 doi:10.1093/bfgp/els012

A genome-wide view of transcription factor gene diversity in chordate evolution: less gene loss in amphioxus?

Jordi Paps, Peter W.H. Holland and Sebastian M. Shimeld

Abstract

Previous studies of gene diversity in the homeobox superclass have shown that the Florida amphioxus *Branchiostoma floridae* has undergone remarkably little gene family loss. Here we use a combined BLAST and HMM search strategy to assess the family level diversity of four other transcription factor superclasses: the Paired/Pax genes, Tbx genes, Fox genes and Sox genes. We apply this across genomes from five chordate taxa, including *B. floridae* and *Ciona intestinalis*, plus two outgroup taxa. Our results show scattered gene family loss. However, as also found for homeobox genes, *B. floridae* has retained all ancient Pax, Tbx, Fox and Sox gene families that were present in the common ancestor of living chordates. We conclude that, at least in terms of transcription factor gene complexity, the genome of amphioxus has experienced remarkable stasis compared to the genomes of other chordates.

Keywords: *Homeobox; Fox; Pax; Sox; T-box*

INTRODUCTION

The loss and gain of genes are major forces in shaping genome evolution, with gene duplication as an important mode of gene gain. This is apparent from the observation that many genes have paralogues that is homologous gene copies in the same genome. Genes can duplicate in several different ways, for example by tandem and segmental duplications, retrotransposition and whole-genome duplication. Transcription factors make interesting case studies for analysing gene duplication, because their evolutionary history can often be traced with clarity. There are relatively few large groups (superclasses) of transcription factors, each characterized by a conserved DNA-binding domain (for example the homeodomain, HMG domain and forkhead domain) and these can be divided into clearly defined gene families which are often recognizable in quite divergent animal phyla. Such gene family identification is usually based on a high level of sequence similarity in the DNA-binding domain, which in turn suggests conservation of sequence-specific DNA recognition. In most transcription factor superclasses, the gene families can be inferred to have diversified by gene duplication early in animal evolution, often before the radiation of the bilaterian phyla. Consistent with this, gene families are generally (though not always consistently) defined as evolving from a single ancestral gene in the common ancestor of the Bilateria. Within individual animal phyla, these gene families often further diversified by additional gene duplication.

Comparison of genome sequences within and between animal phyla has revealed that different gene duplication processes have had different impacts on the genome-wide evolution of transcription factors, both in terms of the timing of the duplications and the subsequent divergence of genes. Tandem duplication seems to be an ongoing process [1]. Despite this, comparison of animal genomes from widely divergent phyla has suggested that tandem

Corresponding author. Sebastian M. Shimeld, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. Tel: +44 (0)1865 281994; Fax: +44 (0)1865 310447; E-mail: sebastian.shimeld@zoo.ox.ac.uk

Jordi Paps obtained his PhD in Genetics from the University of Barcelona and is currently a postdoctoral researcher at the Department of Zoology, University of Oxford. He has worked on phylogenetics of metazoans and opisthokont eukaryotes.

Peter W.H. Holland is Linacre Professor of Zoology, University of Oxford. His research focuses on how the evolution of animal diversity can be related to evolution of the genome. He has an obsession with homeobox genes.

Sebastian M. Shimeld is a Lecturer in the Department of Zoology, University of Oxford, and the Julian Huxley Fellow of Balliol College. His research focuses on the developmental and molecular basis for major transitions in animal evolution.

duplication was particularly important in shaping the diversity of transcription factor families early in animal evolution. Intriguingly, these early gene duplication events were often followed by 'asymmetrical' evolution of the daughter genes; that is, conservation of an ancient 'ancestral' gene, with divergence then fixation of the other copy or copies. Studies comparing such genes between choanoflagellate (the nearest single-celled relatives of the animals), sponge, cnidarian and bilaterian genomes show what appear to be 'bursts' of transcription factor family origin in the lineage leading to animals after the divergence of choanoflagellates, and in the lineage leading to cnidarians and bilaterians after the divergence of sponges [2]. The appearance of bursts of origin could be partly an artefact of our current perspective, coupled with the extinction of intermediate lineages, but comparison of cnidarian and bilaterian genomes certainly shows that many transcription factors families were clearly established by the time these two lineages separated. Generation of new transcription factor families does continue in individual lineages, but this seems to be at a reduced rate compared to early in animal evolution.

In some animal lineages, whole-genome duplication has also contributed to transcription factor gene evolution. This has been particularly well studied in the chordates, with two genome duplications in the early vertebrate lineage and a third in the teleost fish lineage. These ancient genome duplications are recognizable as they have left remnants in vertebrate genomes in the form of genome-wide paralagous gene organization, and they increased the membership of many transcription factor gene families.

A parallel process to gene duplication is gene loss. In contrast to the vast literature on gene duplication, far less attention has been paid to gene loss, even though it could have dramatic effects on shaping gene complements, genome architecture and phenotype. Genes could conceivably be lost through a variety of mutational mechanisms, from immediate loss (from the perspective of an individual) via deletion of a region of genome, to gradual loss over extended time via accumulation of small indels and single nucleotide changes. Both processes require that the loss of the gene is not sufficiently deleterious to prevent transmission to future generations, something that will depend on the interplay between drift and selection, and hence on the function of that gene. Gene loss is harder to recognize than gene gain with patchy data; however the development of whole-genome sequences for multiple chordate lineages has allowed loss to be assessed with reasonable clarity. The logic used to identify gene loss is simple, but crucially dependent on a known phylogeny. Consider three species, where species A and B are more closely related to each other, and C is the 'outgroup' to A plus B. If a gene is present in the genome of species A but not in B, then a gene loss in B can be inferred only when the gene is detected in species C.

