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

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

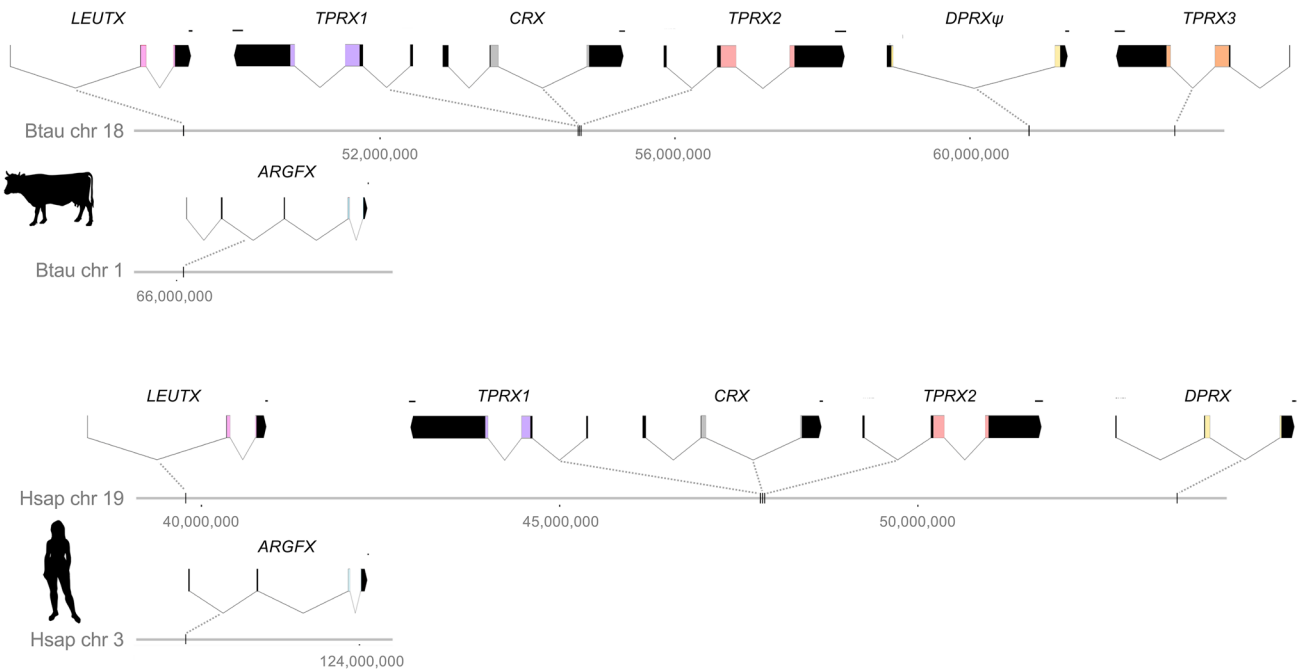

Fig. 1 ETCHbox repertoires of *Homo sapiens* (humans) and *Bos taurus* (cattle), with gene structures as determined using transcriptome assemblies. Horizontal grey bars represent chromosomes, vertical black bars represent the genomic position of ETCHbox genes. For gene structure representations, coding 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*

