbox genes of *Bos taurus* (cattle) using transcriptome data (Fig. 1). *B. taurus* possesses putatively functional *ARGFX*, *LEUTX*, *TPRX1* and *TPRX2* genes, but *DPRX* is a putative pseudogene due to a 2 bp insertion in the homeobox and the loss of exon 1; *PARGFX* has been lost. *B. taurus* also possesses a *TPRX* duplicate, which we refer to as *TPRX3*. Compared with *Homo sapiens* (human), *B. taurus ARGFX* has an additional 5' coding exon, which extends the reading frame. All genes are located in the same syntenic position as in humans, with *LEUTX*, *TPRX1*, *TPRX2* and *DPRX* (and *TPRX3*) in a loose cluster on chromosome 18 (human chromosome 19), and *ARGFX* separate from the cluster on chromosome 1 (human chromosome 3).

We then characterised the ETCHbox gene repertoires in the genomes of 31 further eutherian species, using a combination of phylogenetics, synteny and reciprocal BLAST searches to assign gene identities (Fig. 2, 3; Online Resource File 1). These methods concurred in almost all cases except for *TPRX1* and *TPRX2* genes, where sequence-based methods disagreed with genomic position; for these genes we use genomic position to assign gene name and assess below whether incongruence is due to gene conversion. The only other discordance occurs in *Peromyscus leucopus* (white-footed mouse), where the two genes at the *LEUTX* locus cluster with rodent *TPRX1* genes, although we find no evidence of gene conversion in this case (GENECONV analysis, no gene conversion fragments identified). In phylogenetic analyses, branch lengths are longer for ETCHbox genes than for their paralogues *CRX* and *OTX1*, indicating a higher amino acid substitution rate. Particularly long branches are observed for *Oryctolagus cuniculus* (European rabbit) *TPRX2* and *LEUTX1*, *Microcebus murinus* (gray mouse lemur) *PARGFX*, and *Mus musculus* (house mouse) and *P. leucopus TPRX1* (= *Crxos*) and *TPRX2* (= *Obox*). Loci with a stop codon, frameshift or splice site disruption in, or upstream from, the homeodomain are inferred to be pseudogenes. The ETCHbox genes frequently spawn retrocopies; these were also characterised, with every sampled species possessing at least one ETCHbox retrogene (Online Resource Table S3a); retrocopies are not clustered, and are found

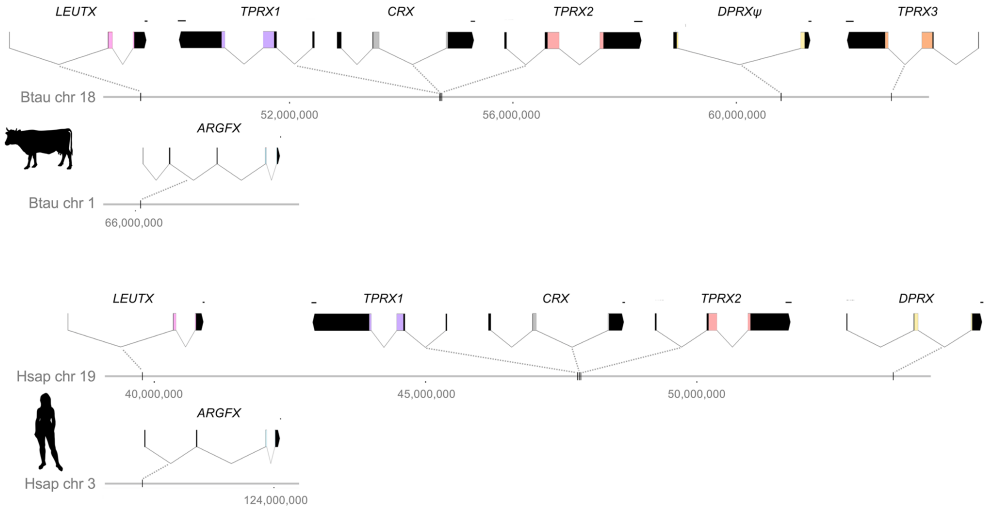


Fig. 1 ETCHbox repertoires of *Homo sapiens* (humans) and *Bos taurus* (cattle), with gene structures as determined using transcriptome assemblies. Horizontal grey bars represent chromosomes, vertical black bars represent the genomic position of ETCHbox genes. For gene structure representations, coding regions are shown in black, homeoboxes in colour. Untranslated regions (UTRs) are not shown. Black scale bars at 3' end of genes = 100 bp. *DPRX*, *LEUTX*, *TPRX1*, *TPRX2* (and *B. taurus TPRX3*) form a loose cluster on a single chromosome (*B. taurus* chromosome 18, *H. sapiens* chromosome 19); *ARGFX* has translocated to another chromosome (*B. taurus* chromosome 1, *H. sapiens* chromosome 3). *TPRX1* and *TPRX2* are located either side of the ETCHbox 'ancestor' *CRX*

Journal of Molecular Evolution (2021) 89:396–414

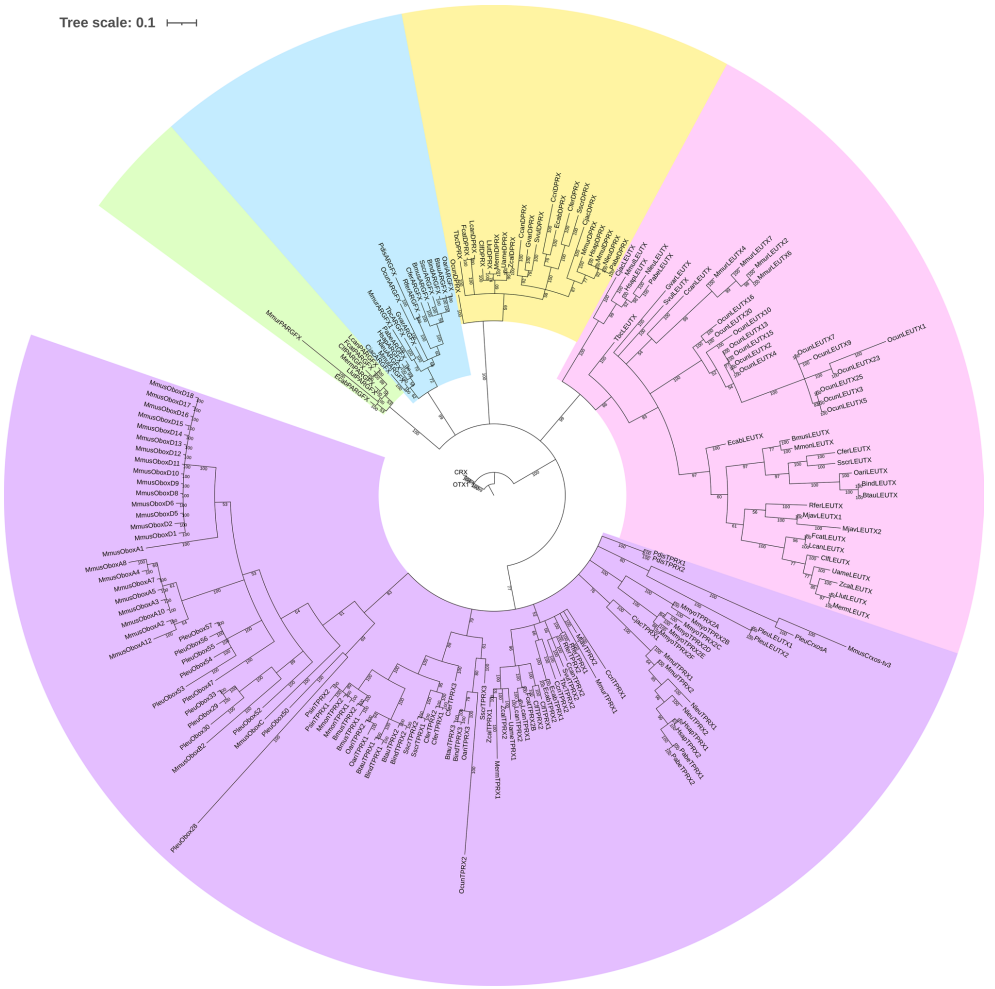


Fig. 2 Bayesian gene tree of putatively functional ETCHbox genes identified in this work. Colours highlight ETCHbox gene families; labels show posterior probabilities. The *ARGFX*, *DPRX*, *LEUTX* and *PARGFX* clades are supported by ≥ 99% probabilities. Due to the limited length of the homeodomain (60 amino acids), gene phylogenies do not always recapitulate known relationships between species. *TPRX* duplicates in Cetartiodactyla are referred to as *TPRX3*. The *TPRX1* and *TPRX2* genes of *Mus musculus* and *Peromyscus leucopus* are referred to as *Crxos* and *Obox*, respectively, reflecting their extensive sequence change compared to the ancestral *TPRX* genes. Abbreviations: Bind = *Bos indicus*, Bmus = *Balaenoptera musculus*, Btau = *Bos taurus*, Ccan = *Castor canadensis*, Ccri = *Condylura cristata*, Cfer = *Camelus ferus*, Cjac = *Callithrix jacchus*, Clf = *Canis lupus familiaris*, Ecab = *Equus caballus*, Fcat = *Felis catus*, Gvar = *Galeopterus variegatus*, Hsap = *Homo sapiens*, Lcan = *Lynx canadensis*, Llut = *Lutra lutra*, Mjav = *Manis javanica*, Merm = *Mustela erminea*, Mmon = *Monodon monoceros*, Mmul = *Macaca mulatta*, Mmur = *Microcebus murinus*, Mmus = *Mus musculus*, Mmyo = *Myotis myotis*, Nleu = *Nomascus leucogenys*, Oari = *Ovis aries*, Ocun = *Oryctolagus cuniculus*, Pabe = *Pongo abelii*, Pdis = *Phyllostomus discolor*, Pleu = *Peromyscus leucopus*, Psin = *Phocoena sinus*, Rfer = *Rhinolophus ferrumequinum*, Sscr = *Sus scrofa*, Svul = *Sciurus vulgaris*, Tbc = *Tupaia belangeri chinensis*, Uame = *Ursus americanus*, Zcal = *Zalophus californianus*

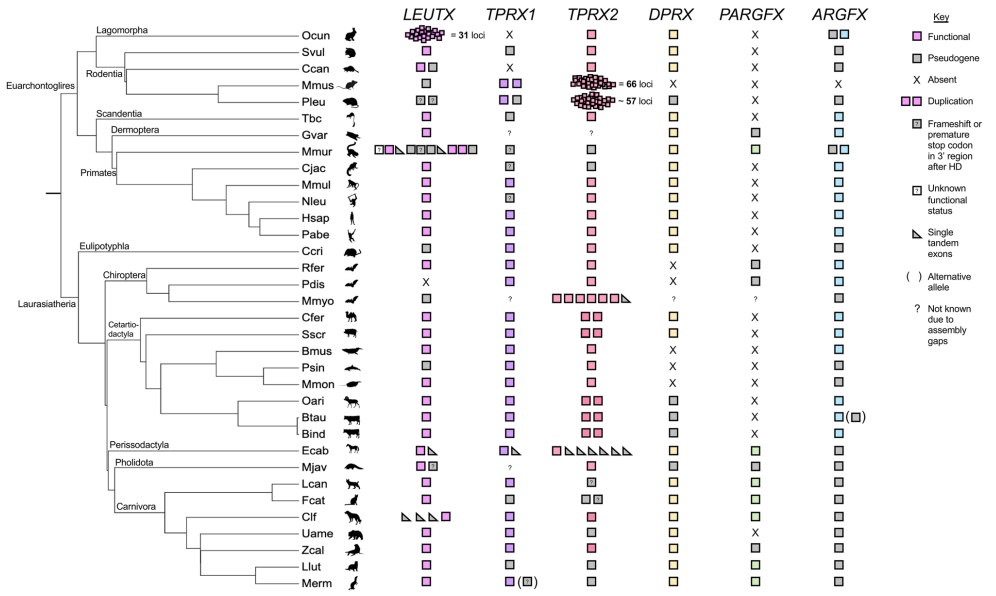


Fig. 3 ETCHbox gene repertoires of 34 eutherian mammals. Phylogenetic relationships are based on TimeTree (Kumar et al. 2017). Coloured boxes = putatively functional gene. Multiple boxes = gene duplicates. Black X = no gene remnants (complete gene loss). Grey boxes = putative pseudogene; grey boxes with a black question mark = complete homeodomain but subsequent frameshift or premature stop codon. White boxes with a question mark = unclear functional status due to incomplete assembly in the region. Grey triangles = tandem single exons. Brackets = polymorphism; question marks = assembly gap such that gene presence or absence cannot be determined. HD = homeodomain. Species abbreviations as in Fig. 2

dispersed around the genome (e.g. *Homo sapiens* and *Bos taurus;* Online Resource Table S3b).

The ETCHbox gene repertoires are highly variable between species, with additional tandem gene duplication, pseudogenisation and gene loss occurring repeatedly across eutherians (Fig. 3). Previous work showed that all six ETCHbox genes were present in the ancestor of the Boreoeutheria (Maeso et al. 2016) so absence at a terminal node implies gene loss. All sampled species have lost at least one ETCHbox gene, and each gene has been lost in at least one sampled species. Some gene losses are inferred to have occurred in the ancestors of large clades (e.g. *ARGFX* in the Carnivora and *DPRX* in the Cetruminantia); many other losses are more recent (e.g. *LEUTX* is lost in *Phocoena sinus* [vaquita] but present in other sampled Cetacea species).

In *Mi. murinus*, we note the first putatively functional *PARGFX* gene reported for any member of the Euarchontoglires. *Mi. murinus PARGFX* is in the expected syntenic position and groups phylogenetically with other *PARGFX* genes, albeit on a long branch (Fig. 2). *Galeopterus variegatus* (Sunda flying lemur) also has a detectable *PARGFX* locus, but it is inferred to be a pseudogene.

Giant Arrays of ETCHbox Genes

We identify several arrays of tandem ETCHbox duplicates, one of the largest of which is an array of *LEUTX* loci in *O. cuniculus*. Previous work detected six loci (Katayama et al. 2018), whereas we detect 27 gene copies in the assembly analysed, of which 14 are putatively functional, 11 are putative pseudogenes and two are of uncertain functional status (Online Resource Fig. S1). We also find four single exons in the cluster, giving a total of 31 loci. This is the largest *LEUTX* expansion discovered and one of the largest ETCHbox expansions, smaller than only those of *Mu. musculus* and *P. leucopus Obox* (*TPRX2*) genes (Royall et al. 2018). An inversion on *O. cuniculus* chromosome 5 has split the array in two, with *LEUTX1* to *LEUTX5* approximately 9 Mb from *LEUTX6* to *LEUTX27*. *Mi. murinus* also has a tandem *LEUTX* expansion of 10 loci, at least three of which are putatively functional, and *Mi. murinus* and *O. cuniculus* are both also notable because they have an *ARGFX* duplication. We find several cases of tandem duplication of single exons, including at *Equus caballus* (domestic horse) *LEUTX*, *TPRX1* and *TPRX2* loci.

It was shown previously that *Mu. musculus* has lost *ARGFX*, *DPRX*, *LEUTX* and *PARGFX* and possesses two *TPRX1* copies (called *Crxos*) and 66 *TPRX2* loci (called *Obox*), all of which are highly divergent in sequence (Maeso et al. 2016; Royall et al. 2018). We asked when the transition to this highly derived state occurred. Our results indicate that this evolved within the rodents. *Sciurus vulgaris* (red squirrel, Sciuridae) and *Castor canadensis* (American beaver, Castoridae) possess putatively functional *DPRX* and *LEUTX* genes, and neither have *TPRX1* or *TPRX2* duplicates (Fig. 3). However, *Peromyscus leucopus* (white-footed mouse, Cricetidae) has two *TPRX1* loci, and no functional *ARGFX*, *DPRX* or *PARGFX*, as in *Mu. musculus*. Furthermore, we detect 57 *P. leucopus TPRX2 (Obox)* loci, of which 12 are putatively functional. Seven of these loci have escaped the *TPRX2* cluster on chromosome 1 and form a separate cluster on chromosome 12. The observation that *P. leucopus TPRX1* and *TPRX2* genes group phylogenetically with *Mu. musculus Crxos* and *Obox*, respectively (Fig. 2), combined with the Notung result that the *TPRX1* duplication and several of the *TPRX2* duplications occurred before the split of *Mu. musculus* and *P. leucopus* (below), suggests that the transition from *TPRX1* and *TPRX2* to *Crxos* and *Obox*-like states occurred before the split of the Muridae and Cricetidae.

Rates of Gene Duplication

We compared rates of gene duplication and loss for each gene by modelling a stochastic birth–death process using CAFE (De Bie et al. 2006; Han et al. 2013), giving maximum likelihood estimates for the rates of ETCHbox gene gain and loss (events per million years; λ and μ, respectively; Table 1). CAFE was also used to infer likely ancestral gene numbers (Online Resource Fig. S2). Rates of gain (λ) and loss (μ) are highly variable between ETCHbox families. *TPRX2* is the gene most prone to duplication ($\lambda = 0.016$) and *PARGFX* most prone to gene loss ($\mu = 0.019$). *ARGFX*, *DPRX* and *PARGFX* have very low rates of gene gain but relatively high rates of loss.

We find evidence that *LEUTX* ($p = 0.003$) and *TPRX2* ($p = 0.000$) duplicate faster than other homeobox genes. The Viterbi assignment method (De Bie et al. 2006) reveals that the high overall duplication rate of *LEUTX* is primarily a result of changes along the *O. cuniculus* ($p = 1.503 \times 10^{-8}$) and *Mi. murinus* ($p = 0.028$) branches; the high duplication rate of *TPRX2* is influenced largely by the branches leading to Cetartiodactyla ($p = 0.027$), *Myotis myotis* (greater mouse-eared bat, $p = 0.001$), Muroidea ($p = 2.774 \times 10^{-10}$), *Mu. musculus* ($p = 1.774 \times 10^{-36}$) and *P. leucopus* ($p = 0.003$).

A high duplication rate for *LEUTX* and *TPRX2* was also supported by analysis incorporating gene trees, implemented using Notung (Chen et al. 2000; Durand et al. 2006) to estimate the number of duplication and loss events and infer their timings (Table 1 and Online Resource Fig. S3). Gene loss is expected to have most functional relevance when a single copy gene transitions to total absence of a functional gene; we find this occurred most for *PARGFX,* in accordance with CAFE results. There are two cases of apparent gene turnover overestimation: Notung reports the *Camelus ferus* (Bactrian camel) *TPRX3* duplication as independent of other Cetartiodactyla *TPRX3* duplicates, and a *DPRX* duplication at the base of the Caniformia followed by multiple losses. These are likely artefacts caused by the rapidly evolving nature of ETCHbox sequences but do not distort the overall inferences from the analysis.