Focusing on chordates in general, and amphioxus in particular, here we examine the extent of gene loss in the three chordate subphyla. We first recap briefly the well-studied homeobox superclass of transcription factor genes. Gene loss has previously been studied in this set of genes, and a remarkable pattern of differential loss has been uncovered. To assess whether this unusual pattern is unique to homeobox genes or a more general feature, we then examine patterns of gene loss in a selection of other transcription factor superclasses, by combining a comprehensive BLAST/HMM search strategy coupled with molecular phylogenetic methods.

HOMEOBOX GENE DIVERSITY IN THE CHORDATES

The homeobox genes are one of the best-studied superclasses of genes, with comprehensive accounts of gene family diversity reported from *Branchiostoma floridae (amphioxus), Ciona intestinalis* (urochordate), *Strongylocentrotus purpuratus* (sea urchin), *Homo sapiens* (human), *Mus musculus* (mouse) and several more taxa [3–7]. A dedicated database, HomeoDB, has also been established to support homeobox gene classification and nomenclature [8, 9]. Comparative genomic and molecular phylogenetic analyses suggest that at least 96 families of homeobox genes existed in the common ancestor of the Bilateria. Table 1 summarizes these data for the chordates, together with *S. purpuratus* and *Drosophila melanogaster*. Comparison of gene complements within the chordates and with outgroups has shown that *B. floridae* retains at least one gene in every homeobox gene family [4, 10]. The two other chordate lineages, however have each lost gene families, with an estimated 7 families lost in the lineage leading to urochordates plus vertebrates (Olfactores), 25 specifically lost in the urochordate lineage and 7 in the vertebrate lineage. Why all these specific gene losses should have occurred is unknown, although an attempt to investigate this has been made

Table 1: Summary of total homeobox gene number, number of Classes and number of families in the taxa analysed

Homeobox	*Drosophila melanogaster* [9]	*Strongylocentrotus purpuratus* [12]	*B. floridae* [9]	*C. intestinalis* [5]	*Danio rerio* [9]	*Mus musculus* [9]	*Homo sapiens* [9]
Number of genes	104	97	133	83	315 (7)	279 (45)	255 (78)
Classes	10	11	12	11	12	12	12
Families	81	72	108 (3)	70	103 (18)	99 (4)	103 (3)
ANTP	47	38	60	39	129 (3)	100 (2)	100 (19)
PRD	28	32	29	19	51 (2)	86 (31)	66 (32)
LIM	6	6	7	6	20	12	12
POU	5	4	7	3	19	16	16 (8)
HNF	0	2	4	2	6	3	3
SINE	3	3	3	3	13	6	6
TALE	8	6	9	7	29	22 (1)	20 (10)
CUT	3	1	4	1	9	7	7 (3)
PROS	1	1	1	2	3	2	2
ZF	2	3	5	1	17	14	14 (1)
CERS	1	1	1	?	3	5	5
Others	0	0	3	0	16 (2)	6 (11)	4 (5)

Numbers in curved brackets indicate genes that we were unable to classify. Numbers in square brackets indicate relevant references.

by examining expression of the homologues in amphioxus [11]. What is most striking is that *B. floridae* has retained such a comprehensive complement of ancient genes.

The observation of relative stasis in homeobox gene complement in amphioxus, compared to vertebrates and urochordates, is particularly interesting because amphioxus has often been considered to have a relatively ancestral body form for chordates, in that it appears to have maintained more of the primitive chordate features that were lost or elaborated in other chordates [10]. It has also been noted that amphioxus appears to retain relatively primitive genome organization in terms of the maintenance of synteny [13]. While the morphological and molecular stasis may be linked, whether this is cause or effect is unclear.

METHODS FOR ASSESSING THE DIVERSITY OF OTHER CHORDATE TRANSCRIPTION FACTOR SUPERCLASSES

Most animal transcription factors fall into a relatively small number of superclasses, defined by the type of DNA-binding domain. Examples other than the homeodomain include the forkhead domain, Paired domain, HMG domain, Tbx domain, bHLH domain, bZip domain, various zinc finger domains and several others. Some of these are relatively resistant to defining orthology and paralogy on a genome-wide scale due to the short size of the domain and/or high levels of sequence similarity. We therefore confined ourselves to several superclasses previously found to be amenable to such analyses, as we reasoned that these offered the best chance of definitively identifying gene loss. These are the Paired, Fox, and Tbx superclasses, and the Sox class within the HMG superclass.

