BMC Genomics

BioMed Central

Research article

Open Access

TOX defines a conserved subfamily of HMG-box proteins

Emmett O'Flaherty and Jonathan Kaye*

Address: Department of Immunology, The Scripps Research Institute, 10550 North Torrey Pines Rd., La Jolla, CA92037, USA

Email: Emmett O'Flaherty - emmett@scripps.edu; Jonathan Kaye* - jkaye@scripps.edu

* Corresponding author

Published: 2 April 2003

BMC Genomics 2003, **4**:13

Received: 10 January 2003

Accepted: 2 April 2003

This article is available from: http://www.biomedcentral.com/1471-2164/4/13

Abstract

Background: HMG-box proteins are a large and diverse superfamily of architectural factors that share one or more copies of a sequence- and structurally-related DNA binding domain. These proteins can modify chromatin structure by bending and unwinding DNA. HMG-box proteins can be divided into two subfamilies based on whether they recognize DNA in a sequence-dependent or sequence-independent manner. We recently identified an HMG-box protein involved in T cell development, designated TOX, which is highly conserved in humans and mice.

Results: We show here that based on sequence alignment, TOX best fits into the sequence-independent HMG-box family. Three other human and murine predicted proteins are identified that share a common HMG-box domain with TOX, as well as other features. The gene encoding one of these additional family members has a distinct but overlapping pattern of tissue expression when compared to TOX. In addition, we identify genes encoding predicted TOX HMG-box subfamily members in pufferfish and mosquito.

Conclusions: We have identified a novel subfamily of HMG-box proteins that is related to the recently described TOX protein. The highly conserved nature of the TOX family of proteins in humans and mice and differences in the pattern of expression between family members suggest non-overlapping functions of individual proteins. In addition, our data suggest that the TOX subtype of HMG-box domain first appeared in invertebrates, was duplicated in early vertebrates and likely took on new functions in mammalian species.

Background

Regulation of DNA-dependent processes such as transcription, replication, and strand repair requires bending and unwinding of compacted chromatin structure. Many of these structural changes are mediated by high mobility group (HMG) proteins, a diverse superfamily of non-histone chromosomal proteins that were originally classified by their electrophoretic mobility [1]. HMG proteins contain DNA-binding domains that allow them to produce specific changes in target DNA structure [2,3]. Three structurally distinct classes of HMG proteins have been defined; the HMG-nucleosomal binding family, the HMG-AT-hook family, and the HMG-box family (whose canonical members are referred to as HMGN, HMGA and HMGB respectively) [4]. In addition, a large number of proteins have been found that contain HMG protein related motifs.

All members of the HMG-box family possess a 70–80 amino acid DNA-binding domain (the HMG-box) related to a motif originally identified in HMGB1 [5]. The HMG-box may also be involved in protein-protein interactions. For example, RAG1 has been reported to interact with the tandem HMG-boxes of HMGB1 or HMGB2 [6]. Where it

has been studied, HMG-boxes have been shown to be structurally related, forming three α-helices in a characteristic L-shaped structure [3,7]. The distortion of DNA by the HMG-box is primarily mediated through contacts with the minor groove of the DNA helix, potentially allowing simultaneous binding of other transcriptional regulators to the DNA [8,9]. The HMG-box family has often been subdivided based on DNA binding properties. One group of HMG-box proteins recognizes structural features of DNA with low or absent sequence specificity. These proteins have broad tissue distribution, and typically contain multiple HMG-box motifs. The second group of HMG-box proteins recognizes DNA in a sequence-specific manner akin to more traditional transcription factors. These proteins have a restricted pattern of expression and typically contain only one HMG-box [10]. Members of both groups of HMG-box proteins also bind altered nucleic acid structures such as 4-way junctions [11–13] and cisplatinated DNA [14].

HMG-box family proteins are found in a variety of eukaryotic organisms and most are known or suspected regulators of gene expression. We recently identified a gene designated *Tox* (for thymocyte selection-associated HMG-box gene), encoding a novel nuclear protein that shows a highly regulated pattern of expression during thymocyte differentiation. The TOX protein is 526 amino acids with an acidic N-terminal domain, a bipartite nuclear localization signal sequence and a single centrally located HMG-box motif [15]. Forced expression of TOX in the thymus of transgenic mice leads to changes in the differentiation program of developing T cells. These studies led us to propose that TOX is involved in regulating gene expression during critical developmental checkpoints in the thymus, and possibly elsewhere in the immune system. For example, TOX may also be involved in germinal center B lymphocyte development and/or function [16].

Several other HMG-box proteins also play important roles in lymphocyte development [17]. Mice doubly deficient in lymphocyte enhancement factor-1 (LEF-1) and T cell factor-1 (TCF-1) have a complete block in development of T cells [18,19], while deficiency of SRY-box containing protein 4 (SOX-4) in mice results in a lack of pro-B cell expansion and mild perturbation of thymocyte development [20]. Unlike these HMG-box proteins, however, we find that the HMG-box of TOX is more closely related to the DNA binding domain of sequence-independent HMG-box proteins. In addition, three other predicted proteins share almost identical HMG-box domains with TOX, defining a new subfamily of HMG-box proteins. The TOX subfamily is highly conserved in mice and humans, and is distributed on four separate chromosomes. Outside the HMG-box domain, these proteins are less well conserved, suggesting that they may have non-overlapping functions. Our analysis also suggests that two exons encoding the HMG-box domain may be the evolutionary unit of the TOX subfamily, found as a single copy in an invertebrate and replicated in early vertebrates. This expanded TOX subfamily likely took on new functions, including specific roles in the mammalian immune system.