Polymorphism in ETCHbox Genes

We find two cases of ETCHbox intraspecific polymorphism where one allele has a frameshift mutation. In *Mustela erminea* (stoat), we identify a putatively functional *TPRX1* in one haplotype of the phased genome assembly while the alternate haplotype has a 'CC' dinucleotide insertion causing a frameshift in exon 3. In the *B. taurus* reference genome (ARS-UCD1.2), we find a 13 bp deletion in *ARGFX* exon 2

Table 1 Duplication and loss in the ETCHbox genes

Gene	λ (gains per gene per million years, CAFE)	μ (losses per gene per million years, CAFE)	Estimated number of gene duplication events (Notung)	Estimated number of gene loss events (Notung)	Number of species with at least one putatively functional gene
ARGFX	3.37E-10	6.70E-03	0	5	17
DPRX	4.26E-11	3.42E-03	1	8	22
LEUTX	8.88E-03	3.32E-03	18	4	29
PARGFX	5.84E-11	1.93E-02	0	10	7
TPRX1	8.49E-04	6.15E-03	1	10	25
TPRX2	1.61E-02	6.33E-03	41	7	28

Probability of duplication or loss (λ and μ) for each ETCHbox gene, estimated by CAFE's stochastic birth and death model (De Bie et al. 2006; Han et al. 2013), together with estimates of numbers of gene duplication and gene loss events, calculated by Notung (Chen et al. 2000; Durand et al. 2006). Pseudogenes are excluded as duplication events

that causes a frameshift and a premature stop codon before the homeobox, making it a putative pseudogene. We do not find this deletion in several other *B. taurus* datasets (Online Resource Table S2) or in the genome of other Bovidae species (Online Resource Table S4).

TPRX1 and TPRX2 have been Subject to Repeated Gene Conversion

Interlocus gene conversion is a naturally occurring 'copy and paste' process that can take place during double-strand break repair, where DNA sequence from one locus is used to replace DNA sequence at a different locus in the same genome (Chen et al. 2007). The incongruence between gene identity inferred from phylogenetics versus gene position for *TPRX1* and *TPRX2* suggests that gene conversion may have occurred between these loci, as suggested previously (Maeso et al. 2016). However, this hypothesis needs further testing, and it is currently unclear whether the complete loci were affected, when it occurred or how often it occurred in evolution.

We first investigated these questions using a phylogenetic approach, searching for incompatibilities between the known species tree and the inferred gene tree. Under the null hypothesis of no gene conversion, *TPRX1* and *TPRX2* genes would form separate clades diverging since the base of eutherians; gene conversion would result in paralogues grouping more closely together. Bayesian nucleotide phylogenies of putatively functional *TPRX* genes reveal eight cases where the *TPRX1* and *TPRX2* genes from a given species group together as pairs of sister sequences, suggesting recent gene conversion events in these lineages (Fig. 4a, blue boxes). There are also indications of further gene conversion events deeper in the phylogeny, notably in the stem lineages of Cetacea, Bovidae, Carnivora and Primates (Fig. 4a, blue dots). Intriguingly, we found evidence for additional gene conversion events when phylogenetic analysis was restricted to the homeobox sequence only. This revealed 13 recent conversion events between *TPRX* loci, with five new cases identified in addition to the eight above (Fig. 4b, pink boxes). Several of the additional events are nested within the clades that showed evidence of older gene conversion (Primates, Cetacea), suggesting successive gene conversion events in evolution. The occurrence of successive gene conversion events is also supported by analysis of pairwise nucleotide identity (Online Resource Fig. S4) which, for example, suggests gene conversion at the base of the Cetartiodactyla, then further events within the Bovidae, Cetacea, *Sus scrofa* (domestic pig) and *C. ferus*.

Since a homeobox-only tree suggests additional episodes of gene conversion, we hypothesised that the 5' region of *TPRX* genes is more prone to gene conversion than the 3' region. To test this, we conducted a sliding window analysis calculating pairwise Kimura 2-parameter (K2P) distances (Kimura 1980) between *TPRX1* and *TPRX2* sequences (Fig. 5). In *Bos indicus* (Zebu cattle), *B. taurus*, *Balaenoptera musculus* (blue whale), *C. ferus*, *Felis catus* (domestic cat), *H. sapiens*, *Lynx canadensis* (Canada lynx), *Monodon monoceros* (narwhal), *Nomascus leucogenys* (northern white-cheeked gibbon), *Ovis aries* (sheep), *P. sinus* and *S. scrofa*, the lowest sequence distances (highest similarities) are at the 5' end, suggesting this region is more prone to homogenisation via gene conversion.

Gene conversion occurring repeatedly in one region of a gene pair is predicted to result in differences between phylogenetic trees built from different sub-regions of the genes. Using GARD, which tests for phylogenetic incongruence within a gene (Kosakovsky Pond et al. 2006a, b), we identified a consistent putative gene conversion breakpoint, located immediately downstream of the homeobox, dividing the gene into two regions with different phylogenetic histories (null model $AIC_C = 54{,}094.1$, breakpoint model $AIC_C = 52{,}467.8$, $\Delta AIC_C = 1626.3$; null model Akaike weight $(w_i) = 0$, breakpoint model Akaike weight $(w_i) = 1$; breakpoint model receives 100% of the weight of the models compared). Bayesian nucleotide sequence phylogenies of partition 1 (including the homeobox) and partition 2 (downstream) show different topologies (Fig. 6; quartet distance = 41,506; quartet divergence = 0.102). Partition 1 trees show more gene conversion events than partition 2 (e.g. Bovidae, Cetacea and Primates in Fig. 6), reinforcing the hypothesis that the 5' region is more prone to gene conversion.

Finally, we used GENECONV (Sawyer 1989) to search for long regions of unusually high sequence identity in multiple sequence alignments as further evidence of gene conversion. In the 53 TPRX sequences analysed, GENECONV identifies eight gene conversion events by global comparisons (Online Resource Table S5) and 613 fragments by pairwise comparisons (Online Resource Table S6), ranging from 9 to 492 bp in length. Pairwise comparisons are particularly powerful for detecting very recent gene conversion. For example, we find evidence for ten 'species-specific' gene conversion events (Online Resource Table S7), eight of which were also detected by phylogenetic methods as forming pairs in the homeobox-only tree (Fig. 4b). Notably, GENECONV and GARD give similar locations for the position of gene conversion breakpoints between 5' and 3' regions, and both show that the upstream region is subject to more frequent gene conversion than the downstream region. Across all species, no fragments have GENECONV breakpoints downstream of position 492 of the 2331 bp multiple sequence alignment (Online Resource Fig. S5); in the majority of species, position 492 is very close to the putative gene conversion breakpoint identified by GARD, and in human they are

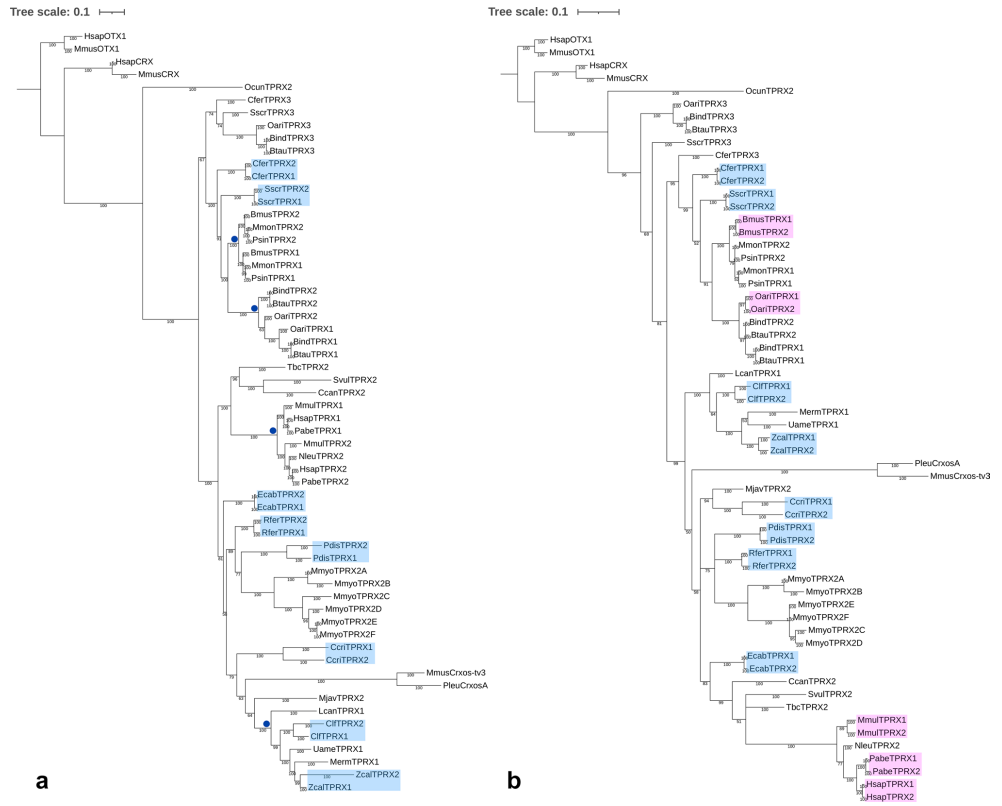


Fig. 4 Bayesian phylogenies of putatively functional *TPRX1*, *TPRX2* and *TPRX3* full gene sequences (**a**) and homeoboxes (**b**). Blue boxes highlight cases where conspecific *TPRX1* and *TPRX2* pairs are more closely related to each other than to other sequences. Pink boxes highlight cases that appear on tree b but not tree a. Blue dots mark putative gene conversion events that occurred deeper in the phylogeny. Putative pseudogenes were excluded. Labels show posterior probabilities. Species abbreviations as in Fig. 2

only nine nucleotides apart (a large insertion in bats means that they are further apart in the full alignment, Online Resource Fig. S6). This corroboration by two methods lends strong support to this partition, which is within exon 3, downstream of the homeobox.

We note that gene conversion continued to occur between the six *My. myotis TPRX2* duplicates following tandem duplication, with 13 fragments identified by GENECONV (Online Resource Table S6). Furthermore, gene conversion in the ETCHbox genes is not limited to *TPRX*. Pairwise analysis using GENECONV identifies 15 gene conversion events between *Mi. murinus LEUTX* tandem duplicates and 367 events between *O. cuniculus LEUTX* duplicates (Online Resource Table S8). Both results are reinforced by GARD (Online Resource Table S9).

Positive Selection in ETCHbox Genes

Using Tajima's relative rate test (Tajima 1993), we find that all ETCHbox sequences have a faster evolutionary rate than their sister gene *CRX* (124 genes analysed; Online Resource Table S10). To investigate if the elevated evolutionary rates are due to positive selection, we used BUSTED (Murrell et al. 2015) to test for episodes of selection that may vary over time and between lineages, and MEME (Murrell et al. 2012) to identify specific sites under selection. These analyses were performed on complete coding sequences of *ARGFX*, *DPRX*, *LEUTX*, *PARGFX* and *TPRX*. However, since gene conversion leads to phylogenetic incongruity, which interferes with detecting positive selection (Anisimova et al. 2003; Shriner et al. 2003; Kosakovsky Pond

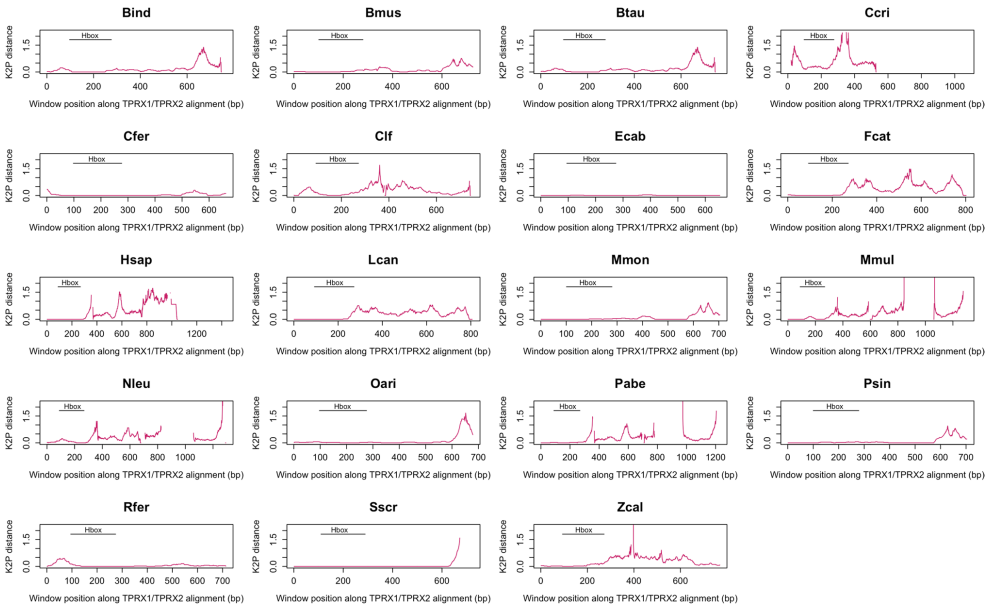


Fig. 5 Sequence similarity between *TPRX1* and *TPRX2* genes within a species. Plots show the Kimura 2-parameter (K2P) distance in 50 bp sliding windows between conspecific *TPRX1* and *TPRX2* genes. Higher K2P values indicate more divergent sequences. Gaps in the trace indicate indels in the alignment. The black bar marked 'Hbox' demarcates the position of the homeobox in each alignment. For many species, the K2P values increase towards the 3' end of the gene, suggesting that the *TPRX* genes have been homogenised by gene conversion less at their 3' ends. Putative pseudogenes were excluded. Species abbreviations as in Fig. 2

et al. 2006b), we divided *TPRX* genes at the gene conversion breakpoint identified by GARD into 5' and 3' regions and performed analyses separately on the two regions.

Using BUSTED (Murrell et al. 2015), we detect evidence of episodic positive selection during the evolution of *ARGFX*, *DPRX*, *LEUTX*, *PARGFX* and both partitions of the *TPRX* genes (LRT $p < 0.05$ for all genes). We also find strong evidence for positive selection in *ARGFX*, *DPRX*, *LEUTX*, *TPRX* partition 1 and *TPRX* partition 2, but not *PARGFX*, using CODEML (Yang 1997, 2007) (Online Resource Table S11), supporting the BUSTED result. Using MEME (Murrell et al. 2012), we find evidence for positive selection acting on between 3 (*PARGFX*) and 31 (*TPRX*) codons in each gene (Online Resource Table S12). The sites deduced to have undergone positive selection are spread across the encoded proteins, and include codons within homeodomains (*ARGFX* 3 sites; *DPRX* 1 site; *LEUTX* 8 sites; *TPRX* 4 sites; Online Resource Fig. S7). The spatial locations of sites under positive selection within homeodomains were inferred by comparative modelling of human ETCHbox homeodomains to a known PRD-class structure using Modeller (Šali and Blundell 1993) (Fig. 7, Online Resource Fig. S8). Sites under positive selection include those within the N-terminal arm of ARGFX (E4), LEUTX (Y1, P4, R7) and TPRX1/2 (Q1), and the recognition helix of ARGFX (S43).

Using RELAX (Wertheim et al. 2015), we also find evidence for relaxed selection in all ETCHbox genes compared with their sister gene *CRX* (Online Resource Table S13), suggesting that a combination of relaxed and positive selection is required to explain the fast evolutionary rate of these genes.

Discussion

After duplication from *CRX* in the lineage leading to eutherians, the Eutherian Totipotent Cell Homeobox (ETCHbox) genes underwent asymmetric evolution and continued to be duplicated and lost (Maeso et al. 2016). The genes are suspected to have important roles in preimplantation development and embryonic genome activation (Jouhilahti et al. 2016; Maeso et al. 2016), making observed variability of ETCHbox gene sets a mystery. Here, we compared the ETCHbox complements of 34

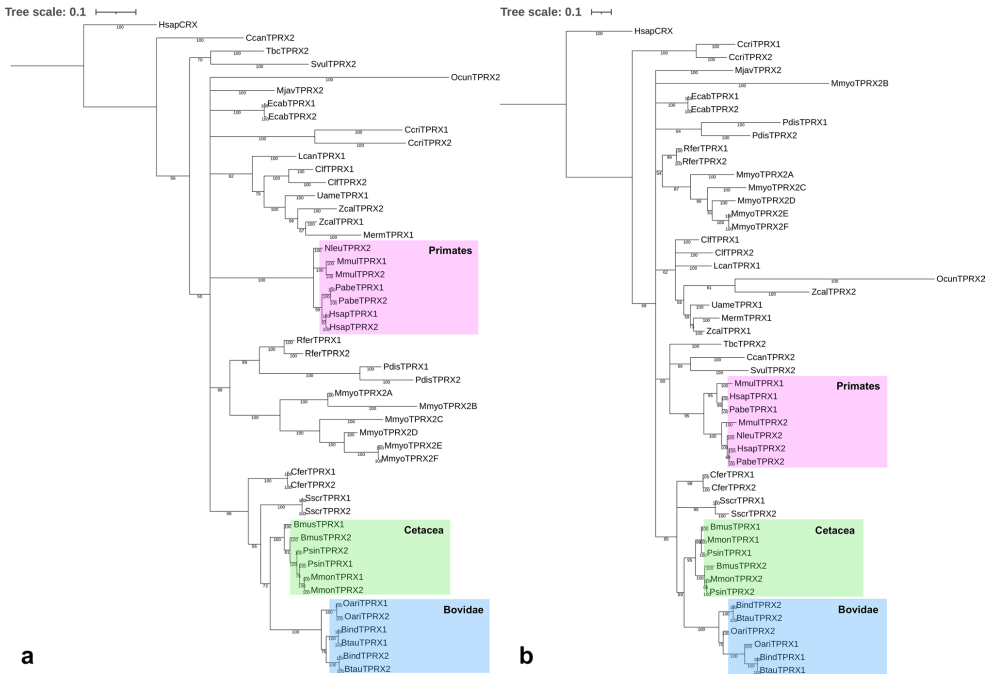


Fig. 6 Bayesian phylogenies inferred using partition 1 (**a**) and partition 2 (**b**) of putatively functional *TPRX1* and *TPRX2* genes split at the gene conversion breakpoint identified by GARD. Boxes highlight the Bovidae, Cetacea and Primates, where topology differs markedly between the two trees. For example, Tree b is consistent with a gene conversion event at the base of the Primates; Tree a has conspecific pairs of *TPRX* genes consistent with additional more recent gene conversion events in the ancestors of these species within the Primates. Species abbreviations as in Fig. 2

species with the aim of illuminating the processes that have sculpted such varied repertoires. Restricting the analyses to genomes sequenced using long-read technologies allowed us to establish with confidence clear examples of gene duplication and secondary loss, something that was challenging in previous work based on lower quality genome assemblies.

We find that, despite extensive and frequent gene loss, all sampled species possess at least two putatively functional ETCHbox genes. This retention suggests that the genes, collectively, are indispensable for eutherian development, and that fluctuations in gene number and rapid sequence evolution are not due to the lack of a function and neutrality. Previous work has shown that some ETCHbox genes can act in an antagonistic fashion, with gene sets upregulated by one gene overlapping with those downregulated by another (Jouhilahti et al. 2016; Maeso et al. 2016). This antagonism could explain why at least two different genes are always retained.