We employed two methods to generate a comprehensive view of these superclasses. First we devised a semi-automated BLAST and HMM based search, which respectively used complete sequences from *D. melanogaster*, *M. musculus* and *H. sapiens* from Pfam [14] as BLAST queries and the Pfam HMM profiles for each superclass for the HMMER searches. The genomes queried were *Danio rerio, C. intestinalis, B. floridae* and *S. purpuratus*. The domains for the sequences were extracted and aligned with MAFFT [15]. The alignments were subjected to molecular phylogenetic analyses using the program RAxML [16]. We analysed 1000 bootstrap replicates and used the evolutionary model LG + Gamma + Invariant; this model was chosen because, in contrast with other models (WAG, JTT, etc.), it incorporates the variability of evolutionary rates across sites in the matrix estimation and is based in a much larger and diverse database than the ones used to estimate previous matrices [17]. The results of these studies are detailed in Tables 2–5, and summarized on Figure 1. The detailed molecular phylogenetic trees on which these summaries are

Table 2: Summary of Pax gene diversity in the genomes analysed

Pax	*Drosophila melanogaster* [9]	*Strongylocentrotus purpuratus* [12, 18]	*B. floridae* [4, 9]	*C. intestinalis* [5]	*Danio rerio* [9, 19]	*Mus musculus* [9]	*Homo sapiens* [9]
Number of genes	10	6	8	4	15	9	9
Families	6	5 (1)	5	4	4	4	4
Pax 2/5/8	1	1	1	1	4	3	3
Pax 3/7	3	0	1	1	4	2	2
Pax 4/6	3	1	2	1	4	2	2
Pax 1/9 (no HD)	1	1	1	1	3	2 [20]	2 [12]
Pox-neuro (no HD)	1	1	2	0	0	0	0
Eyegone	1 [21]	1 [21]	0	0	0	0	0
Others		Sp-paxC (SPU.00276)					

Shaded boxes indicate inferred gene family losses. Numbers in square brackets indicate relevant references. The row labelled 'others' is for genes that we were unable to classify.

Table 3: Summary of Tbx gene diversity in the genomes analysed

T-Box	*Drosophila melanogaster* [29]	*Strongylocentrotus purpuratus* [18, 30]	*Branchiostoma floridae* [31–33]	*Ciona intestinalis* [31, 32]	*Danio rerio* [19, 34]	*Mus musculus* [31, 35]	*Homo sapiens* [31, 35]
Number of genes	8	8	9	9	23	18	16
Families	5	6	8	7	8 (1)	8 (2)	8
Brachyury+Tbx19	1	1	2	1	1	2	2
Tbx1/10	1	1	1	1	1	2	2
Tbx 2/3	1	1	1	1	3	2	2
Tbx4/5	0	0	1	0	3	2	2
Tbx6/16	3 [36]	2	1	2 in [31] 4 in [35] 3 in our study	5 [37]	1	1
Tbx 15/18/22	0	0	1	1	3	3	3
Tbx20	2	2	1	1	1	1	1
Eosmesodermin/ Tbrain/Tbox21	0	1	1	1	5	3	3
Others					Tbx24	Tbx13 (Mm.Tbx7) Tbx14 (Mm.Tbx8)	

Shaded boxes indicate inferred gene family losses. Numbers in square brackets indicate relevant references. The row labelled 'others' is for genes that we were unable to classify.

based can be seen in the Supplementary Data. In parallel we utilized previous studies of chordate transcription factor family diversity based on whole-genome information. While few of these specifically address gene loss, and many do not include all three chordate subphyla, they represent a valuable resource of data due to the species- and gene family specific expertise of the authors, and also act as a test for our own studies. Below we address each gene group in turn.

The Paired domain superclass

The Paired genes (or Pax genes), many of which have important roles in tissue specification during embryonic development, are defined by possession of a 'Paired type' DNA-binding domain of 128 amino acids in the deduced protein. Some of the Pax genes also encode a homeodomain, and as a consequence this subset of Pax genes is also considered in homeobox gene classifications. However, several Pax genes lack a homeobox (e.g. Pax1/9), while of course most homeobox genes lack a Paired domain. Hence, the Pax superclass and the homeodomain superclass overlap. Here we examine the Pax genes in total. The Pax genes have been extensively studied in many animal phyla, defining at least six gene families, Pax1/9, Pax3/7, Pax4/6,

Table 4: Summary of Fox gene diversity in the genomes analysed

Fox	*Drosophila melanogaster* [53]	*Strongylocentrotus purpuratus* [18, 54]	*B. floridae* [40]	*C. intestinalis* [43]	*Danio rerio* [19, 55]	*Mus musculus* [55]	*Homo sapiens* [14, 55]
Number of genes	17	23	42	30	63	42	43
Families	12 (3)	20 (2)	22 (2)	19 (5)	21	20 (2)	20 (2)
FoxA	1	1	2	2	5	3	3
FoxB	2	1	1	1	2	2	2
FoxAB	0	1 [40]	1	0	0	0	0
FoxC	1	1	1	1	2	2	2
FoxD	1	1	1	2	6	4	6
FoxE	0	0	9	1	2	2	3
FoxF	1	1	1	1	2	2	2
FoxG	2	1	1	1	3	1	1
FoxH	0	0	1	2	3	2	1
FoxI	0	1	1	3	5	2	2
FoxJ1	0	1	1	0	3	1	1
FoxJ2/3	0	1	1	1	2	2	2
FoxK	1	1	2	1	2	2	2
FoxL1	0	1	1	0	3	1	1
FoxL2	0	1	2	1	1	1	1
FoxM	0	1	2	1	1	1	1
FoxN1/4	1	2	2	2	2	2	2
FoxN2/3	1	1	2	1	2	1	2
FoxO	1	1	2	1	7	4	4
FoxP	1	1	2	1	6	4	4
FoxQ1	0	1	1	1	3	1	1
FoxQ2	1	1	3	1	1	0	0
Others	d3F, fd19B and fd64A	SpFoxX SpFoxY	FoxIA FoxIB	FoxI to Fox5		MmFoxR1 MmFoxR2 FoxS	FOXR1 FOXR2 FoxS

Shaded boxes indicate inferred gene family losses. Numbers in square brackets indicate relevant references. The row labelled 'others' is for genes that we were unable to classify.