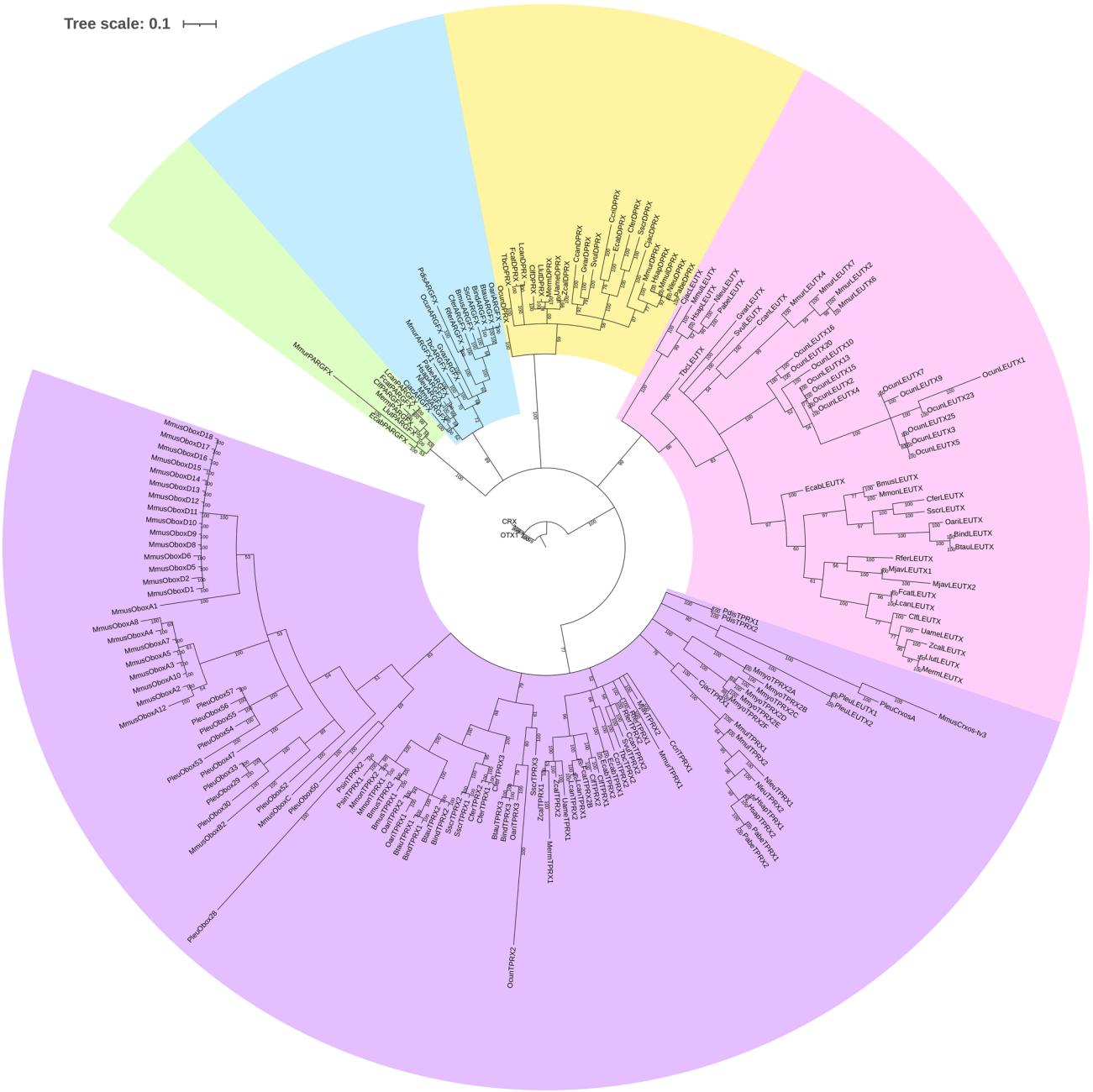

Fig. 2 Bayesian gene tree of putatively functional ETCHbox genes identified in this work. Colours highlight ETCHbox gene families; labels show posterior probabilities. The *ARGFX*, *DPRX*, *LEUTX* and *PARGFX* clades are supported by ≥ 99% probabilities. Due to the limited length of the homeodomain (60 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*