A second line of evidence supporting functionality is that all of the genes have been under recent positive selection, including at residues within the homeodomain. Residues in the N-terminal arm of the ARGFX, LEUTX and TPRX1 homeodomains, identified as having amino acid change driven by positive selection, are suggested by comparative modelling to interact with the minor groove of DNA. Residue 7 specifically, deduced to be under selection in LEUTX, is involved in sequence-specific contact in other homeodomain proteins and therefore may affect binding specificity (Ekker et al. 1994; Damante et al. 1996). ARGFX homeodomain residue S43, also deduced to have been under positive selection, sits within the DNA-binding and specificity-determining recognition helix of PRD-class homeodomains (Bruun et al. 2005). These results suggest that there has been selection for altered DNA-binding properties in ETCHbox homeodomains. In addition, residues in homeodomain helices 1 and 2 are deduced to have been under selection in DPRX, LEUTX and TPRX proteins; modelling suggests

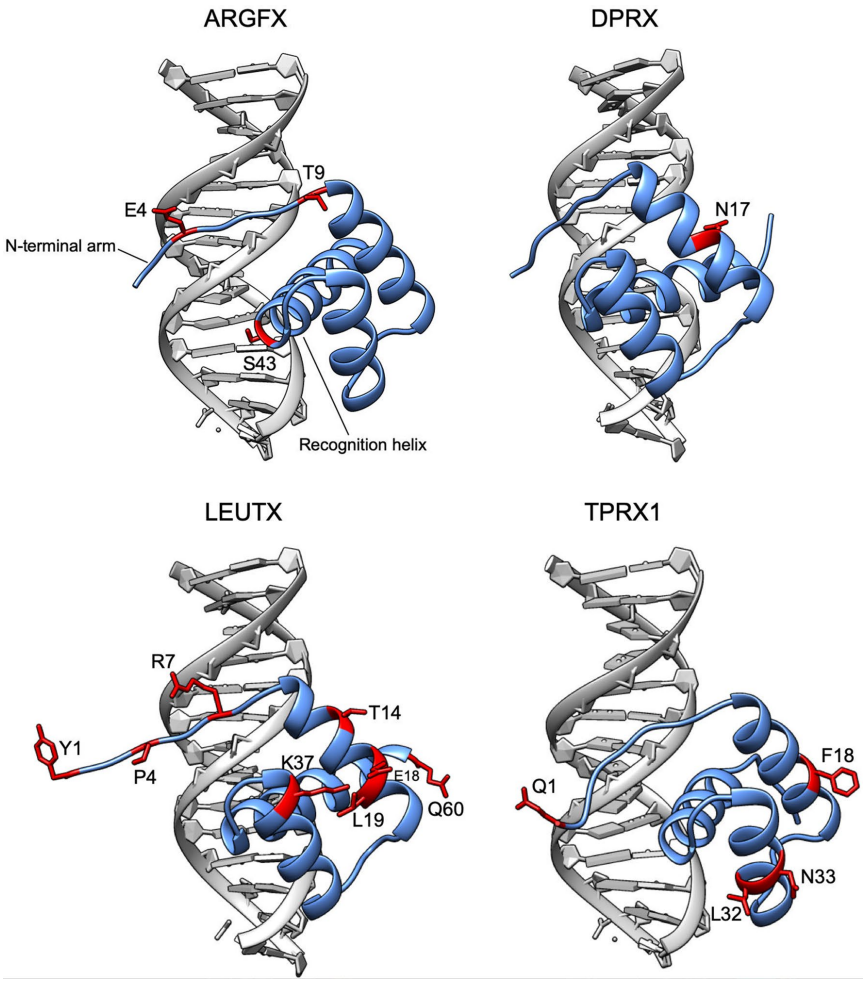


Fig. 7 Models of ETCHbox homeodomain structures including sites under positive selection. Homeodomains of human ETCHbox proteins (blue) are modelled in complex with DNA (grey). Residues under positive selection are coloured red. Amino acid side chains are shown for sites under positive selection only. Letters show the identity of positively selected residues in human, numbers show their position within the homeodomain. TPRX1 and TPRX2 homeodomains are identical due to gene conversion so only one is shown

that most of these residues are on the outer surface of the homeodomain. Since both of these helices have been proposed to mediate protein–protein interactions in some homeodomains, including those of the PRD class (Wilson et al. 1995; Simon et al. 1997; Zaffran and Frasch 2005; Plaza et al. 2008; Altamirano-Torres et al. 2018), we propose that this selection has altered ETCHbox protein–protein binding properties. Madissoon et al. (2016) found that homeodomain differences are not sufficient to explain the differing transcriptional effects of PRD-like genes, implying that other protein domains also contribute to specificity; we therefore suggest that sites outside of the homeodomain that are under selection also influence specificity. Overall, our results suggest that there has been on-going and divergent selection for altered DNA-binding specificity and/or co-factor interactions in the ETCHbox genes, implying that

functions have been modified as part of their rapid evolution during mammalian radiation. Experimental evidence supports these conclusions. For example, Royall et al. (2018) found that, at some point during rodent evolution, *Crxos* (*TPRX1*) likely underwent a change in function to take on part of the role of *ARGFX*.

Though positive selection has contributed to changes in ETCHbox protein sequences, their timing of expression during development has remained relatively stable. Consistent with the results of Maeso et al. (2016), we find that all sampled species possess at least one 'processed pseudogene' derived from an ETCHbox gene; these are generated by retrotransposition exclusively from genes expressed in the germline, including uncommitted early embryonic cells (Vanin 1985; Maestre et al. 1995). This suggests that across large phylogenetic distances the ETCHbox genes retain expression in the very early embryo. This is corroborated by transcriptome data which showed that ETCHbox genes are expressed in preimplantation development in both humans (Euarchontoglires) and cattle (Laurasiatheria) (Maeso et al. 2016).

Despite all eutherian mammals possessing at least two ETCHbox genes, there has been extensive gene loss. We find that of the six ETCHbox genes (*ARGFX*, *DPRX*, *LEUTX*, *PARGFX*, *TPRX1*, *TPRX2*), each has been lost in at least one sampled species, with *PARGFX* lost at the highest rate; furthermore, all sampled species have lost at least one ETCHbox gene. This pattern could be explained through a degree of genetic functional redundancy, whereby multiple genes perform similar functions and can partially substitute for each other, a pattern common after gene duplication (Wagner 1996; Kafri et al. 2009; Zhang 2012). Functional overlap could lead to relaxed selection, allowing repertoires to vary while an overall indispensable function is maintained. This suggestion is consistent with the finding of Maeso et al. (2016) that gene sets regulated by *LEUTX* and *TPRX1* in human cells have a large degree of overlap. Partial redundancy between ETCHbox genes would not be without precedent: it is a common component of biological systems and is known to be a feature of other homeobox duplicates, including members of different *HOX* clusters in mammals (McNulty et al. 2005; Tvrdik and Capecchi 2006; Kafri et al. 2009; Ruff et al. 2015). Genetic redundancy can be evolutionarily stable and may be maintained by selection when, for example, one of the genes occasionally fails to perform a function successfully, or when genes possess other, non-redundant functions which are co-selected with redundant ones (Nowak et al. 1997; Vavouri et al. 2008; Kafri et al. 2009).

Gene duplication is a potential driver of functional innovation. Here we identify large tandem arrays of ETCHbox duplicates in several species, including *O. cuniculus* and *Mi. murinus LEUTX* and *P. leucopus* and *My. myotis TPRX2*. Further arrays have been previously described, such as the 66 Obox (*TPRX2*) loci of *Mu. musculus* (Maeso et al. 2016; Royall et al. 2018). It is likely that the propensity for tandem duplication stems from the position of these genes in a dynamic and unstable genomic region (chromosome 19 in human, 18 in *B. taurus*), in which there is a high density of repetitive sequences, low density of recombination hotspots and elevated gene duplication rates (Castresana 2002; Grimwood et al. 2004; Myers et al. 2005; Maeso et al. 2016), but the selective forces favouring retention of these duplicates are currently unclear. There are three main mechanisms by which duplications could be advantageous in the short term (Innan and Kondrashov 2010): (1) by increasing gene dosage where function is dosage sensitive (Kondrashov and Koonin 2004); (2) by buffering against deleterious mutations (Haldane 1933; Gu et al. 2003); and (3) the immediate emergence of a new function, for example due to the partial duplication of regulatory elements, or alteration of genomic location (Lercher et al. 2003; Lynch and Katju 2004; Katju and Lynch 2006). None of these explanations appears sufficient to explain the giant arrays observed for ETCHbox genes. The alternative is that the initial duplication event is selectively neutral (Innan and Kondrashov 2010), but duplicates are retained following either neofunctionalisation (Ohno 1970) or duplication–degeneration–complementation (DDC) (Force et al. 1999). Current data support this model. The high rates of pseudogenisation in the tandem arrays (45% for *O. cuniculus LEUTX*, 79% for *P. leucopus TPRX2*) suggest that some duplicates are selectively neutral and not actively retained, and previous studies have uncovered functional differences between *Mu. musculus Obox* duplicates, implying that sub- or neofunctionalisation has occurred following expansion of the tandem array (Royall et al. 2018).

Tandem gene duplicates can be subject to gene conversion, and we find overwhelming support that gene conversion has been a major force affecting *TPRX1* and *TPRX2* molecular evolution throughout the Boreoeutheria. Interestingly, these two genes are not directly adjacent to each other, but lie either side of the *CRX* locus. Gene conversion is expected to cause concerted evolution, meaning that instead of gene duplicates accumulating mutations independently they evolve in parallel, maintaining a higher than expected level of sequence similarity (Ohta 1980; Zimmer et al. 1980; Arnheim 1983; Sugino and Innan 2005; Fawcett and Innan 2011). Gene conversion thus restricts the ability of duplicates to neofunctionalise, because their sequence is repeatedly homogenised and divergence is lost (Innan 2003; Teshima and Innan 2008; Fawcett and Innan 2011; Korunes and Noor 2017). As genes diverge, the accumulation of many small mutations or fewer large sequence changes (e.g. transposable element insertion) can cause a threshold to be reached, at which point sequences differ enough that gene conversion no longer occurs; at this stage, independent evolution commences and neofunctionalisation may take place

(Walsh 1987; Teshima and Innan 2008; Fawcett and Innan 2011). The recent gene conversion events and high sequence similarities detected in this work suggest that this threshold is yet to be reached in the *TPRX* genes of most sampled lineages.

It is interesting to consider why *TPRX1* and *TPRX2* seem subject to such frequent gene conversion events, and why this has continued over long time periods across diverse lineages. One possibility is that the genes are dosage-sensitive with a beneficial effect if dosage is increased, as this can cause gene conversion to be favoured by selection (Sugino and Innan 2006). We suggest that gene conversion will affect the strength of selection on *TPRX* genes, whether it be directional or balancing (Fawcett and Innan 2011). For example, gene conversion can lead to faster adaptation because alleles can be transferred between paralogues, enabling the spread of beneficial mutations and elimination of deleterious ones (Winderickx et al. 1993; Chen et al. 2007; Mano and Innan 2008; Korunes and Noor 2017). It is also expected to lead to faster adaptation through increasing effective population size, which enhances the efficiency of selection occurring within a gene family (Mano and Innan 2008). Overall, gene conversion has been a critical factor driving *TPRX1* and *TPRX2* evolution and is predicted to have a dramatic influence on their functional role.

Conclusion

The ETCHbox genes represent an example of the recruitment of eutherian mammal-specific homeobox genes to a very early developmental stage, making them a promising model to study the evolution of young, lineage-specific homeobox genes. Our data show that, unlike the vast majority of homeobox genes, they have been subject to frequent tandem duplications and gene losses over relatively short evolutionary timescales, leading to varied ETCHbox repertoires even amongst closely related species. This includes newly discovered large tandem arrays of homeobox genes. The data also suggest that the ETCHbox genes are indispensable to eutherian preimplantation development, and that positive selection has continued to modify their functions. Finally, we show that gene conversion between *TPRX1* and *TPRX2* has occurred on a striking number of occasions and prevented divergence of their homeodomains; the consequences of this for function are currently unclear. Overall, high rates of gene duplication and loss, extensive divergence, concerted evolution and positive selection have sculpted the varied ETCHbox repertoires that are observed across eutherians; our results support the idea that antagonism and redundancy are key factors in determining these unusual evolutionary patterns.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00239-021-10012-6.

Acknowledgements We thank Yichen Dai, Ignacio Maeso, Peter Mulhair, Rodrigo Pracana, Sebastian Shimeld and Sonia Trigueros for helpful discussions and advice, and Carlos Herrera-Úbeda and Cecy Price for contributions to preliminary work. We also thank two anonymous reviewers for valuable comments that improved the manuscript.

Author Contributions All authors contributed to study conception and design. Material preparation, data collection and analysis were performed by TDL with input from PWHH. TDL and PWHH wrote the manuscript with input from AHR. All authors read and approved the final manuscript.

Funding This work was supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) [Grant Number BB/M011224/1] and an Oxford-Wolfson Marriott BBSRC Graduate Scholarship.

Data Availability The datasets supporting the conclusions of this article are included within the article and its Electronic Supplementary Material.

Declarations

Conflict of interests The authors have no conflict of interests to declare.

Ethical Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr. https://doi.org/10.1109/TAC.1974.1100705

Altamirano-Torres C, Salinas-Hernández JE, Cárdenas-Chávez DL et al (2018) Transcription factor TFIIEβ interacts with two exposed positions in helix 2 of the Antennapedia homeodomain to control homeotic function in Drosophila. PLoS ONE. https://doi.org/10.1371/journal.pone.0205905

Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. https://doi.org/10.1093/nar/25.17.3389

Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics. https://doi.org/10.1017/CBO9780511808999

Arnheim N (1983) Concerted evolution of multigene families. In: Nei M, Koehn RK (eds) Evolution of Genes and Proteins. Sinauer, Sunderland, pp 38–61

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat. https://doi.org/10.1214/aos/1013699998

Berman HM, Westbrook J, Feng Z et al (2000) The protein data bank. Nucleic Acids Res. https://doi.org/10.1093/nar/28.1.235

Bernardo AS, Jouneau A, Marks H et al (2018) Mammalian embryo comparison identifies novel pluripotency genes associated with the naïve or primed state. Biol Open. https://doi.org/10.1242/bio.033282

Blake JA, Ziman MR (2014) Pax genes: regulators of lineage specification and progenitor cell maintenance. Development. https://doi.org/10.1242/dev.091785

Booth HAF, Holland PWH (2007) Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. Gene. https://doi.org/10.1016/j.gene.2006.07.034

Brown SDJ, Collins RA, Boyer S et al (2012) Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Mol Ecol Resour. https://doi.org/10.1111/j.1755-0998.2011.03108.x

Bruun JA, Thomassen EIS, Kristiansen K et al (2005) The third helix of the homeodomain of paired class homeodomain proteins acts as a recognition helix both for DNA and protein interactions. Nucleic Acids Res. https://doi.org/10.1093/nar/gki562

Bürglin TR, Affolter M (2016) Homeodomain proteins: an update. Chromosoma. https://doi.org/10.1007/s00412-015-0543-8

Bürglin TR, Cassata G (2002) Loss and gain of domains during evolution of cut superclass homeobox genes. Int J Dev Biol 46(1):115–123

Burke AC, Nelson CE, Morgan BA, Tabin C (1995) Hox genes and the evolution of vertebrate axial morphology. Development 121(2):333–346

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer, New York

Burset M, Seledtsov IA, Solovyev VV (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. https://doi.org/10.1093/nar/28.21.4364

Castresana J (2002) Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. Nucleic Acids Res. https://doi.org/10.1093/nar/30.8.1751

Chai CL, Zhang Z, Huang FF et al (2008) A genomewide survey of homeobox genes and identification of novel structure of the Hox cluster in the silkworm Bombyx mori. Insect Biochem Mol Biol. https://doi.org/10.1016/j.ibmb.2008.06.008

Chen K, Durand D, Farach-Colton M (2000) NOTUNG: a program for dating gene duplications and optimizing gene family trees. J Comput Biol. https://doi.org/10.1089/106652700750050871

Chen JM, Cooper DN, Chuzhanova N et al (2007) Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet. https://doi.org/10.1038/nrg2193

Cheng WC, Hsiu MHL, Yeh YJ, Li H (2007) Mice lacking the Obox6 homeobox gene undergo normal early embryonic development and are fertile. Dev Dyn. https://doi.org/10.1002/dvdy.21261

Costello I, Nowotschin S, Sun X et al (2015) Lhx1 functions together with Otx2, Foxa2, and Ldb1 to govern anterior mesendoderm, node, and midline development. Genes Dev. https://doi.org/10.1101/gad.268979.115

Dahl E, Koseki H, Balling R (1997) Pax genes and organogenesis. BioEssays. https://doi.org/10.1002/bies.950190905

Damante G, Pellizzari L, Esposito G et al (1996) A molecular code dictates sequence-specific DNA recognition by homeodomains. EMBO J. https://doi.org/10.1002/j.1460-2075.1996.tb00879.x

De Bie T, Cristianini N, Demuth JP, Hahn MW (2006) CAFE: a computational tool for the study of gene family evolution. Bioinformatics. https://doi.org/10.1093/bioinformatics/btl097

Derelle R, Lopez P, Le Guyader H, Manuel M (2007) Homeodomain proteins belong to the ancestral molecular toolkit of eukaryotes. Evol Dev. https://doi.org/10.1111/j.1525-142X.2007.00153.x

Dobin A, Davis CA, Schlesinger F et al (2013) STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. https://doi.org/10.1093/bioinformatics/bts635

Duboule D (1994) Guidebook to the Homeobox Genes. Sambrook & Tooze Publication at Oxford University Press, Oxford

Duboule D (2007) The rise and fall of Hox gene clusters. Development. https://doi.org/10.1242/dev.001065

Dunwell TL, Holland PWH (2017) A sister of NANOG regulates genes expressed in pre-implantation human development. Open Biol. https://doi.org/10.1098/rsob.170027

Durand D, Halldórsson BV, Vernot B (2006) A hybrid micro-macroevolutionary approach to gene tree reconstruction. J Comput Biol. https://doi.org/10.1089/cmb.2006.13.320

Ekker SC, Jackson DG, Von Kassler DP et al (1994) The degree of variation in DNA sequence recognition among four Drosophila homeotic proteins. EMBO J. https://doi.org/10.1002/j.1460-2075.1994.tb06662.x

Estabrook GF, McMorris FR, Meacham CA (1985) Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Syst Zool. https://doi.org/10.2307/sysbio/34.2.193

Fawcett JA, Innan H (2011) Neutral and non-neutral evolution of duplicated genes with gene conversion. Genes (basel). https://doi.org/10.3390/genes2010191

Ferguson L, Marlétaz F, Carter JM et al (2014) Ancient expansion of the Hox cluster in lepidoptera generated four homeobox genes implicated in extra-embryonic tissue formation. PLoS Genet. https://doi.org/10.1371/journal.pgen.1004698

Force A, Lynch M, Pickett FB et al (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics. https://doi.org/10.1093/genetics/151.4.1531

Gouy M, Guindon S, Gascuel O (2010) Sea view version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. https://doi.org/10.1093/molbev/msp259

Graf A, Krebs S, Zakhartchenko V et al (2014) Fine mapping of genome activation in bovine embryos by RNA sequencing. Proc Natl Acad Sci U S A. https://doi.org/10.1073/pnas.1321569111

Grimwood J, Gordon LA, Olsen A et al (2004) The DNA sequence and biology of human chromosome 19. Nature. https://doi.org/10.1038/nature02399

Gruss P, Walther C (1992) Pax in development. Cell. https://doi.org/10.1016/0092-8674(92)90281-G

Gu Z, Steinmetz LM, Gu X et al (2003) Role of duplicate genes in genetic robustness against null mutations. Nature. https://doi.org/10.1038/nature01198

Hahn MW, De Bie T, Stajich JE et al (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res. https://doi.org/10.1101/gr.3567505

Hahn MW, Demuth JP, Han SG (2007) Accelerated rate of gene gain and loss in primates. Genetics. https://doi.org/10.1534/genetics.107.080077

Haldane JBS (1933) The Part Played by Recurrent Mutation in Evolution. Am Nat. https://doi.org/10.1086/280465

Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW (2013) Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol. https://doi.org/10.1093/molbev/mst100