Results and Discussion

TOX resembles sequence-independent HMG-box family members

Although there are few amino acid positions within the HMG-box motif that are conserved throughout the HMG-box protein superfamily, there are highly conserved residues within sequence-dependent and sequence-independent subgroups [21]. The vast majority of HMG-boxes have a tryptophan in helix 2, often as part of a GXXW motif (where X denotes any amino acid) [3,5]. Similarly, TOX has a tryptophan at this position, although as part of the less commonly found AXXW sequence. Figure 1 shows an alignment of the HMG-box region of TOX with HMG-box motifs found in seven other proteins representing the two major subfamilies of HMG-box proteins. Human SRY, murine SOX-4, SOX-17, and LEF-1 contain single HMG-box motifs and bind DNA in a sequence-specific fashion. Murine HMGB1 and yeast NHP6A bind DNA in a sequence-independent fashion. UBF-1 specifically binds the ribosomal RNA gene promoter, although no particular recognition sequence can be identified [22]. HMGB1 and UBF-1 contain multiple HMG-box motifs, although only one is shown (indicated by decimal in Fig. 1).

Positions 5, 10, 65 and 70 (as numbered in Fig. 1) are included in regions that have previously been shown together to be sufficient to confer sequence specificity to an HMG-box protein [23]. Positions 5 and 10 are proline and serine, respectively, in the sequence-independent DNA binders as well as in TOX. In contrast, the sequence specific HMG-box proteins have hydrophobic (V or I) and asparagine residues present at positions 5 and 10, respectively. The asparagine at position 10 makes contact with DNA in the hSRY-DNA complex [24], as does the serine at this position in the NHP6A-DNA complex [7]. A hydrophobic residue at position 32 that can partially intercalate into DNA plays a role in DNA binding of some sequence-independent HMG-box proteins [7,25,26]. Position 32 is the hydrophobic residue phenylalanine in TOX, in contrast to the polar residues found at this position in SRY, SOX proteins and LEF-1. Similarly, TOX is more closely aligned with the sequence-independent HMG-box proteins than the sequence-specific transcription factors at position 65. The tyrosine at this position (shared with TOX) is a DNA contact in the NHP6A-DNA complex [7]. Moreover, there is a conserved proline at position 70 of the sequence-specific DNA binding proteins that truncates the third alpha helix of the HMG-box structure allowing

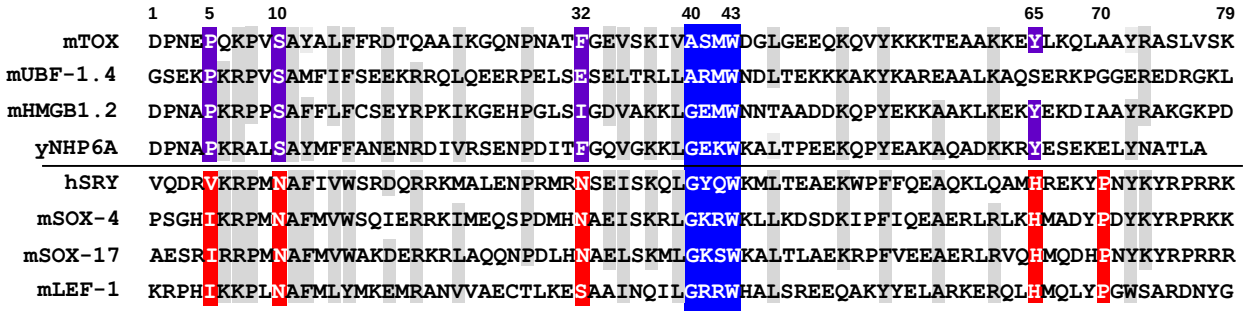

Figure 1

Comparison of HMG-box sequences. The HMG-box region of TOX is aligned with the HMG-box motifs found in seven other proteins representing the two major subfamilies of HMG-box proteins. Similarities between the HMG-boxes of both groups of proteins are highlighted in gray and include matches to the consensus HMG-box domain (accession number pf00505) defined in the Pfam protein families' database. Residues in red and purple distinguish these two subgroups of HMG-box proteins and are discussed in the text. In addition, the consensus sequence GXXW (or more rarely AXXW, as also found in TOX) found commonly in HMG-boxes is shown in blue.

additional DNA contacts [24,27,28]. This proline is absent from TOX and the sequence-independent HMG-box proteins. Based on these observations, TOX appears to best fit into the family of structure-dependent but sequence independent HMG-box DNA binding proteins. However, this has yet to be determined empirically by DNA binding studies.

TOX is a member of a four-protein subfamily with nearly identical HMG-box sequences

Tox is the murine homologue of the human gene KIAA0808, previously isolated from brain tissue [29]. TOX and KIAA0808 are approximately 90% identical at both the nucleotide and amino acid level [15]. Using BLAST program searches [30,31] we have now identified 3 other murine genes that are predicted to encode proteins that have nearly identical HMG-box sequences to TOX (Fig. 2A,2B). Human homologues of the three additional murine genes have also been identified (Fig. 2A,2B). A full length murine cDNA representing Langerhans cell protein 1 (LCP1) was originally cloned from maturing epidermal Langerhans cells, while its human homologue KIAA0737 was cloned from human brain [29] as was a cDNA encoding TNRC9 (also known as CAGF9) [32]. Searches of the EST database confirm that all family members are expressed at the level of mRNA, and we have shown that KIAA0808 is expressed at the level of protein using a cross-reactive anti-TOX antisera (data not shown). *Tox* gene family members are found on four different chromosomes in human (Fig. 2A). Similarly, the murine *Tox* gene family is distributed across four different chromosomes, in regions of synteny conserved between mouse and human (Fig. 2A).

All murine and human TOX subfamily members have a single centrally located HMG-box (Fig. 2A,2B). Within the first 70 amino acids of the HMG-box region only six positions show any variation and of those six, three are shared between all family members except TOX/ KIAA0808 homologues while the remaining three changes are found in the LOC241768/ C20ORF100 pair. Despite high similarity in overall structure between HMG-boxes, individual domains exhibit distinct DNA binding characteristics as discussed above, and can interact with DNA in different orientations. The high degree of conservation of the particular TOX-type HMG-box sequence is therefore likely to reflect a specific DNA and/or protein interaction that is important for function of this protein subfamily.