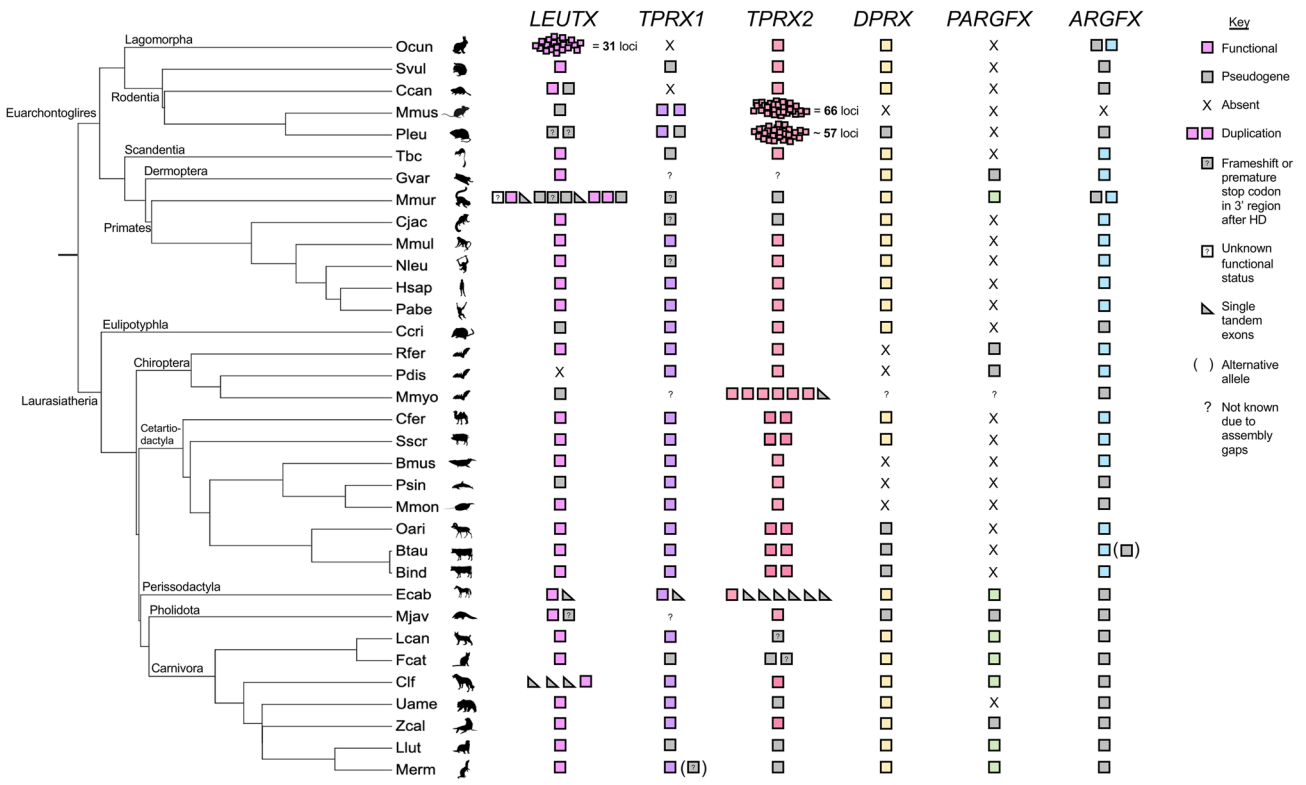

Fig. 3 ETCHbox gene repertoires of 34 eutherian mammals. Phylogenetic relationships are based on TimeTree (Kumar et al. 2017). Coloured boxes = putatively functional gene. Multiple boxes = gene duplicates. Black X = no gene remnants (complete gene loss). Grey boxes = putative pseudogene; grey boxes with 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