Herr W, Sturm RA, Clerc RG et al (1988) The POU domain: a large conserved region in the mammalian pit-1, oct-1, oct-2, and Caenorhabditis elegans unc-86 gene products. Genes Dev. https://doi.org/10.1101/gad.2.12a.1513

Hobert O, Westphal H (2000) Functions of LIM-homeobox genes. Trends Genet. https://doi.org/10.1016/S0168-9525(99)01883-1

Holland PWH (2013) Evolution of homeobox genes. Wiley Interdiscip Rev Dev Biol. https://doi.org/10.1002/wdev.78

Holland PWH, Booth HAF, Bruford EA (2007) Classification and nomenclature of all human homeobox genes. BMC Biol. https://doi.org/10.1186/1741-7007-5-47

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. https://doi.org/10.1093/bioinformatics/17.8.754

Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika. https://doi.org/10.1093/biomet/76.2.297

Innan H (2003) A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. Proc Natl Acad Sci U S A. https://doi.org/10.1073/pnas.1031592100

Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. https://doi.org/10.1038/nrg2689

Jiang Z, Sun J, Dong H et al (2014) Transcriptional profiles of bovine in vivo pre-implantation development. BMC Genomics. https://doi.org/10.1186/1471-2164-15-756

Jouhilahti EM, Madissoon E, Vesterlund L et al (2016) The human PRD-like homeobox gene LEUTX has a central role in embryo genome activation. Development. https://doi.org/10.1242/dev.134510

Kafri R, Springer M, Pilpel Y (2009) Genetic redundancy: new tricks for old genes. Cell. https://doi.org/10.1016/j.cell.2009.01.027

Katayama S, Ranga V, Jouhilahti EM et al (2018) Phylogenetic and mutational analyses of human LEUTX, a homeobox gene implicated in embryogenesis. Sci Rep. https://doi.org/10.1038/s41598-018-35547-5

Katju V, Lynch M (2006) On the formation of novel genes by duplication in the Caenorhabditis elegans genome. Mol Biol Evol. https://doi.org/10.1093/molbev/msj114

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. https://doi.org/10.1007/BF01731581

Kondrashov FA, Koonin EV (2004) A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. Trends Genet. https://doi.org/10.1016/j.tig.2004.05.001

Korunes KL, Noor MAF (2017) Gene conversion and linkage: effects on genome evolution and speciation. Mol Ecol. https://doi.org/10.1111/mec.13736

Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics. https://doi.org/10.1093/bioinformatics/bti079

Kosakovsky Pond SL, Posada D, Gravenor MB et al (2006a) GARD: a genetic algorithm for recombination detection. Bioinformatics. https://doi.org/10.1093/bioinformatics/btl474

Kosakovsky Pond SL, Posada D, Gravenor MB et al (2006b) Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol. https://doi.org/10.1093/molbev/msl051

Kosakovsky Pond SL, Poon AFY, Velazquez R et al (2020) HyPhy 2.5 - a customizable platform for evolutionary hypothesis testing using phylogenies. Mol Biol Evol. https://doi.org/10.1093/molbev/msz197

Kumar S, Stecher G, Suleski M, Hedges SB (2017) TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol. https://doi.org/10.1093/molbev/msx116

Kumar S, Stecher G, Li M et al (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. https://doi.org/10.1093/molbev/msy096

Lercher MJ, Blumenthal T, Hurst LD (2003) Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Res. https://doi.org/10.1101/gr.553803

Letunic I, Bork P (2019) Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. https://doi.org/10.1093/nar/gkz239

Liu X, Wang Y, Gao Y et al (2018) H3K9 demethylase KDM4E is an epigenetic regulator for bovine embryonic development and a defective factor for nuclear reprogramming. Development. https://doi.org/10.1242/dev.158261

Lu A, Guindon S (2014) Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. Mol Biol Evol. https://doi.org/10.1093/molbev/mst198

Lynch M, Katju V (2004) The altered evolutionary trajectories of gene duplicates. Trends Genet. https://doi.org/10.1016/j.tig.2004.09.001

MacLean JA, Chen MA, Wayne CM et al (2005) Rhox: a new homeobox gene cluster. Cell. https://doi.org/10.1016/j.cell.2004.12.022

Madissoon E, Jouhilahti EM, Vesterlund L et al (2016) Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. Sci Rep. https://doi.org/10.1038/srep28995

Maeda RK, Karch F (2009) The bithorax complex of drosophila. An exceptional Hox cluster. Curr Top Dev Biol. https://doi.org/10.1016/S0070-2153(09)88001-0

Maeso I, Dunwell TL, Wyatt CDR et al (2016) Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals. BMC Biol. https://doi.org/10.1186/s12915-016-0267-0

Maestre J, Tchénio T, Dhellin O, Heidmann T (1995) mRNA retroposition in human cells: processed pseudogene formation. EMBO J. https://doi.org/10.1002/j.1460-2075.1995.tb00324.x

Mallo M, Wellik DM, Deschamps J (2010) Hox genes and regional patterning of the vertebrate body plan. Dev Biol. https://doi.org/10.1016/j.ydbio.2010.04.024

Mano S, Innan H (2008) The evolutionary rate of duplicated genes under concerted evolution. Genetics. https://doi.org/10.1534/genetics.108.087676

Mansai SP, Innan H (2010) The power of the methods for detecting interlocus gene conversion. Genetics. https://doi.org/10.1534/genetics.109.111161

McNulty CL, Peres JN, Bardine N et al (2005) Knockdown of the complete Hox paralogous group 1 leads to dramatic hindbrain and neural crest defects. Development. https://doi.org/10.1242/dev.01872

Miyazono KI, Zhi Y, Takamura Y et al (2010) Cooperative DNA-binding and sequence-recognition mechanism of aristaless and clawless. EMBO J. https://doi.org/10.1038/emboj.2010.53

Morino Y, Hashimoto N, Wada H (2017) Expansion of TALE homeobox genes and the evolution of spiralian development. Nat Ecol Evol. https://doi.org/10.1038/s41559-017-0351-z

Mukherjee K, Bürglin TR (2007) Comprehensive analysis of animal TALE homeobox genes: new conserved motifs and cases of accelerated evolution. J Mol Evol. https://doi.org/10.1007/s00239-006-0023-0
Murrell B, Wertheim JO, Moola S et al (2012) Detecting individual sites subject to episodic diversifying selection. PLoS Genet. https://doi.org/10.1371/journal.pgen.1002764
Murrell B, Weaver S, Smith MD et al (2015) Gene-wide identification of episodic selection. Mol Biol Evol. https://doi.org/10.1093/molbev/msv035
Myers S, Bottolo L, Freeman C et al (2005) Genetics: a fine-scale map of recombination rates and hotspots across the human genome. Science. https://doi.org/10.1126/science.1117196
Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. https://doi.org/10.1016/0022-2836(70)90057-4
Nowak MA, Boerlijst MC, Cooke J, Smith JM (1997) Evolution of genetic redundancy. Nature. https://doi.org/10.1038/40618
Ohno S (1970) Evolution by Gene Duplication. Springer-Verlag, Berlin/New York
Ohta T (1980) Evolution and Variation of Multigene Families. Springer-Verlag, Berlin/New York
Pagès H, Aboyoun P, Gentleman R, DebRoy S (2020) Biostrings: Efficient manipulation of biological strings. In: R Package version 2.57.0
Paps J, Xu F, Zhang G, Holland PWH (2015) Reinforcing the egg-timer: recruitment of novel Lophotrochozoa homeobox genes to early and late development in the Pacific oyster. Genome Biol Evol. https://doi.org/10.1093/gbe/evv018
Pertea M, Pertea GM, Antonescu CM et al (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. https://doi.org/10.1038/nbt.3122
Pettersen EF, Goddard TD, Huang CC et al (2004) UCSF Chimera - a visualization system for exploratory research and analysis. J Comput Chem. https://doi.org/10.1002/jcc.20084
Phillips K, Luisi B (2000) The virtuoso of versatility: POU proteins that flex to fit. J Mol Biol. https://doi.org/10.1006/jmbi.2000.4107
Plaza S, Prince F, Adachi Y et al (2008) Cross-regulatory protein-protein interactions between Hox and Pax transcription factors. Proc Natl Acad Sci U S A. https://doi.org/10.1073/pnas.0806106105
R Core Team (2020) R: A language and environment for statistical computing.
Rajkovic A, Yan C, Yan W et al (2002) Obox, a family of homeobox genes preferentially expressed in germ cells. Genomics. https://doi.org/10.1006/geno.2002.6759
Ritz C, Spiess AN (2008) qpcR: An R package for sigmoidal model selection in quantitative real-time polymerase chain reaction analysis. Bioinformatics. https://doi.org/10.1093/bioinformatics/btn227
Robinson JT, Thorvaldsdóttir H, Winckler W et al (2011) Integrative genomics viewer. Nat Biotechnol. https://doi.org/10.1038/nbt.1754
Ronquist F, Teslenko M, van der Mark P et al (2012) MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol. https://doi.org/10.1093/sysbio/sys029
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. https://doi.org/10.1186/s13227-018-0091-4
Ruff JS, Saffarini RB, Ramoz LL et al (2015) Fitness assays reveal incomplete functional redundancy of the hoxa1 and hoxb1 paralogs of mice. Genetics. https://doi.org/10.1534/genetics.115.178079
Saito K, Abe H, Nakazawa M et al (2010) Cloning of complementary DNAs encoding structurally related homeoproteins from preimplantation mouse embryos: their involvement in the differentiation of embryonic stem cells. Biol Reprod. https://doi.org/10.1095/biolreprod.108.075697
Šali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. https://doi.org/10.1006/jmbi.1993.1626
Sand A, Holt MK, Johansen J et al (2014) TqDist: a library for computing the quartet and triplet distances between binary or general trees. Bioinformatics. https://doi.org/10.1093/bioinformatics/btu157
Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol. https://doi.org/10.1093/oxfordjournals.molbev.a040567
Sheng HZ, Zhadanov AB, Mosinger B et al (1996) Specification of pituitary cell lineages by the LIM homeobox gene Lhx3. Science. https://doi.org/10.1126/science.272.5264.1004
Shriner D, Nickle DC, Jensen MA, Mullins JI (2003) Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet Res. https://doi.org/10.1017/S0016672303006128
Sievers F, Wilm A, Dineen D et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. https://doi.org/10.1038/msb.2011.75
Simon KJ, Grueneberg DA, Gilman M (1997) Protein and DNA contact surfaces that mediate the selective action of the Phox1 homeodomain at the c-fos serum response element. Mol Cell Biol. https://doi.org/10.1128/mcb.17.11.6653
Smith MR (2019) Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. Biol Lett. https://doi.org/10.1098/rsbl.2018.0632
Smith MR (2020) Quartet: comparison of phylogenetic trees using quartet and bipartition measures. Zenodo R Package. https://doi.org/10.5281/zenodo.2536318
Stecher G, Tamura K, Kumar S (2020) Molecular evolutionary genetics analysis (MEGA) for macOS. Mol Biol Evol. https://doi.org/10.1093/molbev/msz312
Sugino RP, Innan H (2005) Estimating the time to the whole-genome duplication and the duration of concerted evolution via gene conversion in yeast. Genetics. https://doi.org/10.1534/genetics.105.043869
Sugino RP, Innan H (2006) Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. Trends Genet. https://doi.org/10.1016/j.tig.2006.09.014
Sugiura N (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. Commun Stat Theory Methods. https://doi.org/10.1080/03610927808827599
Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. https://doi.org/10.1093/nar/gkl315
Tajima F (1993) Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135(2):599–607
Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. https://doi.org/10.1080/10635150701472164
Teshima KM, Innan H (2008) Neofunctionalization of duplicated genes under the pressure of gene conversion. Genetics. https://doi.org/10.1534/genetics.107.082933
Töhönen V, Katayama S, Vesterlund L et al (2015) Novel PRD-like homeodomain transcription factors and retrotransposon elements

in early human development. Nat Commun. https://doi.org/10.1038/ncomms9207
Tvrdik P, Capecchi MR (2006) Reversal of Hox1 Gene Subfunctionalization in the Mouse. Dev Cell. https://doi.org/10.1016/j.devcel.2006.06.016
Vanin EF (1985) Processed pseudogenes: characteristics and evolution. Annu Rev Genet. https://doi.org/10.1146/annurev.ge.19.120185.001345
Vavouri T, Semple JI, Lehner B (2008) Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. Trends Genet. https://doi.org/10.1016/j.tig.2008.08.005
Wagenmakers EJ, Farrell S (2004) AIC model selection using Akaike weights. Psychon Bull Rev. https://doi.org/10.3758/BF03206482
Wagner A (1996) Genetic redundancy caused by gene duplications and its evolution in networks of transcriptional regulators. Biol Cybern. https://doi.org/10.1007/BF00209427
Wagner A (2005) Distributed robustness versus redundancy as causes of mutational robustness. BioEssays 27(2):176–188
Walsh JB (1987) Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? Genetics 117(3):543–557
Warnes GR, Bolker B, Bonebakker L et al (2020) Package "gplots": Various R programming tools for plotting data. R Package Version 3:3
Weaver S, Shank SD, Spielman SJ et al (2018) Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. Mol Biol Evol. https://doi.org/10.1093/molbev/msx335
Wellik DM (2007) Hox patterning of the vertebrate axial skeleton. Dev Dyn. https://doi.org/10.1002/dvdy.21286
Wertheim JO, Murrell B, Smith MD et al (2015) RELAX: detecting relaxed selection in a phylogenetic framework. Mol Biol Evol. https://doi.org/10.1093/molbev/msu400
Wilson DS, Guenther B, Desplan C, Kuriyan J (1995) High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. Cell. https://doi.org/10.1016/0092-8674(95)90468-9
Winderickx J, Battlsti L, Hlbiya Y et al (1993) Haplotype diversity in the human red and green opsin genes: evidence for frequent sequence exchange in exon 3. Hum Mol Genet. https://doi.org/10.1093/hmg/2.9.1413
Xu B, Yang Z (2013) PamlX: a graphical user interface for PAML. Mol Biol Evol. https://doi.org/10.1093/molbev/mst179
Yang Z (1997) Paml: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics. https://doi.org/10.1093/bioinformatics/13.5.555
Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. https://doi.org/10.1093/molbev/msm088
Zaffran S, Frasch M (2005) The homeodomain of Tinman mediates homo- and heterodimerization of NK proteins. Biochem Biophys Res Commun. https://doi.org/10.1016/j.bbrc.2005.06.090
Zhang J (2012) Genetic Redundancies and Their Evolutionary Maintenance. In: Soyer OS (ed) Evolutionary Systems Biology. Springer, New York, New York, NY, pp 279–300
Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol. https://doi.org/10.1089/10665270050081478
Zimmer EA, Martin SL, Beverley SM et al (1980) Rapid duplication and loss of genes coding for the α chains of hemoglobin. Proc Natl Acad Sci U S A. https://doi.org/10.1073/pnas.77.4.2158

PRD-Class Homeobox Genes in Bovine Early Embryos: Function, Evolution, and Overlapping Roles

Thomas D. Lewin [1], Ali A. Fouladi-Nashta [2], and Peter W.H. Holland *[,1]

[1]Department of Zoology, University of Oxford, 11a Mansfield Road, Oxford OX1 3SZ, UK

[2]Comparative Biomedical Sciences Department, Royal Veterinary College, Hawkshead Campus, North Mymms, Hatfield, Hertfordshire AL9 7TA, UK

***Corresponding author:** E-mail: peter.holland@zoo.ox.ac.uk.

Associate editor: Camilla Whittington

Abstract

Eutherian Totipotent Cell Homeobox (ETCHbox) genes are mammalian-specific PRD-class homeobox genes with conserved expression in the preimplantation embryo but fast-evolving and highly divergent sequences. Here, we exploit an ectopic expression approach to examine the role of bovine ETCHbox genes and show that ARGFX and LEUTX homeodomain proteins upregulate genes normally expressed in the blastocyst; the identities of the regulated genes suggest that, in vivo, the ETCHbox genes play a role in coordinating the physical formation of the blastocyst structure. Both genes also downregulate genes expressed earlier during development and genes associated with an undifferentiated cell state, possibly via the JAK/STAT pathway. We find evidence that bovine *ARGFX* and *LEUTX* have overlapping functions, in contrast to their antagonistic roles in humans. Finally, we characterize a mutant bovine *ARGFX* allele which eliminates the homeodomain and show that homozygous mutants are viable. These data support the hypothesis of functional overlap between ETCHbox genes within a species, roles for ETCHbox genes in blastocyst formation and the change of their functions over evolutionary time.

Key words: **homeodomain, blastocyst, Crx, genetic redundancy, protein evolution, gene.**

Introduction

Eutherian Totipotent Cell Homeobox (ETCHbox) genes are a group of highly variable PRD-class homeobox genes specific to mammals. In humans, ETCHbox genes are expressed almost exclusively in the preimplantation embryo between the 4-cell stage and blastocyst, and this temporal expression pattern is broadly conserved in cattle (Madissoon et al. 2016; Maeso et al. 2016). The genes are thought to be involved in key transitions during embryonic development, acting as a vital component of regulatory networks active at the 8-cell stage and in the morula (Jouhilahti et al. 2016; Madissoon et al. 2016; Maeso et al. 2016; Mori et al. 2018; Royall et al. 2018; Taubenschmid-Stowers et al. 2022; Mazid et al. 2022).

The six ancestral ETCHbox genes (*ARGFX*, *DPRX*, *LEUTX*, *PARGFX*, *TPRX1*, and *TPRX2*) duplicated from the Otx-family gene *CRX* on the stem lineage leading to eutherians, and the ancestral eutherian ETCHbox repertoire likely consisted of one copy of each gene (Maeso et al. 2016). A curious property of the genes is their high propensity for duplication and loss, and the subsequent variation in the gene repertoires possessed by different eutherian species (Maeso et al. 2016; Royall et al. 2018; Lewin et al. 2021) (fig. 1). While humans (*Homo sapiens*) have a complement similar to that of the eutherian ancestor, having lost only *PARGFX* and undergone no duplications, the ETCHbox cluster of mice (*Mus musculus*) is very different, with *ARGFX*, *DPRX*, and *PARGFX* completely lost, only remnants of *LEUTX* detectable and *TPRX1* (called *Crxos*) and *TPRX2* (known as *Obox*) both duplicated (Royall et al. 2018). Indeed, *M. musculus* possesses a huge array of 66 *Obox* loci (Royall et al. 2018). The ETCHbox repertoire of European cattle (*Bos taurus*) is more similar to that of human, except that *DPRX* is a pseudogene and there is an extra *TPRX* duplicate (Maeso et al. 2016; Lewin et al. 2021).