There is also complete identity of the HMG-boxes between murine and human homologues of a given family member, consistent with formation of the *Tox* subfamily by gene duplication prior to the rodentia and primate split. This is further supported by the genomic organization of the genes. The *Tox* gene has 9 exons spanning more than 300 kb (data not shown). The HMG-box region of TOX is encoded by two exons, as is found in some SOX genes [33,34], although the exact intron break is unique to TOX. Interestingly, the position of the intron/exon boundary encoding the HMG-box is precisely conserved among all family members in mouse and human (Fig. 2B). In general, the loss or gain of an intron represents a major genetic rearrangement and is a relatively rare event when compared with sequence changes [35]. It is highly unlikely that an intron would arise at the same sequence position in different gene lineages, further supporting that the TOX family arose by gene duplication.

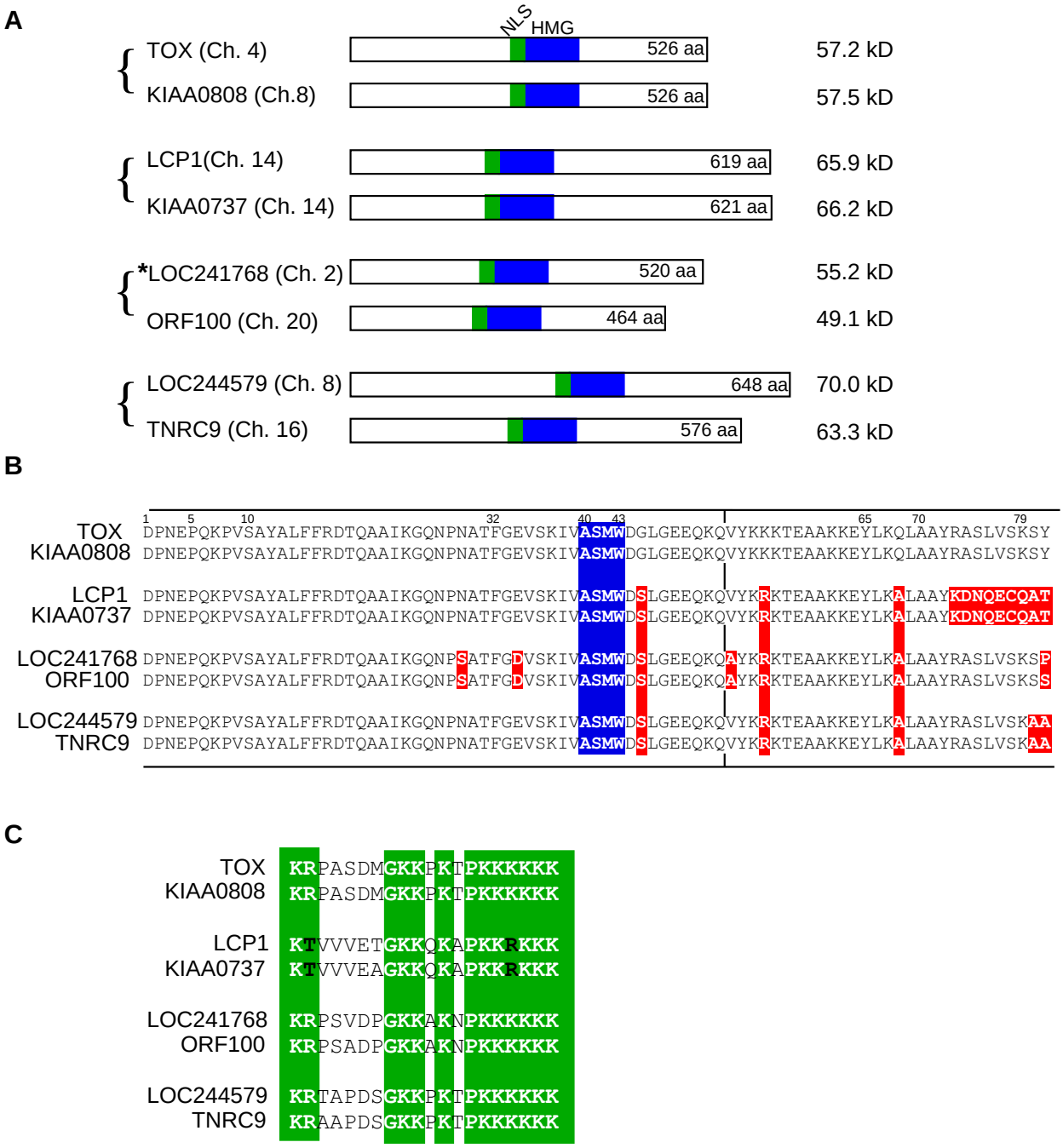

Figure 2

Murine TOX subfamily members and their human homologues. (**A**) The HMG-box and putative NLS regions within each predicted protein are shown in blue and green, respectively. The predicted size and molecular weight of each protein and the chromosomal location of the gene that encodes the protein are also indicated. *Note that the predicted protein LOC241768 as it currently exists in the NCBI database has been modified, and there is some question as to chromosomal location (see Methods). (**B**) Pair-wise comparison of HMG-box domains of murine and human TOX subfamily members. The upper predicted protein of each pair is mouse derived. Residues that differ from TOX are highlighted in red and the AXXW motif is highlighted in blue. The vertical line represents the position of the conserved exon boundary of the respective genes. (**C**) Comparison of predicted NLS of murine and human TOX subfamily members. The consensus motif is highlighted in green.

In addition to the box itself, the TOX family members contain conserved lysine residues at the NH_2-terminus of the HMG-box motif, imbedded in the consensus sequence [KR$(X)_5$GKKXKXPKKKKK] (Fig. 2C). This sequence is the likely nuclear localization signal (NLS) for these proteins [15], similar to the bipartite NLS found in nucleoplasmin [36]. The one exception is the Langerhans cell protein-1 (LCP1)/KIAA0737 pair, which lacks the second basic residue of the motif. Nuclear localization signals are also highly conserved in SRY-box and other HMG-box proteins [37]. Basic residues outside of the HMG box motif may also function to stabilize DNA binding by these proteins [38].