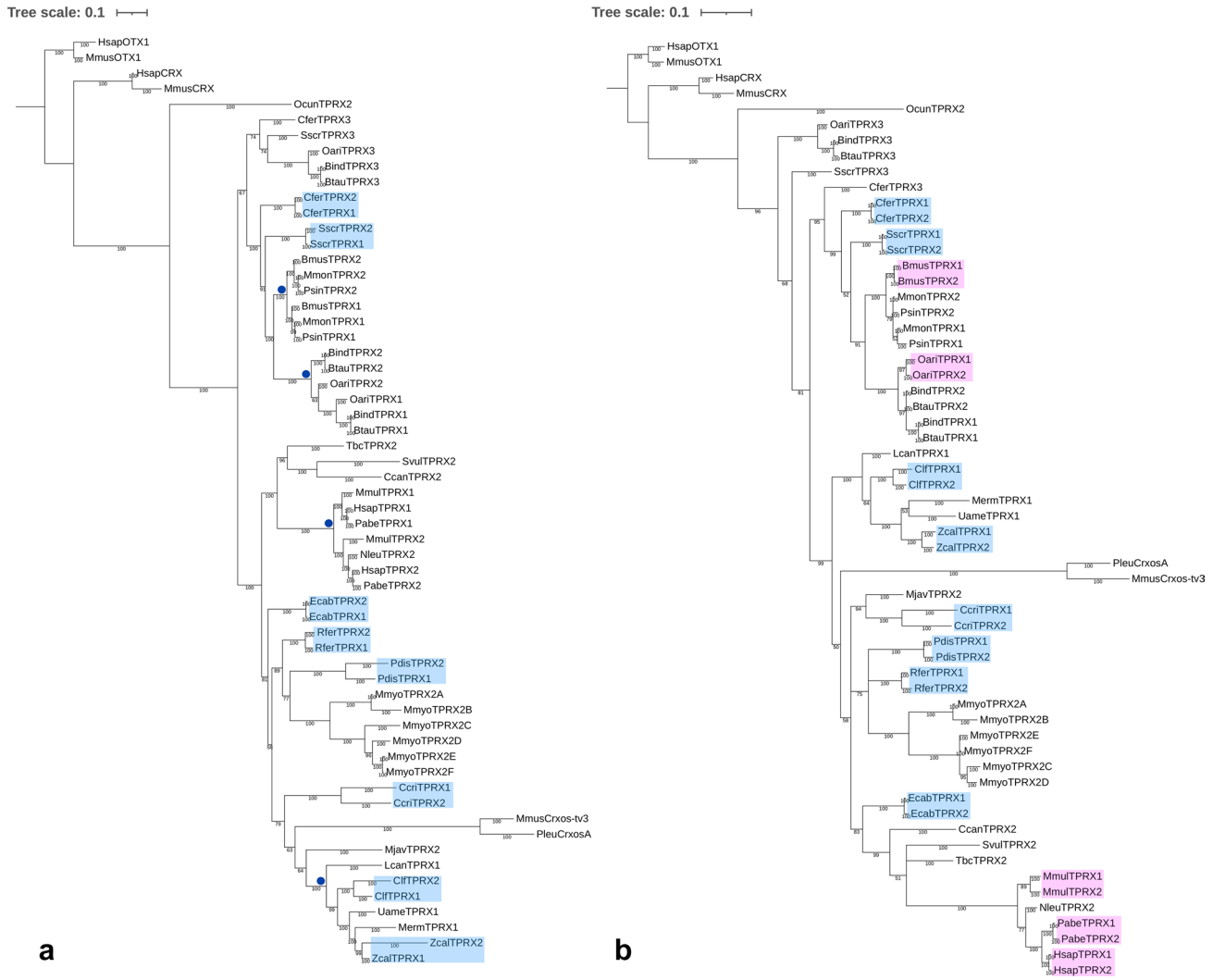

Fig. 4 Bayesian phylogenies of putatively functional *TPRX1*, *TPRX2* and *TPRX3* full gene sequences (**a**) and homeoboxes (**b**). Blue boxes highlight cases where conspecific *TPRX1* and *TPRX2* pairs are more closely related to each other than to other sequences. Pink boxes highlight cases that 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