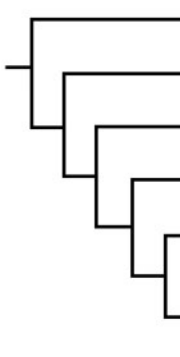
Pax2/5/8, eyegone and Pon, all present in the ancestor of deuterostomes.

Our analysis in the present study confirms this, and identifies at least one member for each of the five families in the *B. floridae* genome. This includes Pox-neuro (Pon gene family) and a previously described Pax1/9 gene [22], both of which lack a homeobox. However we did not identify an eyegone gene in *B. floridae*. In *C. intestinalis* we only found members of four families, with Pon and eyegone missing. This agrees with a previous assessment of *C. intestinalis* Pax gene diversity [5]. Pon and eyegone are also missing from our vertebrate genomes analyses, and have not been reported from any other vertebrate genome thus far [21]. When considering the outgroups, *D. melanogaster* has at least one member of all six families, while *S. purpuratus* has five families but appears to lack a Pax3/7 gene, as previously described [6].

To summarize, all chordates appear to have lost eyegone, however the Florida amphioxus *B. floridae* has retained a full complement of other Pax genes. Some lineages appear to have lost other Pax genes, including Pon in the lineage leading to urochordates and vertebrates. None have been lost specifically on the cephalochordate lineage.

The Tbx gene superclass

The Tbx (or T-box) genes encode a superclass of transcription factors that encode a Tbx type DNA-binding domain first identified in the Brachyury (or T) gene. Many Tbx genes have roles related to the development of limbs and heart. Molecular phylogenetic analyses across the Bilateria have identified a number of clear families, including Brachyury, Tbrain, Tbx20, Tbx15/18/22, Tbx1/10, Tbx4/5, Tbx2/3 and Tbx6/16. Several previous studies have focused on individual genes in *B. floridae* [23–26] and *C. intestinalis* [26, 27]. There has also been an analysis of the full diversity of Tbx genes in the genome of *C. intestinalis* [27]. Our genome-wide analyses confirm and extend the conclusions of



	Pax genes	Pax families	T-box genes	T-box families	Fox genes	Fox families	Sox genes	Sox families	Homeobox genes	Homeobox families	Homeobox classes
Drosophila	10	5	8	5	17	15	8	6	104	81	10
Strongylocentrotus	5	4	8	6	23	22	7	7	96	72	11
Branchiostoma	8	5	9	8	42	24	12	6	133	108 (3)	12
Ciona	4	4	9	7	30	24	7	7	83	70	11
Danio	15	4	23	9	61	22	20	7	315(7)	103 (18)	12
Mus	9	4	18	9	42	22	20	7	279(45)	99 (4)	12
Homo	9	4	16	8	43	22	20	7	255(78)	103 (3)	12

Figure 1: Phylogenetic tree of the species analysed in this study, with Pax, T-box, Fox, Sox and Homeobox genes and family number indicated. Fruit fly lost six Fox families (FoxAB, FoxE, FoxJ1, FoxJ2/3, FoxL2, FoxM), two T/box (Tbx 15/18/22 and Tbx4/5) and SoxH. Deuterostomes gained Tbrain. Sea urchin lost Pax3/7, FoxE, Tbx4/5 and Tbx15/18/22 and gained one Pax gene (Sp-paxC). Chordates gained FoxH, but lost eyegone. No gene family gain or loss can be detected in the lancelet. Olfactores lost Pox-neuro and FoxAB. The sea squirt lost two Fox genes (FoxL1, FoxJ1) and Tbx4/5. Zebrafish lost SoxH, and gained one Sox family (Sox32) and T-box one (tbx24). Mammals lost FoxQ2, but gained two Fox families (FoxR1 and FoxR2). Mice lost two T-box genes (Tbx13 and Tbx14).

these studies (Table 3). We identified clear members of all eight Tbx families in *B. floridae*. Vertebrates also possess at least one member of each family, while *C. intestinalis* is lacking a Tbx4/5 gene, as previously noted [27]. We also failed to find members of the Tbx4/5, Tbx15/18/22 and Tbrain families in *D. melanogaster*, and members of the Tbx4/5 and Tbx15/18/22 in *S. purpuratus* (Table 3). Both Tbx4/5 and Tbx15/18/22 have been previously described in the sea anemone *Nematostella vectensis* [28], indicating these are gene losses from *C. intestinalis*, *D. melanogaster* and *S. purpuratus*.