Outside of the HMG-box and NLS, the murine and human homologues of the three additional pairs of proteins are highly conserved, as we reported for TOX and KIAA0808 [15]. LCP1 and KIAA0737 proteins are approximately 94% identical, while LOC241768 and C20ORF100 share 85% identity. Comparison of LOC244759 and TNRC9 revealed that these proteins were highly similar, with the exception of the C-termini. The C-terminal domain of TNRC9 contains an expansion of trinucleotide repeats coding for glutamine. Long polyglutamine repeats have been found in a number of transcription factors [39]. Notably, the C-terminus of murine HMG-box protein SRY contains a large polyglutamine repeat region that is responsible for the sex-determining function of this protein [40]. However, a number of neurodegenerative disorders, such as Huntington disease and spinobulbar muscular atrophy, are also strongly associated with proteins containing a polyglutamine stretch [41]. Conformational change and protein misfolding in the expanded polyglutamine region is believed to be the molecular basis of the pathogenesis [42]. Whether the presence of polyglutamine repeats in the C-terminus of human TNRC9 is important for function or alternatively hinders its ability to form a functional protein remains to be determined.

TOX family members are approximately 20–30% identical outside the NLS/HMG-box region (Table 1). A detailed alignment of TOX and its close relative LCP1 is shown in Fig. 3A. The HMG-box splits TOX family proteins into two domains of approximately 200–300 amino acids. The N-terminal domains (exclusive of the NLS and HMG-box) are acidic for all human and murine TOX family proteins (pI range of 4–6). In the context of the nucleus, acidic domains are often protein-protein interaction domains. The N-terminal regions are also more similar than C-terminal regions when comparing different TOX subfamily members (for example, Fig. 3A). The C-terminal regions tend to be enriched for the presence of proline and to a lesser extent glutamine residues (data not shown). The C-terminal domains of TOX/ KIAA0808 and LOC244579/ TNRC9 are mildly to strongly basic respectively, while those of LCP1 and KIAA0737 are acidic. Despite overall similarity, the C-terminal domain of LOC241768 is acidic while its human homologue is basic.

Table 1: Amino acid sequence comparison of murine TOX subfamily member proteins outside the HMG-box and NLS regions.

	TOX	LCP1	LOC241768	LOC244579
TOX	100*	28	32	31
LCP1		100	18	25
LOC241768			100	13
LOC244579				100

* Percent identity of the two proteins outside the HMG-box and NLS regions, using an unfiltered BLAST program comparison. Comparisons are always made in relation to the smaller protein of the pair. In all comparisons performed, Expect (E) values were 7e-14 or lower.

The Tox *and LCP1 genes have unique patterns of expression*

Unlike other sequence-independent HMG-box proteins, which show ubiquitous tissue expression, we have previously shown that TOX has a highly regulated tissue and developmental expression profile. The Tox gene is most abundantly expressed in the thymus followed by the liver and brain; it is poorly expressed or absent in other tissues, including heart, kidney, lung, muscle, skin, intestine, spleen stomach and testis [15]. LCP1 is also expressed in thymus, although whether expression is regulated during different developmental stages remains to be determined (Fig. 3B). Unlike Tox, however, the LCP1 gene is most highly expressed in testis (Fig. 3B). In addition, LCP1 is more highly expressed than Tox in skin, consistent with isolation of LCP1 from Langerhans cells. LCP1 is also expressed in most other tissues tested (with the possible exception of muscle) and thus has a more widespread pattern of expression than previously found for the Tox gene.

The TOX subfamily is conserved in early vertebrates and mosquito

It has been suggested that at various times throughout metazoan evolution, HMG-box containing sequences duplicated, in each case leaving one redundant copy, which was free to evolve a new function or be lost from the genome [33]. The spare HMG-box-containing fragment recruited preexisting functional domains and formed mosaic proteins capable of rapidly taking on novel function. In this study, we examined the evolutionary history of the TOX HMG domain. As discussed above, data suggest that duplication of genes encoding the TOX-like HMG-box occurred prior to mammalian radiation.

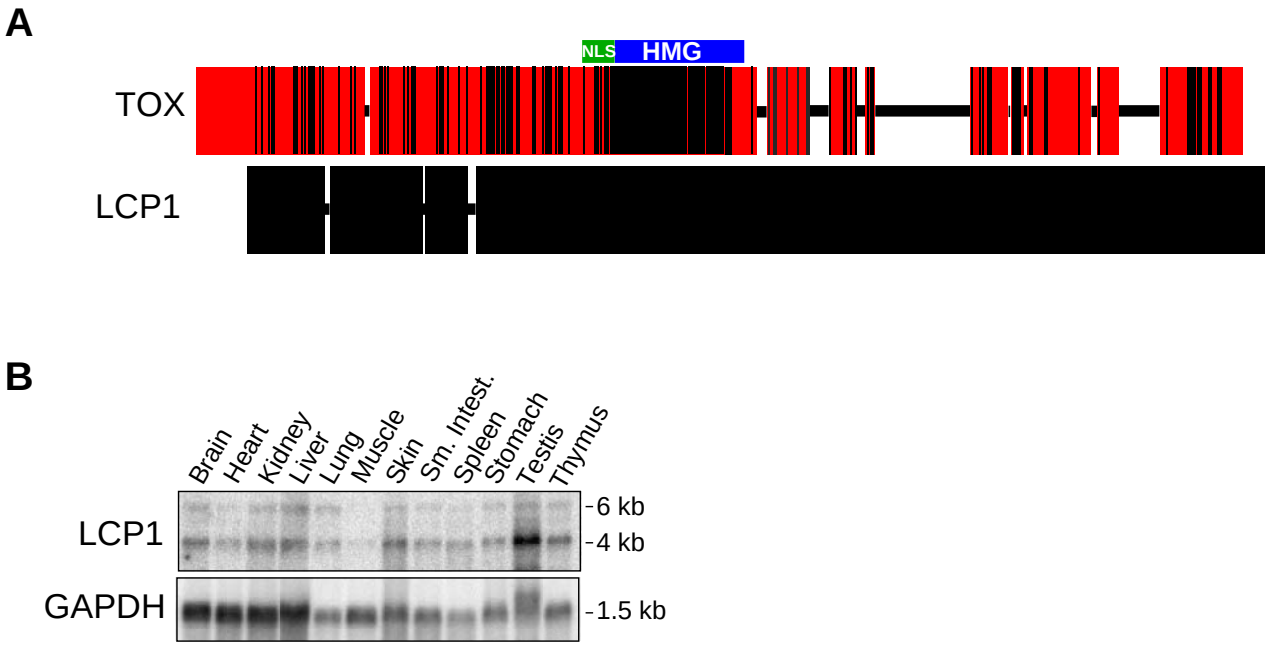

Figure 3
Comparison of TOX and LCP1. **(A)** Graphical representation of an unfiltered BLAST alignment comparing TOX and LCP1 proteins. Breaks in the proteins to allow maximum alignment are represented by a line. Residues in TOX that are identical to the aligned LCP1 protein are shown in black, while differences are shown in red. The NLS and HMG-box regions are indicated by green and blue bars, respectively. (**B**) Northern analysis of LCP1 gene expression in normalized poly-A^+ RNA isolated from various tissues. The signal obtained from a probe to the housekeeping GAPDH gene is also shown as a loading control.