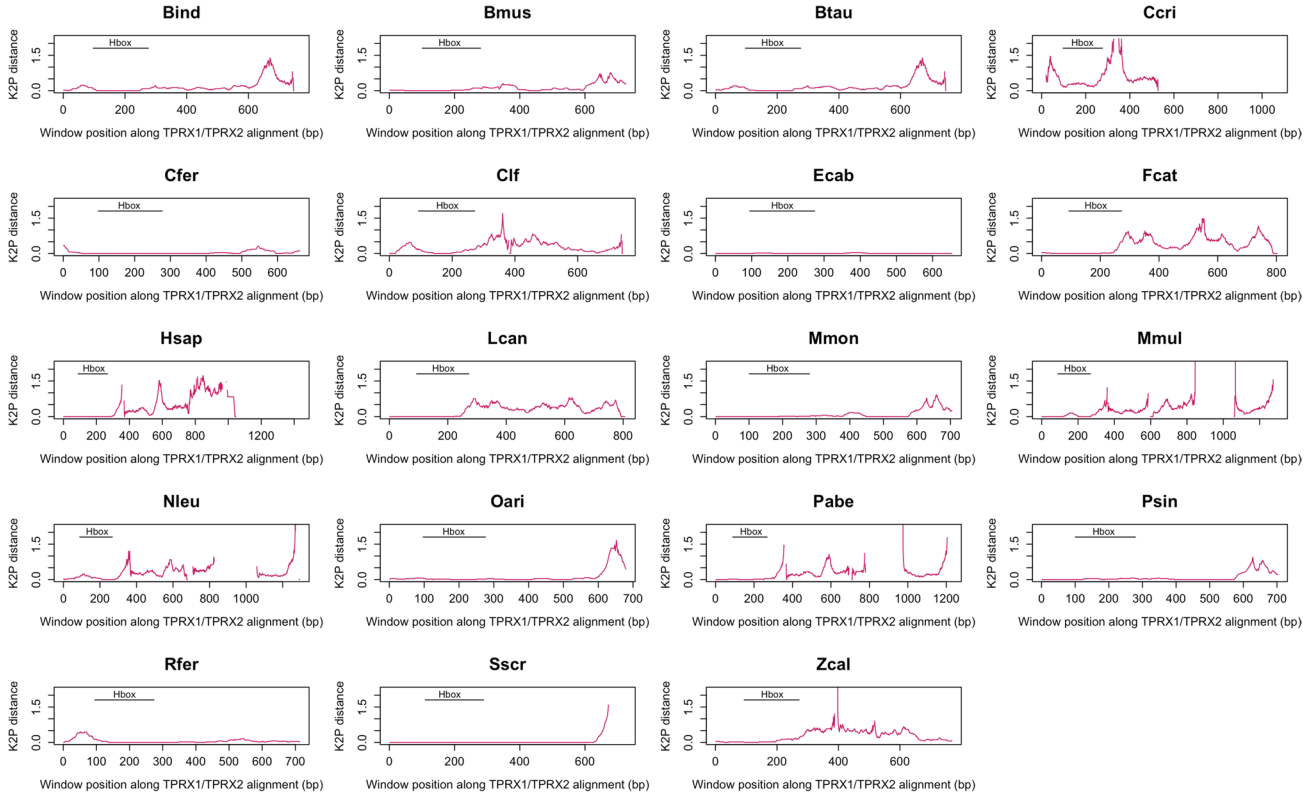

Fig. 5 Sequence similarity between *TPRX1* and *TPRX2* genes within a species. Plots show the Kimura 2-parameter (K2P) distance in 50 bp sliding windows between conspecific *TPRX1* and *TPRX2* genes. Higher K2P values indicate more divergent sequences. Gaps in the trace indicate indels in the alignment. 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