In addition to extensive gain and loss, ETCHbox genes exhibit a much faster rate of sequence change across the eutherians than do other homeobox genes (Maeso et al. 2016; Lewin et al. 2021). They are an archetypal example of asymmetric evolution, as their "parent" *CRX* has been conserved while ETCHbox sequences have changed drastically (Maeso et al. 2016; Katayama et al. 2018; Royall et al. 2018; Lewin et al. 2021). These features are unusual for homeobox genes, which are generally highly conserved because they perform key roles in embryonic development and tissue patterning (Bürglin and Affolter 2016). ETCHbox genes, therefore, pose a conundrum: they show extensive evolutionary change in gene number and amino acid sequence, yet they seem to have conserved temporal expression patterns and are hypothesized to possess key roles in the developing embryo. To resolve this paradox we need to know if similar expression is associated with similar and overlapping gene properties, and/or whether conserved expression between species is masking

© The Author(s) 2022. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries, University of Oxford user on 30 January 2023

Article

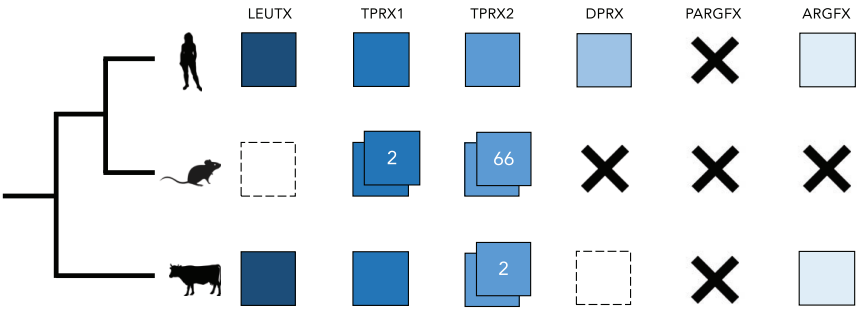


Fig. 1. ETCHbox gene repertoires of human, mouse, and cattle. Colored squares indicate the presence of a functional gene. White squares with dashed borders represent pseudogenes, and a black X represents complete gene loss. Double squares indicate duplication; numbers show the quantity of duplicates of a given gene present in the genome.

changes in gene functions driven by extensive amino acid sequence change.

To address these questions, we focus on *B. taurus* (European cattle) because its repertoire of ETCHbox genes is similar to that of the ancestral eutherian and to that of humans. This allows comparisons to be made between orthologues of *Bos* and human. Furthermore, since humans are within the clade Euarchontoglires and cattle are within Laurasiatheria, this comparison may permit insight into the ancient roles of ETCHbox genes in Eutheria.

There is also intrinsic interest in studying genes that may play regulatory roles in cattle embryonic development. Cattle are enormously important agriculturally: there are an estimated 1–1.5 billion head of cattle worldwide, almost 72 million tonnes of beef and buffalo meat, and 840 million tonnes of milk is produced annually, and demand may increase further if the human population rises as expected (Delgado 2003; FAO 2009, 2019; Ritchie and Roser 2017). It is, therefore, worrying that selective breeding in cattle has sometimes been associated with decreased fertility (Lucy 2001; Sheldon and Dobson 2003; Oltenacu and Algers 2005; Berry et al. 2014), and such infertility is a major cause of cattle culling worldwide (Ahlman et al. 2011; Ansari-Lari et al. 2012; Chiumia et al. 2013; Rilanto et al. 2020). Reduced fertility is caused in large part by a high failure rate of preimplantation embryos, the point at which ETCHbox genes are expressed, with up to 56% of cattle embryos lost before the blastocyst stage (Diskin and Morris 2008). The cost of early embryonic mortality in cattle was estimated at an annual $1.28 trillion worldwide (Perkel et al. 2015) and hence there is a clear imperative to better understand preimplantation development in cattle and identify genes and cellular processes that may be compromised in certain cattle breeds.

In this study, we aimed to examine the function of proteins encoded by ETCHbox genes of cattle, and compare these to their orthologues in humans. The timing and location of ETCHbox gene expression makes them particularly challenging to study in their native setting, and previous work has, therefore, utilized ectopic expression to gain functional insights (Madissoon et al. 2016; Maeso et al. 2016; Royall et al. 2018). This approach, in which the homeobox genes are expressed in cultured cells, asks whether the encoded proteins can elicit changes to cells in which they are not normally expressed. Since homeobox genes encode transcription factors, this can be assayed by analyzing the transcriptome after ectopic expression, a method which has proved highly informative. In this work, we utilize transcriptomic analysis after ectopic expression to understand the function of ETCHbox genes in *B. taurus*, characterize a mutant allele of *B. taurus ARGFX*, and test whether functional roles have changed during mammalian evolution.

Results

Transfection of Homeobox Genes into Bovine Fetal Fibroblasts

In previous work, human ETCHbox gene function was studied by analyzing transcriptomic changes induced by ectopic expression in cultured fibroblasts, in which the genes are not natively expressed (Maeso et al. 2016). To facilitate comparison between orthologous bovine and human ETCHbox genes, here we conducted similar experiments using bovine genes in bovine fetal fibroblast (BFF) cells. First, we established a primary culture of fibroblasts from a biopsy of *B. taurus* fetal skin (supplementary fig. S1, Supplementary Material online). Second, we constructed expression plasmids carrying bovine homeobox genes *ARGFX* or *LEUTX* driven by a constitutive promoter, each connected by a peptide linker to a C-terminal 3xFLAG tag to facilitate downstream analysis. We also constructed an expression plasmid carrying a mutant allele of *ARGFX* identified in the ARS-UCD1.2 reference genome (Lewin et al. 2021). This allele has a 13 base pair (bp) deletion before the homeodomain resulting in a frameshift and truncated predicted protein. Immunocytochemistry

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries of the University of Oxford user on 30 January 2023

showed that bovine ARGFX and LEUTX proteins both localize to the nucleus when transfected into cells, as expected for transcription factors (fig. 2). The mutant ARGFX protein was detected in the nucleus but was also abundant in the cytoplasm, suggesting the presence of a weaker nuclear localization signal. Control transfections were negative.

Transcriptomic Changes Induced by Bovine Homeobox Genes

We used transcriptomic analysis to determine whether ectopic expression of bovine *ARGFX* and *LEUTX* elicited gene expression changes in cells (supplementary table S1, Supplementary Material online). After transfection, RNA-sequencing (RNA-seq) followed by differentially expressed (DE) gene analysis identified which loci in the genome increased or decreased in expression level (fig. 3*A* and *B*; supplementary table S2*a–f*, Supplementary Material online). We found 518 and 725 genes up- and downregulated, respectively, following *ARGFX* ectopic expression, which likely includes direct and indirect targets. *LEUTX* expression caused changes to a larger number of genes: 1,286 increased in expression level and 1,238 decreased. *ARGFX* expression downregulates more genes than it upregulates, and the genes that decreased in expression after *ARGFX* transfection had a greater average fold change and lower *P*-value than those that increased in expression (fig. 3*A*); this suggests that ARGFX primarily has a repressive role in this cellular context. We performed gene ontology (GO) analysis on the DE gene sets (supplementary table S3*a–d*, Supplementary Material online).

We find that 47% of genes upregulated following *ARGFX* ectopic expression are also upregulated after *LEUTX* expression; 61% of genes downregulated by *ARGFX* expression are also downregulated by *LEUTX*. These overlaps between genes upregulated after *ARGFX* and *LEUTX* transfection (Fisher's exact test $P = 1.492 \times 10^{-90}$) or downregulated in response to both genes ($P = 5.176 \times 10^{-251}$) are much greater than overlaps for genes affected antagonistically by *ARGFX* and *LEUTX*, for which there are no significant overlaps (6% genes upregulated in response to *ARGFX* are downregulated in response to *LEUTX*, $P = 1.000$; 3% genes downregulated in response to *ARGFX* are upregulated in response to *LEUTX*; $P = 1.000$) (fig. 3*C*). These analyses suggest a large degree of overlap in the cellular roles of bovine ARGFX and bovine LEUTX: the proteins perform similar functions rather than acting

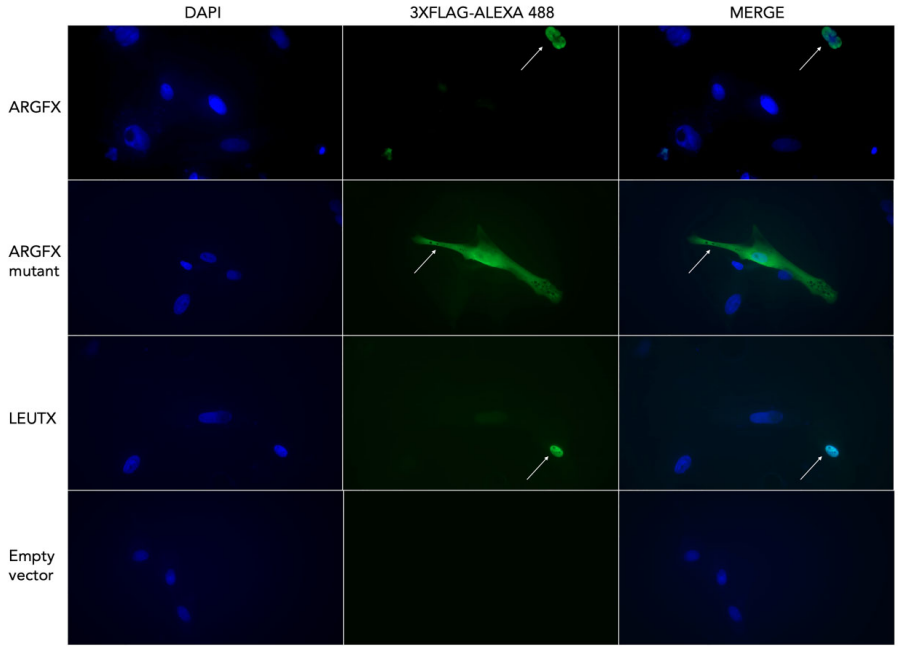


Fig. 2. Immunocytochemistry analysis of BFF cells transfected with bovine ETCHbox genes. White arrows indicate cells with high expression of ETCHbox proteins. DNA is stained with DAPI (blue) to show cell nuclei. ETCHbox proteins are detected using Alexa Fluor 488-labelled antibodies to anti-FLAG antibodies (green). WT ARGFX and LEUTX proteins are localized to the nucleus; the mutant ARGFX is detected in the nucleus and cytoplasm. No green fluorescence is detected in samples transfected with the control plasmid.

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries of the University of Oxford user on 30 January 2023

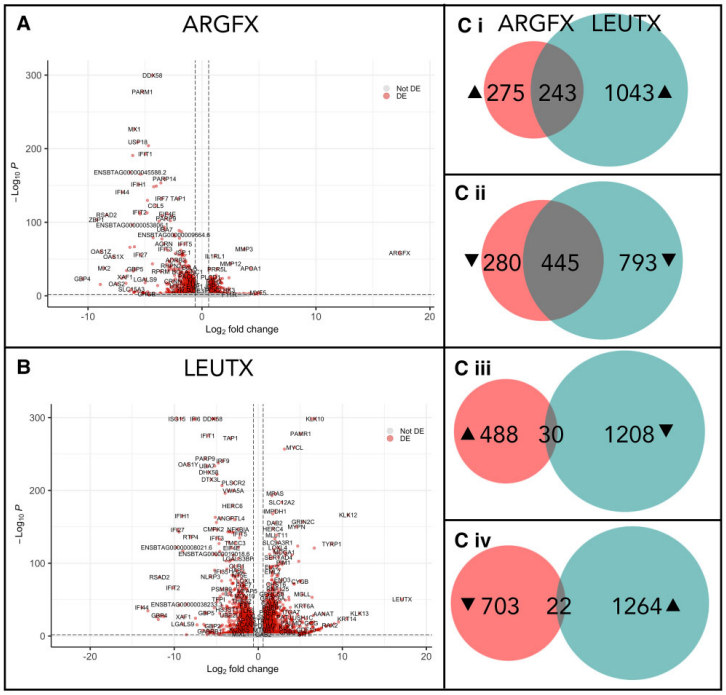


FIG. 3. Transcriptional response to ETCHbox gene ectopic expression. Volcano plots showing genes DE in response to wild type (WT) (*A*) *ARGFX* and (*B*) *LEUTX* ectopic expression. Each point represents a gene. Points in red are considered DE (adjusted $P < 0.05$, fold change [FC] > 1.5). (*C*) Overlaps between sets of genes DE in response to *ARGFX* and *LEUTX* ectopic expression: (i) *ARGFX* (red) up, *LEUTX* (blue) up; (ii) *ARGFX* down, *LEUTX* down; (iii) *ARGFX* up, *LEUTX* down; and (iv) *ARGFX* down, *LEUTX* up.

antagonistically. This is in contrast to data from humans, in which ARGFX and LEUTX act in opposite directions on the same group of genes (Maeso et al. 2016).

Properties of Bovine, Human, and Mouse Homeobox Genes

The comparisons above suggest that bovine and human ETCHbox genes may not be performing identical roles. To investigate further, we compiled lists of genes DE in response to ETCHbox expression in three systems: bovine genes in bovine fetal dermal fibroblasts (this work), human genes in human dermal fibroblasts (Maeso et al. 2016), and mouse genes in mouse embryonic fibroblasts (Royall et al. 2018). We restricted lists of downstream genes to 1:1 orthologues; our dataset includes 21,666 bovine genes, of which 71.7% (15,534) had identifiable 1:1 orthologues in human and 71.6% (15,503) in mouse. Fisher's exact test was then used to examine overlaps: this analysis asks whether overlap in the function of two genes is significantly greater than expected by chance. Indeed, many of the tests give apparently significant overlaps, hence we compared negative logs of adjusted *P*-values, which we call "overlap scores," to extract biologically meaningful information: the higher the overlap score, the more the overlap exceeds the expected similarity. To provide context, previous work suggests that in humans ARGFX acts antagonistically to LEUTX and TPRX1, which have overlapping roles (Maeso et al. 2016).

Considering genes DE in response to bovine *LEUTX* expression, the highest overlap scores are for overlaps between: bovine *LEUTX* down ('down' referring to downstream genes that decrease in expression) genes and human *LEUTX* down genes (overlap score = 62.032), bovine *LEUTX* down genes and human *TPRX1* down genes (overlap score = 56.611), and then bovine *LEUTX* up ('up' referring to downstream genes that increase in expression) genes and human *LEUTX* up genes (overlap score = 26.833) (fig. 4A, points marked "1"). For example, 170 of the 1,019 bovine *LEUTX*-downregulated genes remaining after restriction to 1:1 orthologues are also downregulated by human *LEUTX*. Our test seeks to ask if this number is greater than expected by chance, and we conclude that it is (Fisher's exact test P value $= 9.28 \times 10^{-63}$); from this, we deduce some conservation of function. Overall, these data suggest that *at least some aspects of the function of LEUTX are similar between humans and cattle*. Bovine *LEUTX*-downregulated genes also have a large overlap (overlap score = 29.979) with human *ARGFX* upregulated genes (fig. 4A, point marked "2").

Considering genes downstream of bovine *ARGFX*, we find minimal overlap with gene sets affected by human *ARGFX*, suggesting *bovine and human ARGFX have different properties* (fig. 4A, points marked "3"). Larger overlaps,

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries of the University of Oxford user on 30 January 2023

indeed the most significant overlaps in the whole study, are found when bovine *ARGFX* is compared with human *LEUTX* or *TPRX1*: bovine *ARGFX* down genes and human *LEUTX* down genes (overlap score = 70.383); bovine *ARGFX* down genes and human *TPRX1* down genes (overlap score = 68.166) (fig. 4A, points marked "4"). This suggests *that ARGFX in cattle has similar properties to LEUTX and TPRX1 in humans*. There are also large overlaps between bovine and human *ARGFX* in opposite directions: bovine *ARGFX* down genes and human *ARGFX* up genes (overlap score = 21.460); bovine *ARGFX* up genes and human *ARGFX* down genes (overlap score = 10.320) (fig. 4A, points marked "5"). While this final comparison has a smaller overlap score, it is still the highest overlap for bovine *ARGFX* upregulated genes with any human gene. This suggests that *bovine ARGFX acts in the opposite direction to human ARGFX*.

Comparing bovine and murine genes, the highest similarity of lists of DE genes is between bovine *LEUTX* down and mouse *Crxos*, *OboxA1*, *Oboxa4*, and *Oboxa7* down genes, but these similarities are smaller than bovine/human overlaps (overlap scores = 22.113, 23.008, 27.870, and 26.094, respectively) (fig. 4B, points marked "1"). This suggests that bovine ETCHbox genes are more similar in function to human ETCHbox genes than to those of mice. The highest overlap score for bovine *ARGFX* up genes is murine *Oboxa4* up genes (overlap score = 18.860) and for bovine *ARGFX* down genes is murine *Oboxa4* (overlap score = 16.250) and *Oboxa7* (overlap score = 17.984) down genes (fig. 4B, points marked "2"). These data suggest that both bovine ARGFX and LEUTX act in broadly the same direction as the tested mouse ETCHbox genes.

Relevance of Target Genes to Bovine Development

The bovine *ARGFX* and *LEUTX* genes are expressed in a tight temporal window during preimplantation embryonic development (Maeso et al. 2016). To investigate if the gene expression changes induced by ectopic expression in fibroblasts have functional relevance, we employed temporal expression clustering. This approach tests whether the set of downstream target genes in fibroblast cells are enriched for genes normally expressed in specific temporal patterns in the embryo. Using published transcriptomic data, we determined temporal expression profiles for all bovine genes expressed from oocyte to blastocyst and clustered these into 100 groups sharing similar temporal expression profiles. We then tested whether the gene sets DE in response to *ARGFX* or *LEUTX* expression were enriched for genes in any expression profiles (supplementary table S4, Supplementary Material online).

Genes that increased in expression after ectopic expression of both *ARGFX* and *LEUTX* are enriched for profile 48 (Fisher's exact test adjusted $P = 0.005$ and $P = 0.041$, respectively), which comprises genes that show minimal expression before the blastocyst but are then strongly expressed at the blastocyst stage (fig. 5A; supplementary fig. S2, Supplementary Material online). This supports the above conclusion that ARGFX and LEUTX functions are overlapping. The gene set downregulated by ectopic expression of *LEUTX* is enriched for genes belonging to expression profile 45 (Fisher's exact test adjusted $P = 0.005$). This profile comprises genes expressed specifically at the two-cell stage (fig. 5B). Finally, profile 36 is enriched in the gene set downregulated following *ARGFX* ectopic expression (Fisher's exact test adjusted $P = 0.011$); these genes are activated specifically at the 8-cell stage (fig. 5C).

We note that profile 48, upregulated by *ARGFX* and *LEUTX*, includes several genes with known roles in cell junctions and cell adhesion (*DSC2*, *EPDR1*, *FLRT2*), the cytoskeleton (*ACTG2*, *EPS8L2*, *HCK*, *MIB2*, *MYADM*, *PALLD*, *PLD2*, *SLC9A3R1*, *SMTN*), or the extracellular matrix (*COL15A1*, *COL5A2*, *COL5A3*, *COL6A3*, *CTHRC1*, *PXDN*) (supplementary fig. S2, Supplementary Material online), functions which are central to embryo compaction and the physical formation of the blastocyst. This suggests that bovine *ARGFX* and *LEUTX* regulate aspects of the structural formation of the blastocyst. Moreover, GO cellular component analysis of the shared gene set upregulated by *LEUTX* in both cattle and in humans (supplementary table S5, Supplementary Material online) shows extracellular matrix (GO:0031012) as an enriched function ($P = 7.17 \times 10^{-4}$), suggesting that this functional role may be conserved between the two species.

ETCHbox Genes and Cell Potency

Induced pluripotent stem cells can be generated from mouse and human somatic cells by the addition of just four transcription factors: MYC, KLF4, POU5F1, and SOX2 (Takahashi and Yamanaka 2006; Takahashi et al. 2007). In this work, we find that bovine *ARGFX* and *LEUTX* both downregulate *MYC*, and *ARGFX* also downregulates *KLF4*. Inspecting DE gene lists from published works, we find that human *DPRX*, *LEUTX*, and *TPRX1* all downregulate *KLF4* and that four tested mouse ETCHbox genes (*Crxos*, *Oboxa1*, *Oboxa4*, *Oboxa7*) also downregulated *Sox2* when expressed in fibroblasts (Maeso et al. 2016; Royall et al. 2018). Furthermore, bovine *ARGFX* and *LEUTX* both downregulate the pluripotency-related gene *GBP4* and, indeed, the total bovine *LEUTX*-downregulated gene set is enriched for genes defining an "undifferentiation network signature," which is composed of genes that are considered markers of an undifferentiated cell state (Galan et al. 2013) (Fisher's exact test, $P = 0.002$; supplementary table S6, Supplementary Material online). Overall, both bovine *ARGFX* and *LEUTX* downregulate genes linked with an undifferentiated cell state.