Analysis of the genome of Fugu rubripes (pufferfish) revealed that this duplication was even more ancient.

It has been shown that although the *Fugu rubripes* genome is only one-eighth the size of the human genome, it contains a comparable complement of protein-encoding genes [43]. We found five sequences on different scaffolds in the pufferfish genome that encode predicted proteins containing a TOX-like HMG-box (Fig. 4). The genes present on Scaffold-182 and Scaffold-2474 both encode 500 amino acid proteins. The gene present on Scaffold-16 encodes a 365 amino acid protein, while the Scaffold-5698 gene encodes a protein of 159 amino acids. The Fugu HMG-box sequences are nearly identical to mammalian TOX HMG-boxes, and the exon/intron boundary encoding the box is at the characteristic position (Fig. 4). It has previously been shown that the intron-exon structure of most genes is conserved between Fugu and human [44]. We also found an incomplete HMG-box encoded by a gene on Scaffold 60. We could find no additional exons to complete coding of the box, but whether additional coding sequence is present, the sequence is a pseudogene, or the gene encodes a truncated protein is unclear. In addition, the Fugu proteins we identified contain a lysine-rich region immediately N-terminal to the HMG-box (Fig. 4), similar to the putative NLS sequence previously identified (Fig. 2B).

Fugu proteins share regions of homology with each other outside of the HMG-box domain. Table 2 shows a cross-comparison between family members. Interestingly, the proteins encoded by the genes on Scaffold-182 and Scaffold-2474 are 50% identical. There is also clear similarity between these pufferfish protein sequences and the TOX subfamily in regions outside of the putative NLS and HMG-box. Table 3 shows a comparison of the murine TOX subfamily members and the predicted homologues in Fugu in these regions. The predicted protein encoded by Scaffold-16 is approximately 60% identical to murine LOC244579, while the protein encoded by Scaffold-2474 shares 50% identity with murine LCP1. It should also be noted that the predicted Fugu protein sequences we have

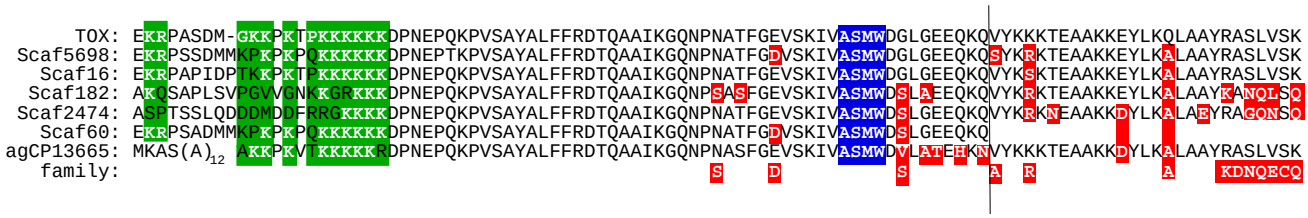

Figure 4

Identification of TOX subfamily predicted proteins in pufferfish and mosquito. Shown is an amino acid comparison of the TOX NLS and HMG-box regions with predicted proteins encoded in the *Fugu rubripes* or *Anopheles gambiae* genomes. *Fugu rubripes* predicted proteins are designated here simply by the scaffold (Scaf) location of the corresponding gene (see Methods). Amino acids that differ from the TOX HMG-box are highlighted in red, the putative NLS regions are highlighted in green, and the AXXW motif is highlighted in blue. The vertical line represents the position of the conserved exon boundary of the respective genes. Alternative amino acids found in other mammalian family members are also shown (family).

identified might not be full length. Thus, additional protein encoding exons may be present at these loci. Zebrafish also contain multiple putative TOX family members (data not shown).

Database searches of genomic sequences from *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (roundworm), and *Strongylocentrotus purpuratus* (sea urchin) did not reveal genes encoding the TOX HMG signature sequence (data not shown). However, analysis of the *Anopheles gambiae* (mosquito) genome revealed a gene (EAA06693) located on chromosome X that encodes a 109 amino acid protein (agCP13665) containing a TOX-like HMG-box motif (Fig. 4). This mosquito HMG-box is approximately 90% identical to that found in the mammalian TOX subfamily, despite the considerable evolutionary history separating the two. In addition, the HMG domain of agCP13665 is encoded by 2 exons, with an identical exon break as that found in mammalian TOX family members (Fig. 4). The highly conserved position of the intron in the TOX HMG domain coding sequence indicates that this is an ancient intron that was present before divergence of vertebrates.

The N-terminal domain of the mosquito protein consists almost entirely of a poly-alanine stretch followed by a lysine-rich sequence similar to the putative NLS of the TOX subfamily (Fig. 4) The C-terminal domain of agCP13665 is also short. Analysis of the genomic sequence surrounding this locus, however, revealed that the computer predicted sequence ends at a consensus splice donor and there is at least one possible additional exon approximately 15 kb downstream. This putative exon encodes a protein region with some similarity to KIAA0737 (data not shown). Thus, the agCP13665 protein may have a more extended C-terminal domain than has currently been identified. It seems likely that this mosquito protein represents an ancestral form of a TOX HMG-box protein that was subsequently duplicated and diversified during evolution of vertebrates.