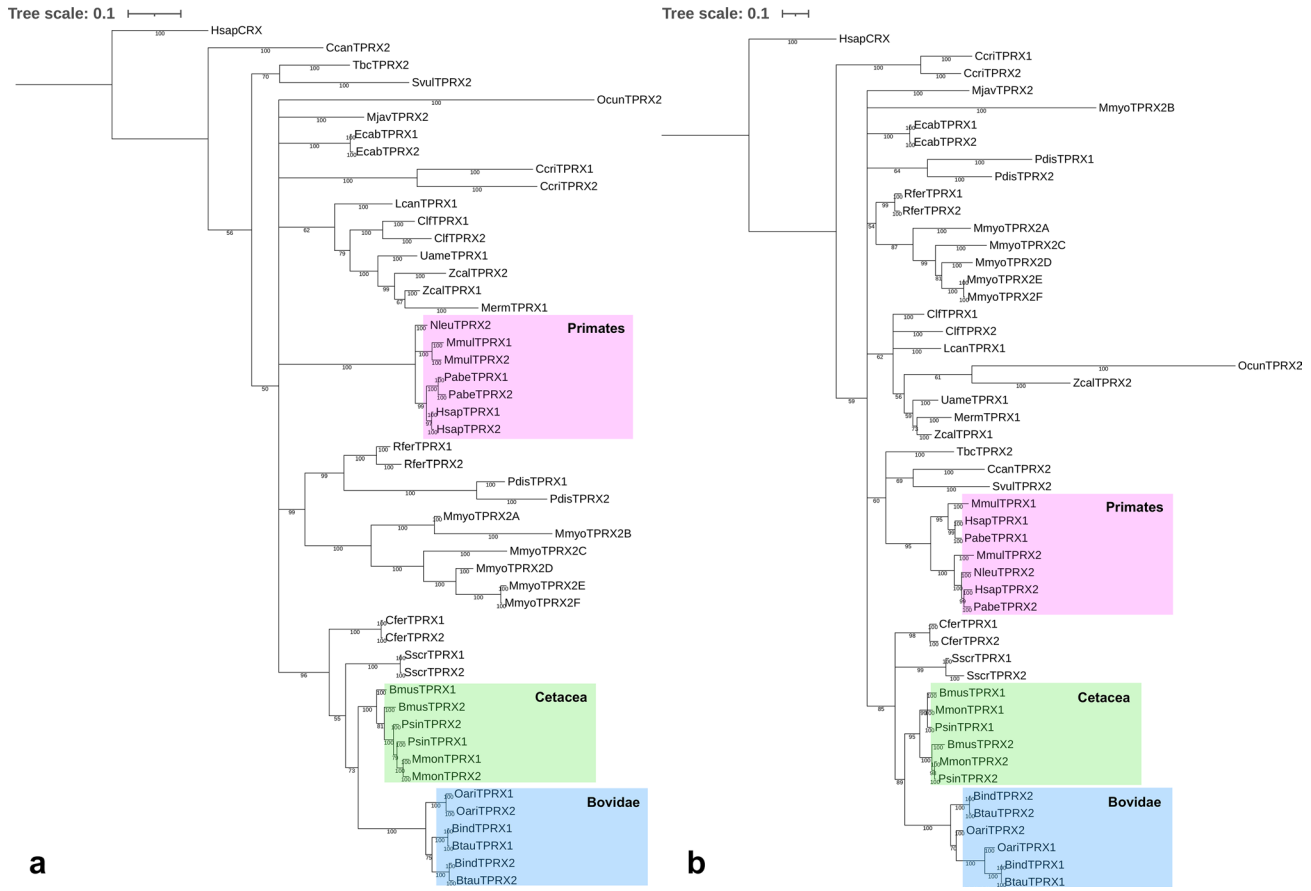

Fig. 6 Bayesian phylogenies inferred using partition 1 (**a**) and partition 2 (**b**) of putatively functional *TPRX1* and *TPRX2* genes split at the gene conversion breakpoint identified by GARD. Boxes highlight the Bovidae, Cetacea and Primates, where topology differs markedly between the two trees. 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