Frequency and Properties of a Mutant Bovine *ARGFX* Allele

We previously identified a 13 bp deletion in the coding region of the *ARGFX* gene of the *B. taurus* reference genome ARS-UCD1.2 and showed that the indel is polymorphic, as several RNA-seq datasets exhibit the wild type (WT) allele

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries of the University of Oxford user on 30 January 2023

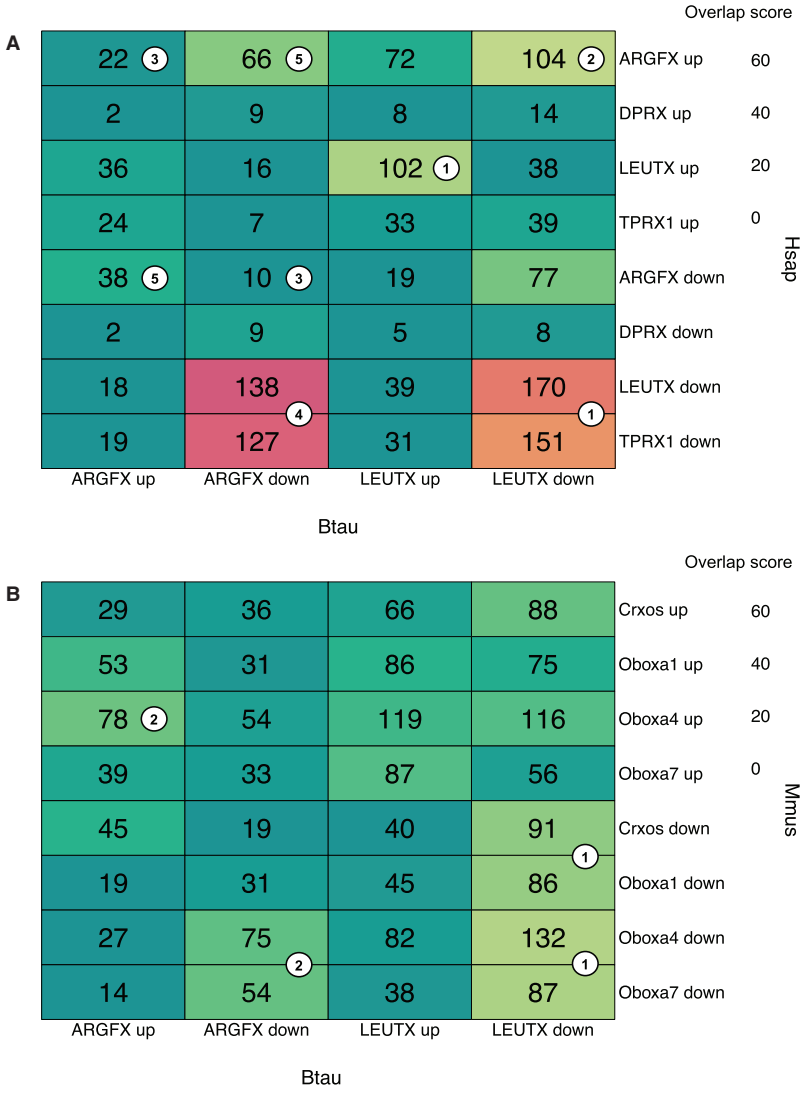


Fig. 4. Overlap between genes DE in response to ectopic expression of (*A*) bovine and human and (*B*) bovine and mouse ETCHbox genes. For each comparison, numbers in the heatmap show the number of shared DE genes. Heatmap color scale reflects overlap scores. The greater the overlap score, the more the overlap of regulated genes exceeds that which is expected by chance. Overlap scores consist of $-\log_{10}$ multiple comparison-adjusted *P*-values from Fisher's exact test for overlaps between gene sets. White dots with numbers mark points referred to in the main text. Btau, *Bos taurus*; Hsap, *Homo sapiens*; Mmus, *Mus musculus*.

(Lewin et al. 2021). Here, we investigate the frequency of the deletion allele and its effect on protein function. The mutation, a 13 bp deletion in exon 2, causes a frameshift that eliminates the homeodomain and truncates the protein (fig. 6A). We used the Bovine Genome Variation Database (Chen et al. 2020), which includes data from 432 *B. taurus* and *B. indicus* samples from 54 breeds, to assess the distribution of the two alleles. We find 25 of 54 (46%) breeds have at least one sampled individual with the mutant allele, and the limited sampling of some breeds

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries of the University of Oxford user on 30 January 2023

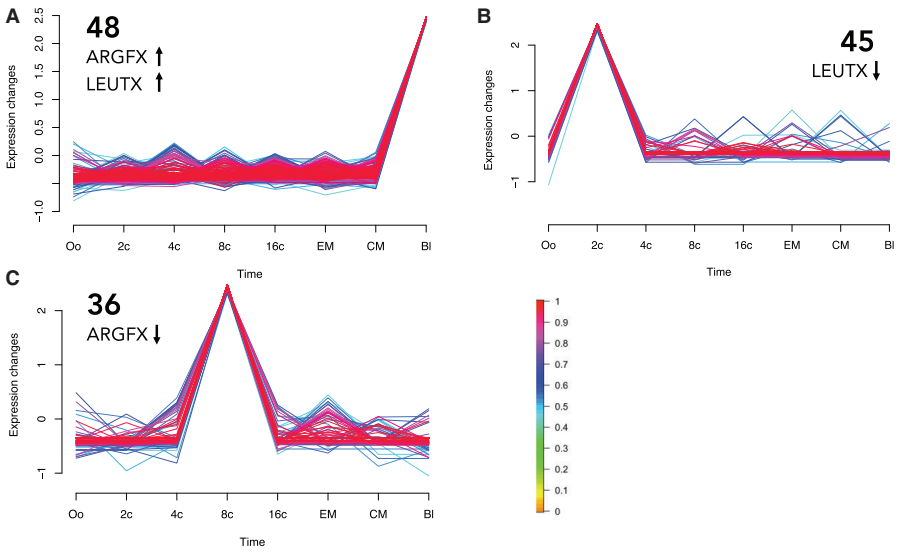


Fig. 5. Temporal expression profiles for bovine genes enriched in the downstream target gene sets of ARGFX and LEUTX. Plots show normalized, standardized expression level as a function of developmental stage. Each line is the expression profile of one gene. Colors of lines show membership score from 0 to 1; membership score is a measure of how well the gene fits the profile. Genes with a score <0.5 were excluded from profiles. (*A*) Profile 48, *ARGFX* up and *LEUTX* up. (*B*) Profile 45, *LEUTX* down. (*C*) Profile 36, *ARGFX* down. Oo, oocyte; 2c, 2-cell; 4c, 4-cell; 8c, 8-cell; 16c, 16-cell; EM, early morula; CM, compact morula; Bl, early blastocyst.

suggests that the incidence may be higher. The reported mutant allele frequency is 0.122 across all samples and is highest in Jersey (0.917, 12 samples) and Hereford cattle (0.667, 21 samples) (fig. 6*B*; supplementary table S7, Supplementary Material online). Consistent with this, the reference genome ARS-UCD1.2 is from a Hereford individual. The deletion is present at least in low frequency in breeds originating from Africa, Asia, and Europe, and is found in both *B. taurus* and *B. indicus*. No other indels in the coding sequence of *ARGFX* are identified in the 432 samples.

To test whether individuals homozygous for the mutant *ARGFX* allele can be viable, we used polymerase chain reaction (PCR) to screen samples from the Jersey breed. From 11 individuals, we found three that are homozygous for the mutant allele, demonstrating that the homozygous mutant is not lethal (fig. 6*C*). Five individuals were heterozygous, two were homozygous WT, and one sample was inconclusive. We sequenced the *ARGFX* gene of all three homozygous mutant samples to confirm that the deletion found by PCR in Jersey cattle is the same 13 bp deletion identified in the reference genome. We also used PCR to genotype the BFFs used for ectopic expression; these were homozygous for the WT *ARGFX* allele (fig. 6*C*).

Analysis of RNA-seq data (Graf et al. 2014) from embryo pools containing both alleles shows that *ARGFX* continues to be expressed from a mutant locus. To understand the impact on protein function, we generated an expression construct containing the mutant *ARGFX* allele, transfected this into bovine fibroblasts, conducted transcriptome sequencing, and compared the sets of genes that changed in expression in response to transfection of WT or mutant *ARGFX* (fig. 6*D*). We found that the mutant allele has a much smaller set of downstream effects than does WT *ARGFX*, but that the DE genes overlapped strongly with those regulated by WT *ARGFX*. Just 18% and 41% of genes up- and downregulated by WT *ARGFX*, respectively, were also regulated by the mutant *ARGFX* but 89% of the genes that increased in expression and 92% of those that decreased in expression after transfection of mutant *ARGFX* had the same response to transfection of WT *ARGFX* (fig. 6*E*). This suggests that the truncated mutant protein is able to perform a subset of the functions of WT ARGFX, but not additional functions. This retention of a limited function despite protein truncation may be facilitated by cofactors that bind to the intact part of the protein.

Discussion

The ETCHbox genes are a group of eutherian-specific homeobox genes that exhibit several puzzling characteristics. They are homeobox genes with a highly specific, conserved temporal expression pattern in the preimplantation

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries of the University of Oxford user on 30 January 2023

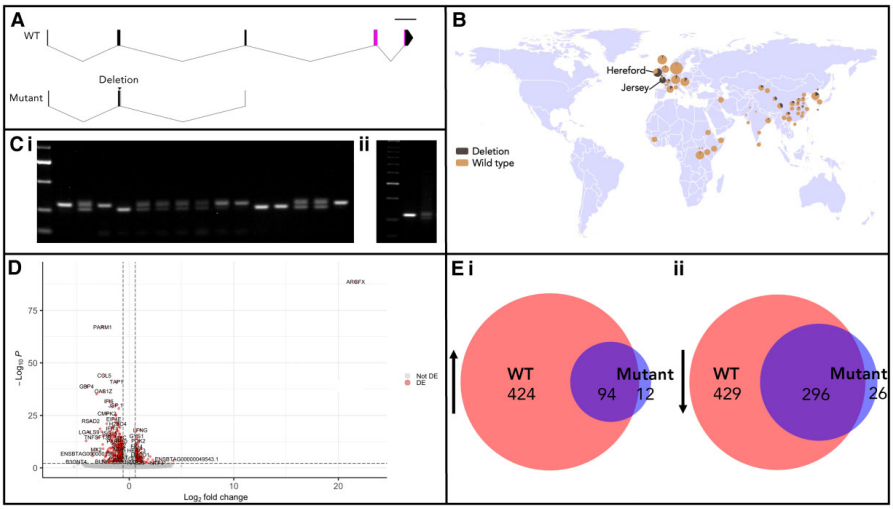


Fig. 6. Characterization of a mutant bovine *ARGFX* allele. (*A*) Structure of WT and mutant genes and deduced proteins. The WT gene comprises five coding exons, with the homeobox (pink) split across exons 4 and 5. The mutant allele has a 13 bp deletion in exon 2 causing a frameshift (gray) and an early stop codon truncating the protein in exon 3 before the homeodomain. Scale bar 1 kb. (*B*) Map showing the distribution of the mutant allele. Dark brown/black = deletion, pale brown = WT. Pie charts show allele frequency for each breed; pie size proportional to sample size. Pie charts are located at the approximate location of the breed's origin. Map generated using Bovine Genome Variation Database (Chen et al. 2020). (*Ci*) Agarose gel shows that individuals can be homozygous for the mutant *ARGFX* allele. PCR primers were used that anneal either side of the deletion site, producing a shorter amplicon in the deletion allele. Lane 1 = Bioline HyperLadder 50 bp. Lanes 2 and 16 = homozygous WT control. Lanes 3 and 15 = heterozygous control. Lanes 4 and 11 are homozygous WT Jersey individuals, lanes 5, 12, and 13 are homozygous mutant Jersey individuals and lanes 6, 7, 8, 9, and 14 are heterozygous Jersey individuals. Lane 10 is inconclusive, with a strong WT band and very weak mutant band. (*Cii*) Agarose gel showing that BFF cells have a WT *ARGFX* gene. Lane 1 = Bioline HyperLadder 50 bp. Lane 2 = BFF band. Lane 3 = heterozygote bands. (*D*) Volcano plot showing genes DE in response to *ARGFX* mutant ectopic expression. Each point represents a gene. Points in red are considered DE (adjusted $P < 0.05$, fold change [FC] > 1.5). (*E*) Overlaps between sets of genes DE in response to WT or mutant *ARGFX* ectopic expression: (i) upregulated by both and (ii) downregulated by both.

embryo, and which are inferred to play roles at key developmental milestones such as embryonic genome activation (Jouhilahti et al. 2016; Madissoon et al. 2016; Maeso et al. 2016). Given this, one might expect them to be highly conserved but, in reality, the genes display rapid sequence evolution, and are also duplicated and lost at an unusually high rate (Maeso et al. 2016; Royall et al. 2018; Lewin et al. 2021).

We utilized ectopic expression to examine the roles of ETCHbox genes in cattle, and find evidence for large functional overlap between bovine *ARGFX* and *LEUTX*. This is consistent with the hypothesis that there is partial redundancy between ETCHbox proteins, and that this is one of the causes of the genes' unusual evolutionary patterns (Lewin et al. 2021). Such redundancy may also explain why a frameshift mutation of *ARGFX* is not lethal when homozygous and has risen to a population allele frequency of over 10%; however, we stress that we do not know if homozygous mutant embryos have the same rate of survival as those with a fully functional copy of *ARGFX*. Regardless, there would be reduced selection against a suboptimal allele if gene function in the preimplantation embryo is redundant with *LEUTX*. Indeed, of the 424 ARGFX up genes and 429 ARGFX down genes which the mutant ARGFX protein fails to properly regulate, 45% and 48%, respectively, are also regulated in the same direction by LEUTX. This provides further evidence that LEUTX is able to directly compensate for the presence of a truncated ARGFX protein. Such redundancy may underpin why ETCHbox genes are lost in evolution at an elevated rate and, in this example, we may be observing this process in action: redundancy between two ETCHbox genes allowing one to move toward pseudogenization.

To test whether ETCHbox gene sequence divergence is accompanied by functional change, we compared the transcriptional response to ectopic expression between humans and cattle. Human and bovine ETCHbox orthologues share 48–77% amino acid sequence identity in their homeodomains and 36–54% identity across the full protein. We were surprised to find that *B. taurus* and *H. sapiens ARGFX* genes show minimal functional overlap, and that *B. taurus ARGFX* likely performs a function similar to that of *H. sapiens LEUTX*, and opposite to that of *H. sapiens ARGFX*. The change in broad functional roles is not seen, however, for LEUTX proteins, which have highly similar effects in human and bovine cells. We also found that

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries of the University of Oxford user on 30 January 2023

bovine ETCHbox genes regulate gene sets that are more similar to those regulated by human ETCHbox genes than murine ETCHbox genes; this is consistent with the greater similarity of bovine ETCHbox sequences to those of humans than those of mice (Royall et al. 2018). In interpreting these results, we note the caveat that the human and cattle experiments were similar but not identical: analysis of human ETCHbox genes used ectopic expression in adult cells and analyzed transcriptome change 48 h post-transfection (Maeso et al. 2016), the bovine experiment used fetal cells, with RNA extracted after 72 h. These differences might cause variation in the gene sets regulated. Overall, however, these data suggest that significant changes in functional roles are a feature of ETCHbox evolution.

We utilized temporal expression profiling of bovine embryonically expressed genes to gain insights into the function of ETCHbox genes in the preimplantation embryo. The logic to this analysis is that if, during normal embryonic development, bovine ETCHbox genes upregulate or downregulate suites of genes at a particular stage, this may be replicated when the genes are ectopically expressed in cell culture. We find one temporal profile enriched in both the *ARGFX* and *LEUTX* upregulated gene sets: this profile consists of genes that have minimal expression up until the blastocyst but are then strongly induced. We suggest, therefore, that bovine *ARGFX* and *LEUTX* are involved in blastocyst formation. Interestingly, many blastocyst genes that are upregulated by *ARGFX* and/or *LEUTX* possess known roles in cell junctions and cell adhesion, the cytoskeleton or the extracellular matrix (supplementary fig. S2, Supplementary Material online). These functions are critical to embryo compaction and the structural changes that occur during morula and blastocyst formation (Albertini et al. 1987; Vestweber et al. 1987; Lehtonen et al. 1988; Enders et al. 1990; Damsky et al. 1993; Koyama et al. 1994; Hardy et al. 1996; Fleming et al. 2000; Fleming 2001; Watson and Barcroft 2001; Eckert and Fleming 2008; Hasley et al. 2017; Aberkane et al. 2018; Lim and Plachta 2021); this suggests that ETCHbox genes are involved in coordinating the physical process of structural formation of the blastocyst. We note, however, that, despite GO analyses suggesting that this role may be conserved in humans, a similar response of "blastocyst genes" was not detected in experiments using human cells (Maeso et al. 2016). One possible explanation for the difference is that blastocyst gene expression could be a downstream effect, not observed in the human experiments when transfected cells were harvested sooner (Maeso et al. 2016). We do not find induction of the proinflammatory genes *S100A9* or *TNFA*, implicated in inducing an inflammation response in uterine cells at implantation (He et al. 2019).

The timing of ETCHbox genes' expression in cattle and human, peaking just before blastocyst formation, and for *ARGFX* continuing into the blastocyst (Maeso et al. 2016), is consistent with the hypothesis that bovine and human ETCHbox genes are involved in coordinating the physical process of blastocyst formation. A function for ETCHbox genes in embryo compaction and the structural formation of the blastocyst is also consistent with their absence in the sister lineage to placentals, marsupials, which do not undergo compaction, possess a morula stage or form an inner cell mass as part of their blastocyst (Renfree and Lewis 1996; Renfree 2010; Renfree and Shaw 2014; Royall et al. 2019). The upregulation of the same temporal profile by both ARGFX and LEUTX supports the contention that bovine ARGFX and LEUTX have overlapping functions.

The bovine *LEUTX*-downregulated gene set is enriched for genes expressed at the 2-cell stage, suggesting that *LEUTX* may downregulate the 2-cell program as part of the exit from the 2-cell stage. Intriguingly, *ARGFX* downregulates a temporal profile comprising genes expressed at the 8-cell stage. This profile is similar to that identified in previous work (Maeso et al. 2016) as downregulated by ectopic expression of human *LEUTX* and *TPRX1*, and upregulated by *ARGFX* (fig. 7); this supports our earlier conclusion that *B. taurus ARGFX* performs the opposite role to human *ARGFX* and is more similar functionally to human *LEUTX*. Overall, the temporal expression profiles enriched in the gene sets regulated by ectopic expression of ETCHbox genes are consistent with the timing of ETCHbox expression in cattle development, indicate that the ectopic expression experiments lead to the activation of realistic embryonic targets, and suggest that ETCHbox genes are involved in blastocyst formation.