Analysis of *Drosophila melanogaster* (fruitfly) sequences did not reveal a gene encoding the conserved TOX-type HMG-box. This is intriguing considering that both mosquito and fly species belong to the same taxonomic order (Diptera), and diverged approximately 250 million years ago, as compared to the 450 million years separating pufferfish and humans [45]. It has been noted that only half the genes in the Drosophila and Anopheles genomes can be interpreted as homologues [46]. We have identified a 250 amino acid protein (CG12104-PA) in the Drosophila database, which contains a partial HMG-box motif with weak similarity to the TOX family. The HMG-box of CG12104-PA is encoded by a single exon, unlike other TOX family members. However, the Drosophila genome is significantly smaller than that of the mosquito, at least in part due to loss of some introns [46]. Thus, it is possible that the CG12104-PA protein originated from an ancestral TOX family member, but has significantly diverged in Drosophila. In support of this possibility, CG12104-PA also contains a region in its C-terminus that shows some similarity to LCP1/ KIAA0737 (data not shown). Given that the mosquito protein may also have similarity in its C-terminal region to KIAA0737, it is possible that the latter human protein is most like the ancestral TOX family protein. Understanding the function of the TOX family of proteins should shed light on the evolutionary pressures that would maintain the TOX HMG-box essentially unchanged from mosquito to humans, but allow such divergence in Drosophila.

Table 2: Amino acid sequence comparison of Fugu TOX subfamily proteins outside the HMG-box and NLS regions.

	Scaf-16	Scaf-182	Scaf-2474	Scaf-5698
Scaf-16	100*	25	23	31
Scaf-182		100	50	0
Scaf-2474			100	0
Scaf-5698				100

* Percent identity of the two proteins as in Table 1. In all comparisons performed, Expect (E) values were 0.01 or lower.

Table 3: Amino acid sequence comparison of mouse and pufferfish TOX subfamily member proteins outside the HMG-box and NLS regions.

	Scaf-16	Scaf-182	Scaf-2474	Scaf-5698
TOX	41*	16	24	33
LCP1	33	40	48	0
LOC241768	19	13	0	40
LOC244579	61	21	16	31

* Percent identity of the two proteins as in Table 1. In all comparisons performed, Expect (E) values were 2e-07 or lower.

Conclusions

In this study, we introduce a subfamily of three additional mammalian proteins, which share a common HMG-box domain with TOX. We include a detailed analysis of the TOX subfamily members and show that the four family members are highly conserved in both mouse and human. Based on sequence alignment comparison with other previously characterized HMG-box proteins, we predict that the TOX HMG-box subfamily will likely fall into the sequence-independent category of HMG-box proteins. In addition, we identify TOX-like HMG-box proteins in mosquito and pufferfish. The identification of these invertebrate and early vertebrate sequences provides valuable insight into the evolution of the TOX subfamily. The documentation of this unique subfamily of HMG box proteins is a necessary initial step in evaluation of their biological function.

Methods

Database and sequence accession numbers

The sequences used in this study can be accessed through the NCBI website at http://www.ncbi.nih.gov, the *Fugu rubripes* website at http://genome.jgi-psf.org/fugu6/fugu6.home.html [47] and the FlyBase Consortium website at http://flybase.org/ [48]. The Fugu scaffold identities presented here refer to assembly v.3.0 of the Fugu genome. HMG-box domains are defined in the protein-domain Pfam database at http://Pfam.wustl.edu. *Yeast sequences*: NHP6A protein (accession NP_015377). *Mouse sequences*: Tox mRNA (NM_145711), TOX protein (NP_663757); LCP1 mRNA (AF228408), LCP1 predicted protein (AAK00713); LOC241768 mRNA (XM_141525.1), LOC244579 mRNA (XM_146430), SOX-4 protein (NP_033264), SOX-17 protein (NP_035571), LEF-1 protein NP_034833), HMGB1 protein (CAA56631), upstream binding factor 1 (UBF-1) protein (NP_035681). *Human sequences*: KIAA0808 mRNA (AB018351), KIAA0808 protein (BAA34528); KIAA0737 (C14ORF92) mRNA (NM_014828), KIAA0737 predicted protein (NP_055643); C20ORF100 mRNA (NM_032883), C20ORF100 predicted protein (NP_116272); TNRC9 mRNA (XM_049037), TNRC9 predicted protein (XP_049037), SEX determining region Y protein (SRY) (XP_010468). *Pufferfish sequences*: Fugu sequences have been designated by the scaffold location of the gene; Scaffold 2474 (coding sequence start at nucleotide position 6733, positive strand), Scaffold 182 (start 228504, negative strand), Scaffold 16 (start 42713, negative strand), Scaffold 5698 (start 5865, negative strand). Analysis of Fugu sequences was performed using tools available on the Fugu website (see above). *Mosquito sequence*: Locus EAA06693, predicted protein agCP13665 (accession EAA06693). *Fly sequences*: CG12104 gene (NP_647629), CG12104 predicted protein (CG12104-PA).

Note that the existing database sequence for LOC241768 encodes a predicted protein of 867 amino acids (accession XP_141525). Based on sequence comparison with human C20ORF100, analysis of exon boundaries, and EST evidence, we believe this sequence contains insertion of genomic sequence in error. In addition, we have removed 8 amino acids within the HMG-box motif from the database protein. This 8 amino acid insertion is not found in other TOX subfamily members or the human homologue C20ORF100, and can be explained by an error in the automated computer prediction of the end of the relevant exon. Our revised exon boundary maintains a consensus splice donor site. Therefore, a revised 520 amino acid protein based on coding sequence from 8 exons is described in this study (see Additional File 1: revised sequence for LOC241768). We also note that there is a difference in length between mouse C20ORF100 and it's human homologue. The mouse gene in the database includes a 3' exon encoding 60 amino acids. Based on a lack of EST evidence and RT-PCR (data not shown), this putative coding region may not be part of the expressed gene. This change is not included in the revised sequence (Additional File 1). LOC241768 was originally placed on murine chromosome 2 in a syntenic region with human C20ORF100. However, in Build 30 of the mouse genome assembly, LOC241768 has not been placed.