The Fox gene superclass

The Fox genes encode a forkhead DNA-binding domain and have been relatively well studied. They are involved in cell growth, proliferation and differentiation during embryonic development. Molecular phylogenetic analysis of (primarily) vertebrate Fox genes originally defined 19 families, FoxA to FoxS [38]. However, analyses involving additional invertebrate genes showed that two of these gene families, FoxR and FoxS, are derived from a vertebrate-specific duplication, while three gene families, FoxL, FoxJ and FoxQ should be split and an additional family, FoxAB, defined. Overall, 21 Fox gene families can be listed that existed in the common ancestor of the Bilateria: FoxAB, FoxA, FoxB, FoxC, FoxD, FoxE, FoxF, FoxG, FoxI, FoxJ1, FoxJ2/3, FoxK, LoxL1, FoxL2, FoxM, FoxN1/4, FoxN2/3, FoxO, FoxP, FoxQ1 and FoxQ2 [39–42].

Genome-wide analyses have previously been undertaken of the Fox genes in both *C. intestinalis* and *B. floridae* [40, 43]. In the former species, three families (FoxAB, FoxL1 and FoxJ1) appear to be missing, while in the latter species members of all 21 gene families were identified. Our searches (Table 4) confirmed this analysis, identifying members of every Fox gene family in the *B. Floridae* genome and of all families except FoxAB, FoxL1 and FoxJ1 in *C. intestinalis*. Since the 21 gene families were present in the common ancestor of bilaterians (and hence chordates) these must represent gene losses in the urochordate. One of these gene families, FoxAB, is also missing from vertebrate genomes and it is most parsimonious to conclude that this was lost in the stem lineage leading to urochordates plus vertebrates. FoxQ2 is absent from the two mammalian genome analysed, but is present in *D. rerio*. We conclude that this gene was lost separately in the urochordate and mammalian lineages.

We also failed to identify a member of the FoxE family in *S. purpuratus*, as previously reported [44]. FoxE genes have been indentified in more basal animals [2], so we infer that this is a gene loss. We also failed to find a FoxH gene in *S. purpuratus*, however as FoxH genes have not been identified conclusively outside the chordates, we cannot infer whether the lack of a FoxH gene in *S. purpuratus* represents

a secondary loss, or represents evolution of this gene family in the chordate lineage. Finally *D. melanogaster* appears to have lost many Fox genes. Two of the unclassified *D. melanogaster* Fox genes detected in our analysis have been previously proposed to belong to families FoxL1 and FoxQ2, respectively [39], however even with these re-assigned *D. melanogaster* has lost several Fox gene families.

The Sox gene class

The Sox genes are part of the HMG superclass, defined by possession of an HMG DNA-binding domain. Sox genes modulate various facets of development: while some are related to sex determination (SRY or SoxA), others play roles in neuronal development. Previous studies have divided HMG domain-containing genes into two groups: the Sox group genes (including TCF/LEF, BBX/HBP, Capicua and MATA) and the HMG/UBF group (including SSRP, mTFA and Polybromo) [45]. We focused on the former, and confined our analysis to the Sox genes themselves. Phylogenetic analyses have suggested several gene families in this group, named SoxA to SoxH. However, it should be noted that Sox gene nomenclature is often inconsistent with the definition of a gene family as primitively shared by all members of the Bilateria. For example, SoxA (also known as SRY) is the non-recombinant allele on the Y chromosome of a SoxB1 paralogue found only in placental mammals, while SoxG is also a paralogue of SoxB1 found only in vertebrates [46]. SoxH has been recently claimed to predate the bilaterian origin [47], while SoxB is split into two gene families, SoxB1 and SoxB2 [48]. Thus only the SoxB1, SoxB2, SoxC, SoxD, SoxE and SoxF gene families conform to the definition of each deriving from a single gene in the common ancestor of the Bilateria, however we also include SoxH as we can infer its origin predates the radiation of chordates [47].

B. floridae Sox genes have not been comprehensively classified, though a few individual genes have been described from this and other amphioxus species [49–51]. *C. intestinalis* Sox genes have been well-described, including orthologues of most families [52]. Our study identified SoxB1, SoxB2, SoxC, SoxD, SoxE and SoxF genes in all the species examined, including *B. floridae* (Table 5). Within the SoxB1 family, SoxA and SoxG genes were only identified respectively in vertebrates and mammals, as discussed above, while the SoxH family was identified in all species except *D. rerio* and *D. melanogaster*. Thus our analysis illustrates general stability of the pan-bilaterian Sox families through the species analysed. No family losses can be inferred, with the exception of the aforementioned SoxH in zebrafish and fruit fly [47], and some new families do appear to have evolved in the vertebrate lineage.

OVERVIEW AND CONCLUSIONS

Gene numbers

Figures 1 and 2 summarize our findings. Figure 1 illustrates total gene number in each class or superclass, and for the Pax, T-box, Fox and Sox genes also shows family number. In terms of raw gene number it is apparent that vertebrate genomes nearly always contain more members of a given group of transcription factor genes than do invertebrate genomes. This is true at the family level (Tables 2–5), class level (Table 2) and superclass level (Figure 1), though there are a few exceptions (for example, Pax genes in *D. melanogaster*). The extra vertebrate genes derive primarily from within-family expansion, largely due to genome duplications, though we again note some exceptions (for example, the evolution of the SoxA family within the vertebrate lineage). This is also not to say that within-family duplication is absent elsewhere; Tables 2–5 show many scattered instances of gene duplication in individual families within one lineage.