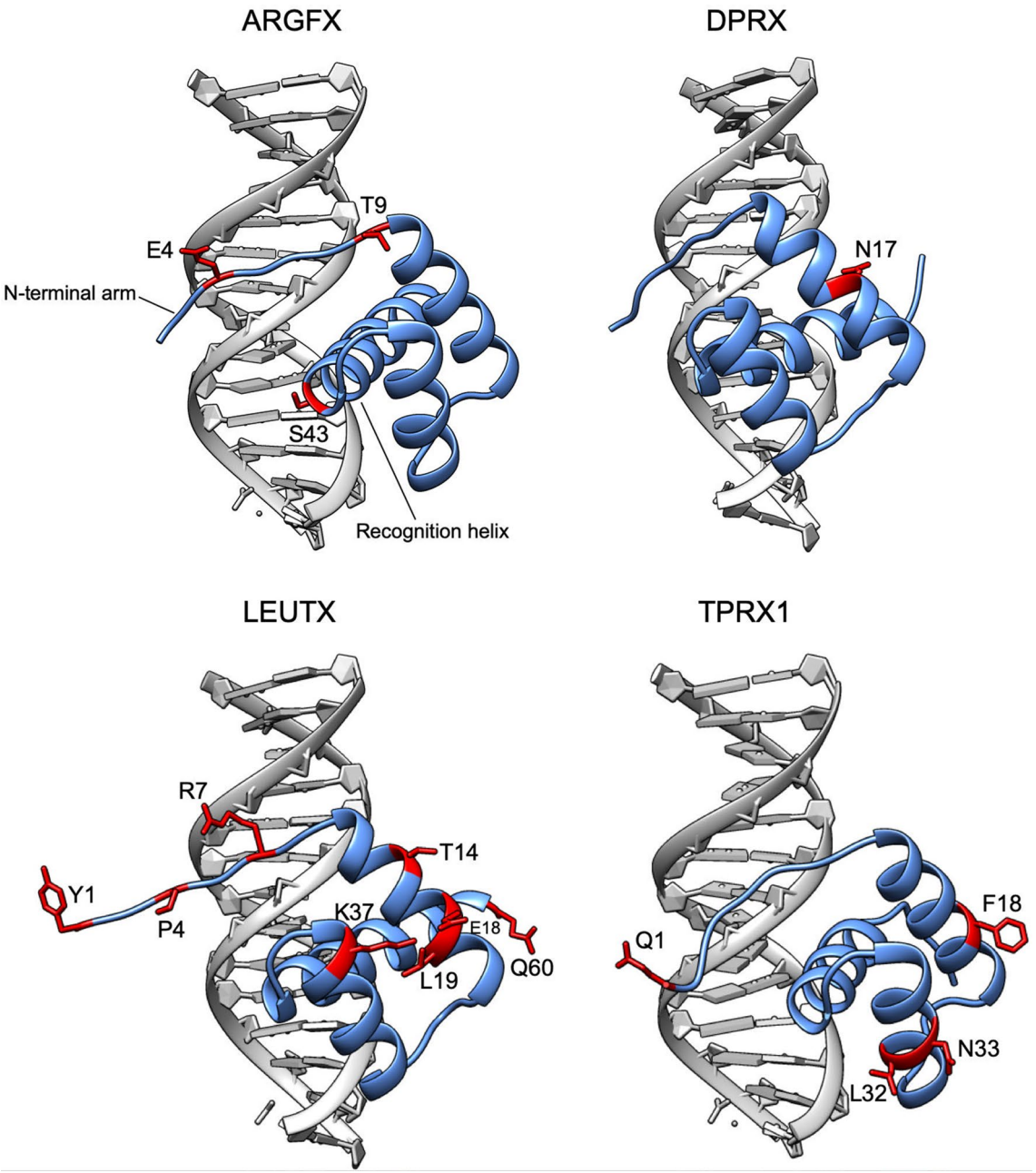

Fig. 7 Models of ETCHbox homeodomain structures including sites under positive selection. Homeodomains of human ETCHbox proteins (blue) are modelled in complex with DNA (grey). Residues under positive selection are coloured red. Amino acid side chains are shown for sites under positive selection only. 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.

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