Note that the existing database sequence for LOC244579 encodes a predicted protein of 95 amino acids. Based on sequence comparison with human TNRC9, we believe that this represents an error generated by automated computational analysis. The correct sequence for LOC244579, which encodes a protein of 648 amino acids, can be located using gi:20888005. Database entries for these anomalous sequences are currently under review by NCBI (personal communication).

We also note the presence of a nucleotide sequence on chromosome 1 which is approximately 97% identical to LCP1 (LOC226876). However, this gene does not have an intron in the HMG-box coding region that is shared with other members of the TOX subfamily. In addition, expression of this gene is not supported by EST evidence. It is possible that this is a retroprocessed pseudogene similar to those seen within the GAPDH family [49].

Northern Analysis

A 340-bp fragment of an LCP1 cDNA was radiolabeled with [α^{32}P]-dCTP using a random primer labeling kit (Roche, Indianapolis, IN). This 3' LCP1 probe does not include the HMG-box encoding region and will not detect genes encoding other subfamily members. The probe was hybridized to a poly(A)$^+$ RNA Northern blot of mouse tissues (Origene Technologies, Rockville, MD) in ULTRAhyb hybridization buffer (Ambion, Austin, TX) overnight at 42°C and washed according to the manufacturer's instructions. Amounts of mRNA on this blot have been normalized previously with a β-actin probe (Origene Technologies). In addition, we have performed an independent normalization with a GAPDH probe. Blots were visualized using a Storm 860 imaging system (Molecular Dynamics, Sunnyvale, CA). Quantification of probe signal was preformed using NIH Image software. Results of hybridization of this same tissue blot with a *Tox* probe have been published [15].

Authors' contributions

EOF performed sequence comparison analysis, Northern Blot analysis, and drafted the manuscript. JK conceived of the study, performed sequence alignments, completed figures and edited the manuscript. Both authors read and approved the final manuscript.

Additional material

Additional File 1

Revised nucleic acid and predicted protein sequence for LOC241768
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-4-13-S1.pdf]

Acknowledgements

We are grateful to the other members of the laboratory (Parinaz Aliahmad, Olivia Goularte, Peggy Han and Jian-Ming Yang) for valuable discussion during the preparation of this manuscript and to Olivia Goularte for technical assistance. This work was supported by a grant from the National Institutes of Health (AI44110) to J.K. This is manuscript 15468-IMM from the Scripps Research Institute.

References

1. Goodwin GH and Johns EW **Isolation and characterisation of two calf-thymus chromatin non-histone proteins with high contents of acidic and basic amino acids** *Eur J Biochem* 1973, **40:**215-219
2. Bustin M **Regulation of DNA-dependent activities by the functional motifs of the high-mobility-group chromosomal proteins** *Mol Cell Biol* 1999, **19:**5237-5246
3. Thomas JO and Travers AA **HMG1 and 2, and related 'architectural' DNA-binding proteins** *Trends Biochem Sci* 2001, **26:**167-174
4. Bustin M **Revised nomenclature for high mobility group (HMG) chromosomal proteins** *Trends Biochem Sci* 2001, **26:**152-153
5. Baxevanis AD and Landsman D **The HMG-1 box protein family: classification and functional relationships** *Nucleic Acids Res* 1995, **23:**1604-1613
6. Aidinis V, Bonaldi T, Beltrame M, Santagata S, Bianchi ME and Spanopoulou E **The RAG1 homeodomain recruits HMG1 and HMG2 to facilitate recombination signal sequence binding and to enhance the intrinsic DNA-bending activity of RAG1-RAG2** *Mol Cell Biol* 1999, **19:**6532-6542
7. Allain FH, Yen YM, Masse JE, Schultze P, Dieckmann T, Johnson RC and Feigon J **Solution structure of the HMG protein NHP6A and its interaction with DNA reveals the structural determinants for non-sequence-specific binding** *EMBO J* 1999, **18:**2563-2579
8. Read CM, Cary PD, Crane-Robinson C, Driscoll PC and Norman DG **Solution structure of a DNA-binding domain from HMG1** *Nucleic Acids Res* 1993, **21:**3427-3436
9. Bewley CA, Gronenborn AM and Clore GM **Minor groove-binding architectural proteins: structure, function, and DNA recognition** *Annu Rev Biophys Biomol Struct* 1998, **27:**105-131
10. Soullier S, Jay P, Poulat F, Vanacker JM, Berta P and Laudet V **Diversification pattern of the HMG and SOX family members during evolution** *J Mol Evol* 1999, **48:**517-527
11. Pöhler JRG, Norman DG, Bramham J, Bianchi ME and Lilley DMJ **HMG box proteins bind to four-way DNA junctions in their open conformations** *EMBO J* 1998, **17:**817-826
12. Webb M and Thomas JO **Structure-specific binding of the two tandem HMG boxes of HMG1 to four-way junction DNA is mediated by the A domain** *J Mol Biol* 1999, **294:**373-387
13. Zlatanova J and van Holde K **Binding to four-way junction DNA: a common property of architectural proteins?** *FASEB J* 1998, **12:**421-431
14. Pil PM and Lippard SJ **Specific binding of chromosomal protein HMG1 to DNA damaged by the anticancer drug cisplatin** *Science* 1992, **256:**234-237
15. Wilkinson B, Chen JY, Han P, Rufner KM, Goularte OD and Kaye J **TOX: an HMG box protein implicated in the regulation of thymocyte selection** *Nat Immunol* 2002, **3:**272-280
16. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore

T, Hudson J., Jr., Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM and et al. **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling** *Nature* 2000, **403:**503-511
17. Kuo CT and Leiden JM **Transcriptional regulation of T lymphocyte development and function** *Annu Rev Immunol* 1999, **17:**149-187
18. Schilham MW and Clevers H **HMG box containing transcription factors in lymphocyte differentiation** *Semin Immunol* 1998, **10:**127-132
19. Staal FJ and Clevers H **Tcf/Lef transcription factors during T-cell development: unique and overlapping functions** *Hematol J* 2000, **1:**3-6
20. Schilham MW, Oosterwegel MA, Moerer P, Ya J, de Boer PA, van de Wetering M, Verbeek S, Lamers WH, Kruisbeek AM, Cumano A and Clevers H **Defects in cardiac outflow tract formation and pro-B-lymphocyte expansion in mice lacking Sox-4** *Nature* 1996, **380:**711-714
21. Masse JE, Wong B, Yen YM, Allain FH, Johnson RC and Feigon J **The S. cerevisiae architectural HMGB protein NHP6A complexed with DNA: DNA and protein conformational changes upon binding** *J Mol Biol* 2002, **323:**263-284
22. Learned RM, Learned TK, Haltiner MM and Tjian RT **Human rRNA transcription is modulated by the coordinate binding of two factors to an upstream control element** *Cell* 1986, **45:**847-857
23. Read CM, Cary PD, Preston NS, Lnenicek-Allen M and Crane-Robinson C **The DNA sequence specificity of HMG boxes lies in the minor wing of the structure** *EMBO J* 1994, **13:**5639-5646
24. Werner MH, Huth JR, Gronenborn AM and Clore GM **Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex** *Cell* 1995, **81:**705-714
25. Murphy F. V. th, Sweet RM and Churchill ME **The structure of a chromosomal high mobility group protein-DNA complex reveals sequence-neutral mechanisms important for non-sequence-specific DNA recognition** *EMBO J* 1999, **18:**6610-6618
26. Ohndorf UM, Rould MA, He Q, Pabo CO and Lippard SJ **Basis for recognition of cisplatin-modified DNA by high-mobility-group proteins** *Nature* 1999, **399:**708-712
27. Ner SS **HMGs everywhere** *Curr. Biol.* 1992, **2:**208-210
28. Love JJ, Li X, Case DA, Giese K, Grosschedl R and Wright PE **Structural basis for DNA bending by the architectural transcription factor LEF-1** *Nature* 1995, **376:**791-795
29. Nagase T, Ishikawa K, Suyama M, Kikuno R, Miyajima N, Tanaka A, Kotani H, Nomura N and Ohara O **Prediction of the coding sequences of unidentified human genes. XI. The complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro** *DNA Res* 1998, **5:**277-286
30. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ **Basic local alignment search tool** *J Mol Biol* 1990, **215:**403-410
31. Madden TL, Tatusov RL and Zhang J **Applications of network BLAST server** *Methods Enzymol* 1996, **266:**131-141
32. Margolis RL, Abraham MR, Gatchell SB, Li SH, Kidwai AS, Breschel TS, Stine OC, Callahan C, McInnis MG and Ross CA **cDNAs with long CAG trinucleotide repeats from human brain** *Hum Genet* 1997, **100:**114-122
33. Bowles J, Schepers G and Koopman P **Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators** *Dev Biol* 2000, **227:**239-255
34. Hosking BM, Wyeth JR, Pennisi DJ, Wang SC, Koopman P and Muscat GE **Cloning and functional analysis of the Sry-related HMG box gene, Sox18** *Gene* 2001, **262:**239-247
35. Gilbert W, de Souza SJ and Long M **Origin of genes** *Proc Natl Acad Sci U S A* 1997, **94:**7698-7703
36. Robbins J, Dilworth SM, Laskey RA and Dingwall C **Two interdependent basic domains in nucleoplasmin nuclear targeting sequence: identification of a class of bipartite nuclear targeting sequence** *Cell* 1991, **64:**615-623
37. Poulat F, Girard F, Chevron MP, Goze C, Rebillard X, Calas B, Lamb N and Berta P **Nuclear localization of the testis determining gene product SRY** *J Cell Biol* 1995, **128:**737-748
38. Yen YM, Wong B and Johnson RC **Determinants of DNA binding and bending by the Saccharomyces cerevisiae high mobility group protein NHP6A that are important for its biological activities. Role of the unique N terminus and putative intercalating methionine** *J Biol Chem* 1998, **273:**4424-4435
39. Gerber HP, Seipel K, Georgiev O, Hofferer M, Hug M, Rusconi S and Schaffner W **Transcriptional activation modulated by homopolymeric glutamine and proline stretches** *Science* 1994, **263:**808-811
40. Bowles J, Cooper L, Berkman J and Koopman P **Sry requires a CAG repeat domain for male sex determination in Mus musculus** *Nat Genet* 1999, **22:**405-408
41. Ross CA **Polyglutamine pathogenesis: emergence of unifying mechanisms for Huntington's disease and related disorders** *Neuron* 2002, **35:**819-822
42. Masino L and Pastore A **Glutamine repeats: structural hypotheses and neurodegeneration** *Biochem Soc Trans* 2002, **30:**548-551
43. Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B and Aparicio S **Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome** *Nature* 1993, **366:**265-268
44. Coutelle O, Nyakatura G, Taudien S, Elgar G, Brenner S, Platzer M, Drescher B, Jouet M, Kenwrick S and Rosenthal A **The neural cell adhesion molecule L1: genomic organisation and differential splicing is conserved between man and the pufferfish Fugu** *Gene* 1998, **208:**7-15
45. De Gregorio E and Lemaitre B **The mosquito genome: the post-genomic era opens** *Nature* 2002, **419:**496-497
46. Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM, Mueller HM, Dimopoulos G, Law JH, Wells MA, Birney E, Charlab R, Halpern AL, Kokoza E, Kraft CL, Lai Z, Lewis S, Louis C, Barillas-Mury C, Nusskern D, Rubin GM, Salzberg SL, Sutton GG, Topalis P, Wides R, Wincker P, Yandell M, Collins FH, Ribeiro J, Gelbart WM, Kafatos FC and Bork P **Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster** *Science* 2002, **298:**149-159
47. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MD, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJ, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D and Brenner S **Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes** *Science* 2002, **297:**1301-1310
48. The-FlyBase-Consortium **The FlyBase database of the Drosophila genome projects and community literature** *Nucleic Acids Res* 2002, **30:**106-108
49. Garcia-Meunier P, Etienne-Julan M, Fort P, Piechaczyk M and Bonhomme F **Concerted evolution in the GAPDH family of retrotransposed pseudogenes** *Mamm Genome* 1993, **4:**695-703