We find that human and bovine ETCHbox genes downregulate the expression of Yamanaka factors, and that the total bovine *LEUTX*-downregulated gene set is enriched for genes of the "undifferentiation network signature" of the early preimplantation embryo (Galan et al. 2013). Cells of the preimplantation embryo undergo a reduction in their potency as the totipotent blastomeres of early development undergo lineage specification and differentiation to form the blastocyst (Lu and Zhang 2015; Hu 2019); the downregulation of genes associated with an undifferentiated state is, therefore, consistent with the timing of ETCHbox genes' expression just prior to blastocyst formation.

Finally, we also examined which gene targets had the strongest response to ETCHbox expression. Eight of the ten downregulated genes with the highest fold change were common between *ARGFX* and *LEUTX*; seven of these eight genes (*GBP4*, *IFI44*, *MX2*, *OAS1X*, *OAS1Z*, *RSAD2* [*VIPERIN*], *ZBP1*) are known to be interferon-stimulated genes (ISGs), which are regulated via the JAK/STAT signaling pathway (Indraccolo et al. 2007; DeFilippis et al. 2010; Hu et al. 2011; Seo et al. 2011; Kane et al. 2013; Schneider et al. 2014; Vladimer et al. 2014; Hernandez et al. 2015; Harwardt et al. 2016; DeDiego et al. 2019; Kuepper et al. 2019; Tirumurugaan et al. 2020). Two more ISGs, *XAF1* and *IFIT2*, are in the top 10 downregulated genes for *ARGFX* and *LEUTX*, respectively (Sun et al. 2008; Zhou et al. 2013). Moreover, the overall gene sets downregulated in response to *ARGFX* and *LEUTX* expression are each

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries of the University of Oxford user on 30 January 2023

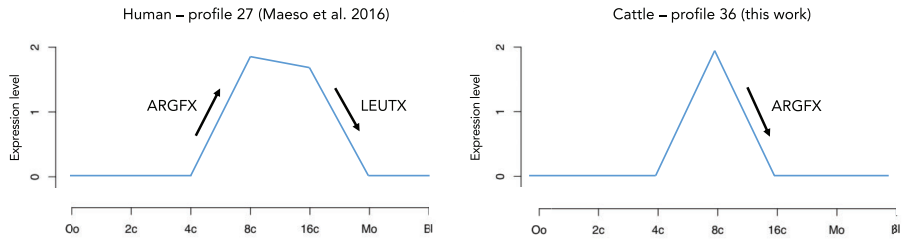


FIG. 7. Temporal expression profiles for genes upregulated in response to *ARGFX* expression in human (Maeso et al. 2016) and downregulated in response to *ARGFX* expression in bovine (this work). Oo, oocyte; 2c, 2-cell; 4c, 4-cell; 8c, 8-cell; 16c, 16-cell; Mo, morula; Bl, early blastocyst.

Table 1. Interferon-related GO Terms Enriched in *ARGFX* and *LEUTX*-Downregulated Gene Sets.

Interferon-related GO term	Bovine *ARGFX* FDR-adjusted *P*-value	Bovine *LEUTX* FDR-adjusted *P*-value	Bovine *LEUTX* and Human *LEUTX* Overlap FDR-adjusted *P*-value
Cellular response to interferon-beta (GO:0035458)	NS	NS	2.89×10^{-3}
Cellular response to interferon-gamma (GO:0071346)	NS	NS	1.24×10^{-3}
Cellular response to type I interferon (GO:0071357)	1.32×10^{-2}	1.66×10^{-2}	5.63×10^{-4}
Interferon-gamma-mediated signaling pathway (GO:0060333)	NS	NS	1.31×10^{-5}
Positive regulation of interferon-alpha production (GO:0032727)	NS	NS	1.28×10^{-2}
Positive regulation of interferon-beta production (GO:0032728)	NS	NS	3.99×10^{-6}
Positive regulation of type I interferon production (GO:0032481)	NS	NS	1.48×10^{-6}
Regulation of interferon-alpha production (GO:0032647)	NS	NS	1.30×10^{-3}
Regulation of interferon-beta production (GO:0032648)	NS	NS	1.34×10^{-7}
Regulation of type I interferon production (GO:0032479)	NS	NS	3.63×10^{-8}
Response to interferon-alpha (GO:0035455)	1.74×10^{-3}	2.71×10^{-3}	8.50×10^{-3}
Response to interferon-beta (GO:0035456)	2.54×10^{-4}	3.91×10^{-3}	7.75×10^{-5}
Response to interferon-gamma (GO:0034341)	4.69×10^{-3}	1.18×10^{-3}	5.27×10^{-5}
Response to type I interferon (GO:0034340)	8.14×10^{-4}	3.50×10^{-3}	6.51×10^{-7}
Response to type III interferon (GO:0034342)	NS	NS	3.40×10^{-2}
Type I interferon signaling pathway (GO:0060337)	4.57×10^{-2}	3.82×10^{-2}	6.83×10^{-3}

NOTE.—GO biological process terms related to interferon signaling enriched in gene sets downregulated by *ARGFX* and *LEUTX*. *P*-values are from binomial tests with FDR corrections for multiple testing. FDR, false discovery rate; GO, gene ontology; NS, not significant.

enriched for six GO Biological Process terms related to the interferon response (table 1). Overall, these data suggest that the expression of ISGs is downregulated by bovine *ARGFX* and *LEUTX*.

Through what mechanism are ETCHbox genes regulating ISGs? We suggest that ETCHbox genes are capable of downregulating the JAK/STAT pathway: we find that bovine *LEUTX* and/or *ARGFX* stimulate downregulation of expression of *JAK2*, *STAT1*, *STAT2* and upregulation of the JAK/STAT inhibitors *PIAS3* and *PIAS4* (supplementary table S2, Supplementary Material online). Since cell surface receptors compete for JAKs, meaning that JAK concentration is often a limiting factor for JAK/STAT signaling (Haan et al. 2006; Schneider et al. 2014), this downregulation of *JAK2* may result in a direct downregulation of JAK/STAT signaling potential. We note that there is also downregulation of the JAK/STAT inhibitors SOCS1-5; while this may initially seem counterintuitive, these are also ISGs and are upregulated by JAK/STAT, so a decrease in their expression is consistent with reduced JAK/STAT signaling (Matsumoto et al. 1997; Krebs and Hilton 2001; Bousoik and Montazeri Aliabadi 2018). The JAK/STAT signaling pathway plays important roles in preimplantation development at the time of normal ETCHbox expression. For example, in mice the pluripotency factor POU5F1 acts via JAK/STAT3 signaling to induce pluripotency (Stirparo et al. 2021), and JAK/STAT3 signaling alone can induce naïve pluripotency during reprograming (Yang et al. 2010; Van Oosten et al. 2012). In cattle, the JAK/STAT pathway is vital for inner cell mass gene expression and

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries of the University of Oxford user on 30 January 2023

development (Meng et al. 2015). Indeed, it has been found that JAK/STAT signaling is essential for pluripotency maintenance, self-renewal and the prevention of differentiation (Niwa et al. 1998; Nichols et al. 2001; Do et al. 2013; Stirparo et al. 2021); the downregulation of JAK/STAT signaling is, therefore, consistent with the above result that genes involved in the maintenance of an undifferentiated cell state are also downregulated by ETCHbox genes. Together with the temporal profile analyses, these data suggest that ETCHbox genes are involved in the promotion of blastocyst structural formation, and downregulate genes involved in the undifferentiation network in cells, possibly via the downregulation of JAK/STAT signaling. GO terms related to interferon signaling are strongly enriched in the set of genes commonly downregulated by human and bovine *LEUTX* (table 1 and supplementary table S8, Supplementary Material online); this suggests that this function is conserved between humans and cattle.

Conclusions

Fast-evolving homeobox genes pose a conundrum in developmental biology since they contrast so markedly from the evolutionary conservation typical of homeobox genes. In this work, we utilized an ectopic expression approach to compare the functions of bovine ETCHbox genes to those of humans and mice and found evidence suggesting that, remarkably, the function of bovine *ARGFX* is highly different to that of its human orthologue and is more similar to that of human *LEUTX*. This suggests that the gene's function has undergone a transition since the divergence of humans and cattle. In addition, we found that bovine *ARGFX* and *LEUTX* upregulate blastocyst-expressed genes, including genes involved in cell adhesion, the extracellular matrix and the cytoskeleton; we suggest that these genes have overlapping roles in orchestrating structural changes during blastocyst formation. Our results also suggest that ETCHbox genes downregulate JAK/STAT signaling, and that this may result in the downregulation of genes associated with an undifferentiated cell state. Finally, the existence of a frameshift allele that truncates the bovine ARGFX protein and eliminates many of its functions, but is nonlethal when homozygous, supports the hypothesis of redundancy between ETCHbox proteins.

Materials and Methods

Cell Isolation and Culture

Primary BFFs were isolated from the torso skin of a 6-month-old *B. taurus* fetus from a pregnant female sacrificed at an abattoir, using a method similar to that used to isolate human dermal fibroblasts (Vangipuram et al. 2013). Briefly, a 1 cm^2 section of skin was excised, transported on ice and washed 3 times with sterile phosphate-buffered saline (PBS) (Gibco #10010015). Subcutaneous fat was scraped off and the tissue was minced with a scalpel. Five 2 mm^2 pieces were placed in each well of a 0.001% poly-L-lysine-coated (Sigma-Aldrich, #P4832-50ML) 6-well plate, covered with a sterile glass coverslip, and 2 ml BFF media added. BFF media is Dulbecco's modified Eagle medium (Gibco #41965039) with 10% heat-inactivated fetal bovine serum (Gibco #10500064), 1% penicillin-streptomycin (Gibco #15140122), and 1% amphotericin B (Gibco #15290026). Outgrowth was allowed for 18 days at 37 °C with 5% CO_2, with the culture media changed every 3 days before cells were washed three times in 2 ml PBS and detached with 2 ml TrypLE Express enzyme (Gibco #12604013) at 37 °C with 5% CO_2 for 20 min. TrypLE Express was neutralized with 5 ml BFF medium, the coverslip removed, cells and medium transferred to a 50 ml Falcon tube and centrifuged for 5 min at 1,200 rpm. Cells were washed twice with 10 ml BFF medium and centrifuged, then seeded into 0.001% poly-L-lysine-coated T75 flasks. To maintain growth, fibroblasts were incubated at 37 °C with 5% CO_2 and passaged one flask to four every 3–4 days at 70–80% confluency. Genomic DNA was extracted using a DNeasy Blood and Tissue Kit (Qiagen #69504), and the *cytochrome c oxidase subunit I* gene amplified by PCR (primer pair 1; supplementary table S9, Supplementary Material online) and sequenced to verify species identity of the cultured cells (supplementary fig. S3, Supplementary Material online). Cells were tested and negative for mycoplasma contamination.

Ectopic Expression

Bos taurus ARGFX and *LEUTX* intron/exon structures were determined previously (Lewin et al. 2021). Where multiple isoforms were present, the isoform with the highest FPKM (Fragments per kilobase of transcript per million mapped reads) value and possessing an exon structure matching that of the human gene was used. *ARGFX* and *LEUTX* complete coding sequences with a C-terminal GGGGSGGGGS linker and 3xFLAG were synthesized by ThermoFisher GeneArt and Twist Bioscience, respectively (supplementary fig. S4, Supplementary Material online). Sequences were cloned into plasmid pSF-CMV-Puro-COOH (OXGENE #OG3422) under the control of a CMV promoter using XhoI (Thermo Scientific #FD0695) and NotI (Thermo Scientific #FD0593) (primer pair 2). An *ARGFX* mutant containing the 13 bp deletion found in the ARS-UCD1.2 reference genome was produced using a Q5 Site-Directed Mutagenesis Kit (New England BioLabs #E0554S). The deletion was induced with primer pair 3, then primer pair 4 used to replace the premature stop codon caused by the frameshift with the GGGGSGGGGS linker and in-frame 3xFLAG. Constructs were validated by Sanger sequencing (primer pair 5).

Cells were transfected with either the WT bovine *ARGFX* gene, *LEUTX* gene, mutant *ARGFX* gene or empty pSF-CMV-Puro-COOH vector as a control. For transfection, 1 million cells (passage <6) were resuspended in

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries of the University of Oxford user on 30 January 2023

98 µl Opti-MEM (Gibco #11058021), combined with 2 µl of 5 µg/µl endotoxin-free plasmid, electroporated with a 2 mm gap using a NEPA21 Super Electroporator (NEPAGENE) and seeded in 2 ml BFF medium. The medium was changed after 24 h, and puromycin (Sigma Aldrich #P8833) added at 4 µg/ml to select for transfected cells. After 72 h, RNA was extracted using an RNeasy Plus Micro kit (Qiagen #74034). RNA purity was measured using a Nanodrop ND-1000 spectrophotometer and RNA integrity checked using an Agilent 2100 Bioanalyzer.

For one sample of each construct, immunocytochemistry was undertaken to check for protein expression using an antibody to 3xFLAG, as described in previous work (Maeso et al. 2016) with the following differences: primary antibody = monoclonal mouse anti-FLAG M2 antibody (Sigma Aldrich #F1804-50UG) 1:100, 4 h incubation time; secondary antibody = goat antimouse IgG (H + L) superclonal recombinant secondary antibody with Alexa Fluor 488 (Invitrogen #A28175) 1:1000, 1 h incubation time. Cells were mounted using SlowFade Gold Antifade Mountant with DAPI (Invitrogen #S36938) to stain DNA and visualized with a Zeiss Axioskop 2 Plus fluorescence microscope and Zeiss Axiocam 202 mono camera. Images were processed using ImageJ.

Analysis of RNA-seq Data

RNA sequencing was performed on four biological replicates for each plasmid using the Illumina NovaSeq 6000 Sequencing System (Novogene), giving between 77.7 million and 120.9 million paired-end 150 bp reads per sample. Reads were filtered to remove those containing adapters, those with $N > 10\%$ and those of low quality (Qscore < 5). Quality control was performed using FastQC version 0.11.8 (Andrews 2010), MultiQC version 1.8 (Ewels et al. 2016), and Trimmomatic version 0.39 used to trim reads (Bolger et al. 2014). The pseudoaligner kallisto version 0.46.1 (Bray et al. 2016) was used to align reads to the *B. taurus* ARS-UCD1.2 transcriptome and quantify abundances; *B. taurus* ETCHbox sequences were added to the transcriptome before pseudoalignment. STAR version 2.7.9a (Dobin et al. 2013), StringTie version 2.1.7 (Pertea et al. 2015), and the Integrative Genomics Viewer (Robinson et al. 2011) were used to align, assemble, and visualize reads of *ARGFX* and *ARGFX* mutant samples to verify that the correct version was expressed. R version 4.1.0 was used for data analysis (R Core Team, 2020). Tximport version 1.20.0 was used to create count tables (Soneson et al. 2016), and DESeq2 version 1.32.0 used for DE analysis (Love et al. 2014). The false discovery rate (FDR = 0.05) method was used to correct for multiple testing. Genes with an adjusted P-value < 0.05, fold change $>$ 1.5, and mean TPM (transcripts per million) > 2 were considered DE. One of the four *ARGFX* mutant samples was excluded as it was an outlier on PCA plots, clustering with *ARGFX* WT samples (supplementary fig. S5, Supplementary Material online). Volcano plots were created using R package EnhancedVolcano version 1.10.0 (Blighe et al. 2021) and Venn diagrams with BioVenn (Hulsen et al. 2008). GO analysis (Ashburner et al. 2000) was completed using PANTHER version 16.0 (Thomas et al. 2003; Mi et al. 2013, 2021) using binomial tests and an FDR correction for multiple testing (FDR = 0.05).

We tested whether there were significant overlaps between the genes DE in response to the expression of *B. taurus*, *M. musculus*, and *H. sapiens* ETCHbox genes. To achieve this, we combined the bovine data from this work with lists of genes reported to be DE in response to *H. sapiens ARGFX*, *DPRX*, *LEUTX*, and *TPRX1* ectopic expression and mouse *Crxos*, *Oboxa1*, *Oboxa4*, and *Oboxa7* ectopic expression (Maeso et al. 2016, Royall et al. 2018). Gene lists were first restricted to 1:1 orthologues, identified using OMA (Altenhoff et al. 2021; supplementary table S10*a* and *b*, Supplementary Material online). Fisher's exact test was then used to test for overlaps in the gene sets up- or downregulated by bovine *ARGFX* or *LEUTX* and each of either the human or mouse ETCHbox genes. Many of these tests produce significant overlaps so, to identify the "most significant" overlaps and facilitate clearer comparisons, we calculated an overlap score consisting of the $-\log_{10}$ of the multiple comparison-corrected P-value. The greater the overlap score, the more the overlap in the genes' activity exceeds that expected by chance. For these tests, the Benjamini-Yekutieli FDR method (Benjamini and Yekutieli 2001) (FDR = 0.05) was used to correct for multiple testing because it does not assume independence of tests.

To gain insights into the role of bovine ETCHbox genes in the preimplantation embryo, temporal expression profile clustering was employed. To obtain a dataset for clustering, raw reads from a published bovine embryonic RNA-seq dataset (Jiang et al. 2014), covering oocyte, 2-cell, 4-cell, 8-cell, 16-cell, early morula, compact morula and blastocyst stages, were downloaded from NCBI SRA (PRJNA254699) and mapped with STAR (Dobin et al. 2013), then transcripts were quantified with StringTie (Pertea et al. 2015) as above. Mfuzz version 2.52.0 (Kumar and Futschik 2007) was then used to cluster bovine embryonically expressed genes into temporal expression profiles. The fuzzy c-means algorithm of Mfuzz was selected as it is more robust to the noise present in RNA-seq data than is k-means clustering. Clusters were restricted to genes with a membership value >0.5, and Fisher's exact test was used to test for enrichment of genes of each cluster in the sets of genes DE in response to ETCHbox expression. The Benjamini and Hochberg (1995) FDR method (FDR = 0.05) was used to correct for multiple testing.

Fisher's exact test with a Benjamini and Hochberg (1995) adjustment for multiple testing was used to test for enrichment of 'undifferentiation' marker genes in the lists of genes DE in response to bovine ETCHbox ectopic expression. The undifferentiation network signature consists of 266 genes (Galan et al. 2013), of which 234 had identifiable 1:1 orthologues in bovine.

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries of the University of Oxford user on 30 January 2023

Frequency of an *ARGFX* Mutant Allele

The Bovine Genome Variation Database (Chen et al. 2020) was used to assess the allele frequency and distribution of a mutant *ARGFX* allele identified in the *B. taurus* ARS-UCD1.2 reference genome. To test whether individuals can be homozygous for the mutant form of *ARGFX* or whether it is lethal, PCR (GoTaq G2 DNA polymerase, Promega #M7841) was used to characterize the *ARGFX* gene of 11 individuals of the Jersey breed, which possess a high frequency of the mutant allele. Nasal fluid was collected from *B. taurus* individuals using GenoTube Livestock swabs (ThermoFisher Scientific #9062010). To extract genomic DNA, a 5 mm^2 fragment was cut from the swab with sterile scalpels and minced, and then treated as a tissue sample for extraction using the DNeasy Blood and Tissue Kit (QIAGEN #69504). A primer set (pair 6) was designed with primers either side of the deletion, giving a 13 bp shorter amplicon for a mutant allele versus WT. To distinguish bands with a 13 bp difference in length, a 3% TBE (Tris–Borate–EDTA) gel was used. To test which form of *ARGFX* is present in the BFF cells, genomic DNA was extracted with a DNA Isolation Kit (Roche #11814770001) using the standard protocol, and the above primer sets used.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank David Marden (Elm Farm) and Paul Redmore (Neston Park Home Farm) for access to Jersey cows for DNA extraction. We also thank Dunbia Cardington for access to tissues to facilitate fibroblast isolation, and Aziz Aboobaker, Yichen Dai, Yi-Jyun Luo, Peter Mulhair, Rodrigo Pracana, Amy Royall, Sebastian Shimeld, Tim Davies, and Sonia Trigueros for helpful discussions and advice. We appreciate the efforts of two anonymous reviewers whose comments improved the manuscript. This work was supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) (grant number BB/M011224/1) and an Oxford-Wolfson Marriott BBSRC Graduate Scholarship.