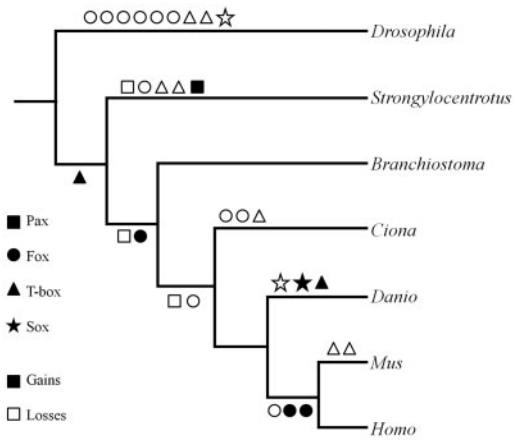


Figure 2: Phylogenetic tree of the species analysed in this study, with inferred gene family losses indicated. Filled symbols in the branches indicate gene family gains, while empty ones indicate losses.

Table 5: Summary of Sox gene diversity in the genomes analysed

Sox	*Drosophila melanogaster* [56]	*Strongylocentrotus purpuratus* [57]	*Branchiostoma floridae* [57]	*Ciona intestinalis* [52, 57]	*Danio rerio*	*Mus musculus* [14]	*Homo sapiens* [14]
Number of genes	8	7	13	7	20	20	20
Families	6	7	7	7	6 (1)	7	7
BI	1	1	3	1	2 (+2 SoxG [46])	3 (+SRY + SoxG)	3 (+SRY + SoxG)
B2	3	1	1	1	3	2	2
C	1	1	3	1	4	3	3
D	1	1	1	1	2	3	3
E	1	1	2	1	4	3	3
F	1	1	2	1	2	3	3
H (Sox30) [47]	0	1	1	1	0	1	1
Others					Sox32		

Shaded boxes indicate inferred gene family losses. Numbers in square brackets indicate relevant references. The row labelled 'others' is for genes that we were unable to classify.

Gene loss

The overarching goal of our study was to examine patterns of transcription factor gene loss within the chordates, and especially to compare the three chordate subphyla: cephalochordates (e.g. amphioxus *B. floridae*), urochordates (tunicates, e.g. *C. intestinalis*) and vertebrates. This study was prompted by the remarkable previous finding that, within the homeobox gene superclass, amphioxus has experienced no gene family loss at all since the common ancestor of chordates, in very marked contrast to urochordates and vertebrates. Could this pattern extend to other groups of genes? Has the amphioxus genome really experienced more stasis in the composition of transcription factor genes?

Our first data set focused on Pax genes, where we find that five gene families inferred to be present in the chordate ancestor are retained in the amphioxus genome. A sixth family, eyegone, may have been lost by all chordates. The urochordate and vertebrate genomes have also lost an additional gene family, Pon, and this is independent from the homeobox gene loss since the Pon gene lacks homeobox sequences. For Tbx genes, we inferred that eight gene families were present in the chordate ancestor. Again, all are retained by amphioxus. The urochordate has lost one, but vertebrates retain all. Moving to the larger Fox superclass, we infer there were 22 gene families in the chordate ancestor. Once again, amphioxus retains all of them. Urochordate and vertebrate genomes have experienced moderate gene family loss (three and one, respectively), although there is complexity in the picture with new gene families in vertebrates and loss in particular lineages. Finally, the Sox genes, with just six gene families in the chordate ancestor have been retained in all three chordate subphyla with the exception of SoxH in *D. rerio*.

The overall pattern of gene loss is plotted on a phylogeny of the chordates, plus outgroups, in Figure 2. For the Pax, Tbx, Fox and Sox genes, our analyses echo the previous conclusion for homeobox genes. In each of these cases, the Florida amphioxus *B. floridae* has retained every gene family inferred to have been present in the common ancestor of the Bilateria; the only exception to this is the Pax gene eyegone. The other chordate lineages have lost more genes in most cases, although of course the picture is further complicated by additional gene duplications in vertebrates. Amphioxus also shows relative stasis in other aspects of genome evolution, including a high level of conserved synteny with respect to vertebrates, while the urochordate lineage shows elevated rates of gene loss and genome reorganization [13, 58]. Detailed evolutionary studies of other gene families will be needed to see just how representative our study is of broader patterns of gene loss in the chordates. From current data we conclude that, at least in terms of transcription factor gene complexity, the genome of amphioxus has experienced remarkable stasis compared to the genomes of other chordates.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- The Homeobox gene superclass shows little if any gene loss in amphioxus.
- Genome searches and molecular phylogenetics show a similar pattern for Sox, Tbx and Fox genes.
- Only one gene in all of these groups has been lost from amphioxus; the Pax gene eyegone, also missing in all other chordates.
- This may reflect a more general relative stasis of the amphioxus genome compared to other chordates.

Acknowledgements

The authors thank Dr Nacho Maeso for his helpful comments on the analyses and article.

FUNDING

The programme Beatriu de Pinós of the Generalitat de Catalunya (2009 BP-DGR to J.P.); the European Research Council under EU FP7 ERC grant [(268513)11, to P.W.H.H].

References

1. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science* 2000;**290**:1151–5.
2. Larroux C, Luke GN, Koopman P, *et al*. Genesis and expansion of metazoan transcription factor gene classes. *Mol Biol Evol* 2008;**25**:980–96.
3. Holland PW, Booth HA, Bruford EA. Classification and nomenclature of all human homeobox genes. *BMC Biol* 2007;**5**:47.
4. Takatori N, Butts T, Candiani S, *et al*. Comprehensive survey and classification of homeobox genes in the genome of amphioxus, Branchiostoma floridae. *Dev Genes Evol* 2008;**218**:579–90.
5. Wada S, Tokuoka M, Shoguchi E, *et al*. A genomewide survey of developmentally relevant genes in Ciona intestinalis. II. Genes for homeobox transcription factors. *Dev Genes Evol* 2003;**213**:222–34.
6. Howard-Ashby M, Materna SC, Brown CT, *et al*. Identification and characterization of homeobox transcription factor genes in Strongylocentrotus purpuratus, and their expression in embryonic development. *Dev Biol* 2006;**300**:74–89.
7. Zhong YF, Holland PW. The dynamics of vertebrate homeobox gene evolution: gain and loss of genes in mouse and human lineages. *BMC Evol Biol* 2011;**11**:169.
8. Zhong YF, Butts T, Holland PW. HomeoDB: a database of homeobox gene diversity. *Evol Dev* 2008;**10**:516–8.
9. Zhong Y-F, Holland PWH. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evol Dev* 2011;**13**:567–8.
10. Holland PWH. From genomes to morphology: a view from amphioxus. *Acta Zool* 2010;**91**:81–6.
11. Butts T, Holland PWH, Ferrier DEK. Ancient homeobox gene loss and the evolution of chordate brain and pharynx development: deductions from amphioxus gene expression. *Proc R Soc B: Biol Sci* 2010;**277**:3381–3389.
12. Howard-Ashby M, Materna SC, Brown CT, *et al*. Identification and characterization of homeobox transcription factor genes in Strongylocentrotus purpuratus, and their expression in embryonic development. *Dev Biol* 2006;**300**:74–89.
13. Putnam NH, Butts T, Ferrier DE, *et al*. The amphioxus genome and the evolution of the chordate karyotype. *Nature* 2008;**453**:1064–71.
14. Finn RD, Mistry J, Tate J, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2010;**38**:D211–22.
15. Katoh K, Misawa K, Kuma Ki, *et al*. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;**30**:3059–66.
16. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006;**22**:2688–90.
17. Le SQ, Gascuel O. An improved general amino acid replacement matrix. *Mol Biol Evol* 2008;**25**:1307–20.
18. Cameron RA, Samanta M, Yuan A, *et al*. SpBase: the sea urchin genome database and web site. *Nucleic Acids Res* 2009;**37**:D750–4.
19. Flicek P, Amode MR, Barrell D, *et al*. Ensembl 2011. *Nucleic Acids Res* 2011;**39**:D800–6.
20. Tremblay P, Gruss P. Pax: genes for mice and men. *Pharmacol Therapeut* 1994;**61**:205–26.
21. Friedrich M, Caravas J. New insights from hemichordate genomes: prebilaterian origin and parallel modifications in the paired domain of the Pax gene eyegone. *J Exp Zool B Mol Dev Evol* 2011;**316**:387–92.
22. Holland ND, Holland LZ, Kozmik Z. An amphioxus Pax gene, AmphiPax-1, expressed in embryonic endoderm, but not in mesoderm: implications for the evolution of class I paired box genes. *Mol Mar Biol Biotechnol* 1995;**4**:206–14.
23. Horton AC, Gibson-Brown JJ. Evolution of developmental functions by the Eomesodermin, T-brain-1, Tbx21 subfamily of T-box genes: insights from amphioxus. *J Exp Zool* 2002;**294**:112–21.
24. Horton AC, Mahadevan NR, Minguillon C, *et al*. Conservation of linkage and evolution of developmental function within the Tbx2/3/4/5 subfamily of T-box genes: implications for the origin of vertebrate limbs. *Dev Genes Evol* 2008;**218**:613–28.
25. Mahadevan NR, Horton AC, Gibson-Brown JJ. Developmental expression of the amphioxus Tbx1/10 gene illuminates the evolution of vertebrate branchial arches and sclerotome. *Dev Genes Evol* 2004;**214**:559–66.
26. Minguillon C, Logan M. The comparative genomics of T-box genes. *Brief Funct Genomic Proteomic* 2003;**2**:224–33.
27. Takatori N, Hotta K, Mochizuki Y, *et al*. T-box genes in the ascidian Ciona intestinalis: characterization of cDNAs and spatial expression. *Dev Dyn* 2004;**230**:743–53.