Author Contributions

T.D.L. and P.W.H.H. conceived and designed the study. A.A.F.-N. and T.D.L. generated fetal fibroblast cell line. Wet laboratory experiments and data analysis were performed by T.D.L. with input from P.W.H.H. T.D.L. and P.W.H.H. wrote the manuscript with input from A.A.F.-N. All authors read and approved the final manuscript.

Data Availability

Raw and processed sequencing datasets have been deposited to the NCBI Gene Expression Omnibus (GEO) (www.ncbi.nlm.nih.gov/geo) under accession GSE192356.

References

R Core Team. 2020. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing

Aberkane A, Essahib W, Spits C, De Paepe C, Sermon K, Adriaenssens T, Mackens S, Tournaye H, Brosens JJ, Van De Velde H. 2018. Expression of adhesion and extracellular matrix genes in human blastocysts upon attachment in a 2D co-culture system. *Mol Hum Reprod*. **24**(7):375–387.

Ahlman T, Berglund B, Rydhmer L, Strandberg E. 2011. Culling reasons in organic and conventional dairy herds and genotype by environment interaction for longevity. *J Dairy Sci*. **94**(3):1568–1575.

Albertini DF, Overstrom EW, Ebert KM. 1987. Changes in the organization of the actin cytoskeleton during preimplantation development of the pig embryo. *Biol Reprod*. **37**(2):441–451.

Altenhoff AM, Train C-M, Gilbert KJ, Mediratta I, Mendes de Farias T, Moi D, Nevers Y, Radoykova H-S, Rossier V, Warwick Vesztrocy A, *et al*. 2021. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res*. **49**(D1):D373–D379.

Andrews S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Babraham Institute. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ansari-Lari M, Mohebbi-Fani M, Rowshan-Ghasrodashti A. 2012. Causes of culling in dairy cows and its relation to age at culling and interval from calving in Shiraz, Southern Iran. *Vet Res Forum* **3**(4):233–237.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet*. **25**(1):25–29.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* **57**(1):289–300.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. **29**(4):1165–1188.

Berry DP, Wall E, Pryce JE. 2014. Genetics and genomics of reproductive performance in dairy and beef cattle. *Animal* **8**(Suppl. 1):105–121.

Blighe K, Rana S, Lewis M. 2021. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. R-Package version 1.10.0. https://github.com/kevinblighe/EnhancedVolcano.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15):2114–2120.

Bousoik E, Montazeri Aliabadi H. 2018. "Do We Know Jack" about JAK? A closer look at JAK/STAT signaling pathway. *Front Oncol*. **8**:287.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. **34**(5):525–527.

Bürglin TR, Affolter M. 2016. Homeodomain proteins: an update. *Chromosoma* **125**(3):497–521.

Chen N, Fu W, Zhao J, Shen J, Chen Q, Zheng Z, Chen H, Sonstegard TS, Lei C, Jiang Y. 2020. BGVD: an integrated database for bovine sequencing variations and selective signatures. *Genom Proteom Bioinform*. **18**(2):186–193.

Chiumia D, Chagunda MGG, Macrae AI, Roberts DJ. 2013. Predisposing factors for involuntary culling in Holstein-Friesian dairy cows. *J Dairy Res*. **80**(1):45–50.

Damsky C, Sutherland A, Fisher S. 1993. Extracellular matrix 5: adhesive interactions in early mammalian embryogenesis, implantation, and placentation. *FASEB J*. **7**(14):1320–1329.

DeDiego ML, Nogales A, Martinez-Sobrido L, Topham DJ. 2019. Interferon-induced protein 44 interacts with cellular fk506-binding

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries of the University of Oxford user on 30 January 2023

protein 5, negatively regulates host antiviral responses, and supports virus replication. *mBio* **10**(4):e01839-19.
DeFilippis VR, Alvarado D, Sali T, Rothenburg S, Früh K. 2010. Human cytomegalovirus induces the interferon response via the DNA sensor ZBP1. *J Virol*. **84**(1):585–598.
Delgado CL. 2003. Rising consumption of meat and milk in developing countries has created a new food revolution. *J Nutr*. **133**(11 Suppl. 2):3907S–3910S.
Diskin MG, Morris DG. 2008. Embryonic and early foetal losses in cattle and other ruminants. *Reprod Domest Anim*. **43**(Suppl. 2): 260–267.
Do DV, Ueda J, Messerschmidt DM, Lorthongpanich C, Zhou Y, Feng B, Guo G, Lin PJ, Hossain MZ, Zhang W, *et al.* 2013. A genetic and developmental pathway from STAT3 to the OCT4-NANOG circuit is essential for maintenance of ICM lineages in vivo. *Genes Dev*. **27**(12):1378–1390.
Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1):15–21.
Eckert JJ, Fleming TP. 2008. Tight junction biogenesis during early development. *Biochim Biophys Acta* **1778**(3):717–728.
Enders AC, Lantz KC, Schlafke S. 1990. The morula-blastocyst transition in two old world primates: the Baboon and Rhesus Monkey. *J Med Primatol*. **19**(8):725–747.
Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**(19):3047–3048.
Fleming TP. 2001. Cell adhesion in the preimplantation mammalian embryo and its role in trophectoderm differentiation and blastocyst morphogenesis. *Front Biosci*. **6**(1):D1000-7.
Fleming TP, Ghassemifar MR, Sheth B. 2000. Junctional complexes in the early mammalian embryo. *Semin Reprod Med*. **18**(2):185–194.
Galan A, Diaz-Gimeno P, Poo ME, Valbuena D, Sanchez E, Ruiz V, Dopazo J, Montaner D, Conesa A, Simon C. 2013. Defining the genomic signature of totipotency and pluripotency during early human development. *PLoS One* **8**(4):e62135.
FAO. 2009. Global agriculture towards 2050. High Level Expert Forum - How to Feed the World in 2050. Rome, Italy. Available online at: https://www.fao.org/fileadmin/templates/wsfs/docs/Issues_papers/HLEF2050_Global_Agriculture.pdf [Date accessed: 17/12/2021].
Graf A, Krebs S, Zakhartchenko V, Schwalb B, Blum H, Wolf E. 2014. Fine mapping of genome activation in bovine embryos by RNA sequencing. *Proc Natl Acad Sci U S A*. **111**(11):4139–4144.
Haan C, Kreis S, Margue C, Behrmann I. 2006. Jaks and cytokine receptors – an intimate relationship. *Biochem Pharmacol*. **72**(11): 1538–1546.
Hardy K, Warner A, Winston RML, Becker DL. 1996. Expression of intercellular junctions during preimplantation development of the human embryo. *Mol Hum Reprod*. **2**(8):621–632.
Harwardt T, Lukas S, Zenger M, Reitberger T, Danzer D, Übner T, Munday DC, Nevels M, Paulus C. 2016. Human cytomegalovirus immediate-early 1 protein rewires upstream STAT3 to downstream STAT1 signaling switching an IL6-type to an IFNγ-like response. *PLoS Pathog*. **12**(7):e1005748.
Hasley A, Chavez S, Danilchik M, Wühr M, Pelegri F. 2017. Vertebrate embryonic cleavage pattern determination. *Adv Exp Med Biol*. **953**:117–171.
He B, Zhang H, Wang J, Liu M, Sun Y, Guo C, Lu J, Wang H, Kong S. 2019. Blastocyst activation engenders transcriptome reprogram affecting X-chromosome reactivation and inflammatory trigger of implantation. *Proc Natl Acad Sci U S A*. **116**(33):16621–16630.
Hernandez JM, Elahi A, Clark W, Humphries LA, Wang J, Achille A, Seto E, Shibata D. 2015. The tumor suppressive effects of HPP1 are mediated through JAK-STAT-interferon signaling pathways. *DNA Cell Biol*. **34**(8):541–549.
Hu K. 2019. On mammalian totipotency: what is the molecular underpinning for the totipotency of zygote? *Stem Cells Dev*. **28-**(14):897–906.
Hu Y, Wang J, Yang B, Zheng N, Qin M, Ji Y, Lin G, Tian L, Wu X, Wu L, *et al.* 2011. Guanylate binding protein 4 negatively regulates virus-induced type I IFN and antiviral response by targeting IFN regulatory factor 7. *J Immunol*. **187**(12):6456–6462.
Hulsen T, de Vlieg J, Alkema W. 2008. BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics* **9**:488.
Indraccolo S, Pfeffer U, Minuzzo S, Esposito G, Roni V, Mandruzzato S, Ferrari N, Anfosso L, Dell'Eva R, Noonan DM, *et al.* 2007. Identification of genes selectively regulated by IFNs in endothelial cells. *J Immunol*. **178**(2):1122–1135.
Jiang Z, Sun J, Dong H, Luo O, Zheng X, Obergfell C, Tang Y, Bi J, O'Neill R, Ruan Y, *et al.* 2014. Transcriptional profiles of bovine in vivo pre-implantation development. *BMC Genomics* **15**(1):756.
Jouhilahti E-M, Madissoon E, Vesterlund L, Töhönen V, Krjutškov K, Plaza Reyes A, Petropoulos S, Månsson R, Linnarsson S, Bürglin T, *et al.* 2016. The human PRD-like homeobox gene LEUTX has a central role in embryo genome activation. *Development* **143**(19):3459–3469.
Kane M, Yadav SS, Bitzegeio J, Kutluay SB, Zang T, Wilson SJ, Schoggins JW, Rice CM, Yamashita M, Hatziioannou T, *et al.* 2013. MX2 is an interferon-induced inhibitor of HIV-1 infection. *Nature* **502**(7472):563–566.
Katayama S, Ranga V, Jouhilahti EM, Airenne TT, Johnson MS, Mukherjee K, Bürglin TR, Kere J. 2018. Phylogenetic and mutational analyses of human LEUTX, a homeobox gene implicated in embryogenesis. *Sci Rep*. **8**(1):17421.
Koyama H, Suzuki H, Yang X, Jiang S, Foote RH. 1994. Analysis of polarity of bovine and rabbit embryos by scanning electron microscopy. *Biol Reprod*. **50**(1):163–170.
Krebs DL, Hilton DJ. 2001. SOCS proteins: negative regulators of cytokine signaling. *Stem Cells* **19**(5):378–387.
Kuepper MK, Bütow M, Herrmann O, Ziemons J, Chatain N, Maurer A, Kirschner M, Maié T, Costa IG, Eschweiler J, *et al.* 2019. Stem cell persistence in CML is mediated by extrinsically activated JAK1-STAT3 signaling. *Leukemia* **33**(8):1964–1977.
Kumar L, Futschik ME. 2007. Mfuzz: a software package for soft clustering of microarray data. *Bioinformation* **2**(1):5–7.
Lehtonen E, Ordónez G, Reima I. 1988. Cytoskeleton in preimplantation mouse development. *Cell Differ*. **24**(3):165–77.
Lewin TD, Royall AH, Holland PWH. 2021. Dynamic molecular evolution of mammalian homeobox genes: duplication, loss, divergence and gene conversion sculpt PRD-class repertoires. *J Mol Evol*. **89**(6):396–414.
Lim HYG, Plachta N. 2021. Cytoskeletal control of early mammalian development. *Nat Rev Mol Cell Biol*. **22**(8):548–562.
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. **15**(12):550.
Lu F, Zhang Y. 2015. Cell totipotency: molecular features, induction, and maintenance. *Natl Sci Rev*. **2**(2):217–225.
Lucy MC. 2001. Reproductive loss in high-producing dairy cattle: where will it end? *J Dairy Sci* **84**(6):1277–1293.
Madissoon E, Jouhilahti EM, Vesterlund L, Töhönen V, Krjutškov K, Petropoulous S, Einarsdottir E, Linnarsson S, Lanner F, Månsson R, *et al.* 2016. Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. *Sci Rep*. **6**:28995.
Maeso I, Dunwell TL, Wyatt CDR, Marlétaz F, Veto B, Bernal JA, Quah S, Irimia M, Holland PWH. 2016. Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals. *BMC Biol*. **14**:45.
Matsumoto A, Masuhara M, Mitsui K, Yokouchi M, Ohtsubo M, Misawa H, Miyajima A, Yoshimura A. 1997. CIS, a cytokine inducible SH2 protein, is a target of the JAK-STAT5 pathway and modulates STAT5 activation. *Blood* **89**(9):3148–3154.
Mazid MA, Ward C, Luo Z, Liu C, Li Y, Lai Y, Wu L, Li J, Jia W, Jiang Y, *et al.* 2022. Rolling back human pluripotent stem cells to an eight-cell embryo-like stage. *Nature* 605:315-324. doi:10.1038/s41586-022-04625-0.

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries of the University of Oxford user on 30 January 2023

Meng F, Forrester-Gauntlett B, Turner P, Henderson H, Oback B. 2015. Signal inhibition reveals JAK/STAT3 pathway as critical for bovine inner cell mass development. *Biol Reprod*. **93**(6):132.
Mi H, Ebert D, Muruganujan A, Mills C, Albou LP, Mushayamaha T, Thomas PD. 2021. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res*. **49**(D1):D394–D403.
Mi H, Muruganujan A, Thomas PD. 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res*. **41**(Database issue):D377–D386.
Mori Y, Sakuraoka M, Suzuki T, Sho S, Sugawara S, Hiraide M, Suguru S, Kobayashi M. 2018. Exogenous TPRX1 homeoprotein modulates the gene expression of lineage-specific transcription factors in human embryonal carcinoma cells. *Biotechnol Biotechnol Equip*. **33**(1):498–503.
Nichols J, Chambers I, Taga T, Smith A. 2001. Physiological rationale for responsiveness of mouse embryonic stem cells to gp130 cytokines. *Development* **128**(12):2333–2339.
Niwa H, Burdon T, Chambers I, Smith A. 1998. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev*. **12**(13):2048–2060.
Oltenacu PA, Algers B. 2005. Selection for increased production and the welfare of dairy cows: are new breeding goals needed? *Ambio* **34**(4-5):311–315.
Perkel KJ, Tscherner A, Merrill C, Lamarre J, Madan P. 2015. The ART of selecting the best embryo: a review of early embryonic mortality and bovine embryo viability assessment methods. *Mol Reprod Dev*. **82**(11):822–838.
Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. **33**(3):290–295.
FAO. 2019. Production: Crops and Livestock Products. FAOSTAT. Latest update: 15/09/2021. Available online at: https://www.fao.org/faostat/en/#data [Date accessed: 17/12/2021].
Renfree MB. 2010. Review: Marsupials: placental mammals with a difference. *Placenta* **31**(Suppl.):S21–S26.
Renfree MB, Lewis AMD. 1996. Cleavage in vivo and in vitro in the marsupial *Macropus eugenii*. *Reprod Fertil Dev*. **8**(4):725–742.
Renfree MB, Shaw G. 2014. Embryo-endometrial interactions during early development after embryonic diapause in the marsupial tammar wallaby. *Int J Dev Biol*. **58**(2-4):175–181.
Rilanto T, Reimus K, Orro T, Emanuelson U, Viltrop A, Mõtus K. 2020. Culling reasons and risk factors in Estonian dairy cows. *BMC Vet Res*. **16**(1):173.
Ritchie H, Roser M. 2017. Meat and Dairy Production. Our World Data. Available online at: https://ourworldindata.org/meat-production [Date accessed 17/12/2021.
Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol*. **29**(1):24–26.
Royall AH, Frankenberg S, Pask AJ, Holland PWH. 2019. Of eyes and embryos: subfunctionalization of the CRX homeobox gene in mammalian evolution. *Proc Biol Sci*. **286**(1907):20190830.
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.
Schneider WM, Chevillotte MD, Rice CM. 2014. Interferon-stimulated genes: a complex web of host defenses. *Annu Rev Immunol*. **32**:513–545.
Seo JY, Yaneva R, Cresswell P. 2011. Viperin: a multifunctional, interferon-inducible protein that regulates virus replication. *Cell Host Microbe* **10**(6):534–539.
Sheldon IM, Dobson H. 2003. Reproductive challenges facing the cattle industry at the beginning of the 21st century. *Reprod Suppl*. **61**:1–13.
Soneson C, Love MI, Robinson MD. 2016. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**:1521.
Stirparo GG, Kurowski A, Yanagida A, Bates LE, Strawbridge SE, Hladkou S, Stuart HT, Boroviak TE, Silva JCR, Nichols J. 2021. OCT4 induces embryonic pluripotency via STAT3 signaling and metabolic mechanisms. *Proc Natl Acad Sci U S A*. **118**(3):e2008890118.
Sun Y, Qiao L, Xia HHX, Lin MCM, Zou B, Yuan Y, Zhu S, Gu Q, Cheung TK, Kung HF, *et al.* 2008. Regulation of XAF1 expression in human colon cancer cell by interferon β: activation by the transcription regulator STAT1. *Cancer Lett*. **260**(1-2):62–71.
Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**(5):861–872.
Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**(4):663–676.
Taubenschmid-Stowers J, Rostovskaya M, Santos F, Ljung S, Argelaguet R, Krueger F, Nichols J, Reik W. 2022. 8C-like cells capture the human zygotic genome activation program. *Cell Stem Cell* **29**(3):449–459.e6.
Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res*. **13**(9):2129–2141.
Tirumurugaan KG, Pawar RM, Raj GD, Thangavelu A, Hammond JA, Parida S. 2020. RNAseq reveals the contribution of interferon stimulated genes to the increased host defense and decreased PPR viral replication in cattle. *Viruses* **12**(4):463.
Vangipuram M, Ting D, Kim S, Diaz R, Schüle B. 2013. Skin punch biopsy explant culture for derivation of primary human fibroblasts. *J Vis Exp*. **77**:e3779.
Van Oosten AL, Costa Y, Smith A, Silva JCR. 2012. JAK/STAT3 signalling is sufficient and dominant over antagonistic cues for the establishment of naive pluripotency. *Nat Commun*. **3**:817.
Vestweber D, Gossler A, Boller K, Kemler R. 1987. Expression and distribution of cell adhesion molecule uvomorulin in mouse preimplantation embryos. *Dev Biol*. **124**(2):451–456.
Vladimer GI, Górna MW, Superti-Furga G. 2014. IFITs: emerging roles as key anti-viral proteins. *Front Immunol*. **5**:94.
Watson AJ, Barcroft LC. 2001. Regulation of blastocyst formation. *Front Biosci*. **6**(1):D708–D730.
Yang J, Van Oosten AL, Theunissen TW, Guo G, Silva JCR, Smith A. 2010. Stat3 activation is limiting for reprogramming to ground state pluripotency. *Cell Stem Cell* **7**(3):319–328.
Zhou X, Michal JJ, Zhang L, Ding B, Lunney JK, Liu B, Jiang Z. 2013. Interferon induced IFIT family genes in host antiviral defense. *Int J Biol Sci*. **9**(2):200–208.

Downloaded from https://academic.oup.com/mbe/article/39/5/msac098/6581424 by Bodleian Libraries of the University of Oxford user on 30 January 2023