28. Yamada A, Pang K, Martindale MQ, *et al*. Surprisingly complex T-box gene complement in diploblastic metazoans. *Evol Dev* 2007;**9**:220–30.
29. Berns N, Kusch T, Schröder R, *et al*. Expression, function and regulation of Brachyenteron in the short germband insect Tribolium castaneum. *Dev Genes Evol* 2008;**218**: 169–79.
30. Wei Z, Angerer RC, Angerer LM. A database of mRNA expression patterns for the sea urchin embryo. *Dev Biol* 2006;**300**:476–84.
31. Minguillon C, Logan M. The comparative genomics of T-box genes. *Brief Funct Genomic Proteomic* 2003;**2**:224–33.
32. Takatori N, Hotta K, Mochizuki Y, *et al*. T-box genes in the ascidian Ciona intestinalis: characterization of cDNAs and spatial expression. *Dev Dynam* 2004;**230**:743–53.
33. Ruvinsky I, Silver LM, Gibson-Brown JJ. Phylogenetic analysis of T-box genes demonstrates the importance of amphioxus for understanding evolution of the vertebrate genome. *Genetics* 2000;**156**:1249–57.
34. Ruvinsky I, Oates AC, Silver LM, *et al*. The evolution of paired appendages in vertebrates: T-box genes in the zebrafish. *Dev Genes Evol* 2000;**210**:82–91.
35. Wilson V, Conlon F. The T-box family. *Genome Biol* 2002; **3**:3008.3001–7.
36. Reim I, Frasch M. The Dorsocross T-box genes are key components of the regulatory network controlling early cardiogenesis in Drosophila. *Development* 2005;**132**:4911–25.
37. Lardelli M. The evolutionary relationships of zebrafish genes tbx6, tbx16/spadetail and mga. *Dev Genes Evol* 2003;**213**: 519–22.
38. Kaestner KH, Knochel W, Martinez DE. Unified nomenclature for the winged helix/forkhead transcription factors. *Genes Dev* 2000;**14**:142–6.
39. Mazet F, Yu JK, Liberles DA, *et al*. Phylogenetic relationships of the Fox (Forkhead) gene family in the Bilateria. *Gene* 2003;**316**:79–89.
40. Yu JK, Mazet F, Chen YT, *et al*. The Fox genes of Branchiostoma floridae. *Dev Genes Evol* 2008;**218**:629–38.
41. Shimeld SM, Degnan B, Luke GN. Evolutionary genomics of the fox genes: origin of gene families and the ancestry of gene clusters. *Genomics* 2010;**85**:256–60.
42. Wotton KR, Shimeld SM. Comparative genomics of vertebrate fox cluster loci. *BMC Genomics* 2006;**7**:271–8.
43. Yagi K, Satou Y, Mazet F, *et al*. A genomewide survey of developmentally relevant genes in Ciona intestinalis. III. Genes for Fox, ETS, nuclear receptors and NFkappaB. *Dev Genes Evol* 2003;**213**:235–44.
44. Tu Q, Brown CT, Davidson EH, *et al*. Sea urchin Forkhead gene family: phylogeny and embryonic expression. *Dev Biol* 2006;**300**:49–62.
45. Soullier S, Jay P, Poulat F, *et al*. Diversification pattern of the HMG and SOX family members during evolution. *J Mol Evol* 1999;**48**:517–27.
46. Okuda Y, Yoda H, Uchikawa M, *et al*. Comparative genomic and expression analysis of group B1 sox genes in zebrafish indicates their diversification during vertebrate evolution. *Dev Dynam* 2006;**235**:811–25.
47. Han F, Wang Z, Wu F, *et al*. Characterization, phylogeny, alternative splicing and expression of Sox30 gene. *BMC Mol Biol* 2010;**11**:98.
48. Royo JL, Maeso I, Irimia M, *et al*. Transphyletic conservation of developmental regulatory state in animal evolution. *Proc Natl Acad Sci USA* 2011;**108**:14186–91.
49. Holland LZ, Schubert M, Holland ND, *et al*. Evolutionary conservation of the presumptive neural plate markers AmphiSox1/2/3 and AmphiNeurogenin in the invertebrate chordate amphioxus. *Dev Biol* 2000;**226**:18–33.
50. Lin Y, Chen D, Fan Q, *et al*. Characterization of SoxB2 and SoxC genes in amphioxus (Branchiostoma belcheri): implications for their evolutionary conservation. *Sci China C Life Sci* 2009;**52**:813–22.
51. Meulemans D, Bronner-Fraser M. The amphioxus SoxB family: implications for the evolution of vertebrate placodes. *Int J Biol Sci* 2007;**3**:356–64.
52. Yamada L, Kobayashi K, Degnan B, *et al*. A genomewide survey of developmentally relevant genes in Ciona intestinalis. IV. Genes for HMG transcriptional regulators, bZip and GATA/Gli/Zic/Snail. *Dev Genes Evol* 2003;**213**: 245–53.
53. Lee H-H, Frasch M. Survey of forkhead domain encoding genes in the Drosophila genome: classification and embryonic expression patterns. *Dev Dynam* 2004;**229**:357–66.
54. Tu Q, Brown CT, Davidson EH, *et al*. Sea urchin forkhead gene family: phylogeny and embryonic expression. *Dev Biol* 2006;**300**:49–62.
55. Kaestner KH, Knöchel W, Martínez DE. Unified nomenclature for the winged helix/forkhead transcription factors. *Genes Dev* 2000;**14**:142–6.
56. Guth S, Wegner M. Having it both ways: sox protein function between conservation and innovation. *Cell Mol Life Sci* 2008;**65**:3000–18.
57. Phochanukul N, Russell S. No backbone but lots of sox: invertebrate sox genes. *Int J Biochem Cell Biol* 2010;**42**: 453–64.
58. Hughes AL, Friedman R. Loss of ancestral genes in the genomic evolution of Ciona intestinalis. *Evol Dev* 2005;**7**: 196–200.