Phylogenetic Analysis of T-Box Genes Demonstrates the Importance of Amphioxus for Understanding Evolution of the Vertebrate Genome

Ilya Ruvinsky,[1] Lee M. Silver and Jeremy J. Gibson-Brown

Lewis Thomas Laboratory, Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544

Manuscript received September 20, 1999
Accepted for publication June 29, 2000

ABSTRACT

The duplication of preexisting genes has played a major role in evolution. To understand the evolution of genetic complexity it is important to reconstruct the phylogenetic history of the genome. A widely held view suggests that the vertebrate genome evolved via two successive rounds of whole-genome duplication. To test this model we have isolated seven new T-box genes from the primitive chordate amphioxus. We find that each amphioxus gene generally corresponds to two or three vertebrate counterparts. A phylogenetic analysis of these genes supports the idea that a single whole-genome duplication took place early in vertebrate evolution, but cannot exclude the possibility that a second duplication later took place. The origin of additional paralogs evident in this and other gene families could be the result of subsequent, smaller-scale chromosomal duplications. Our findings highlight the importance of amphioxus as a key organism for understanding evolution of the vertebrate genome.

COMPARISONS of the genomes of a wide variety of organisms have revealed that the evolution of genome complexity has not proceeded by nucleotide substitution alone, but rather has relied on extensive gene duplication (Haldane 1932; Ohno 1967; Nei 1969). These duplications could have involved individual genes or small chromosomal segments or encompassed the entire genome. In a classic treatise, Ohno (1970) proposed that the latter process may have been of particular importance during chordate evolution. By comparing the amount of DNA present in the nuclei of diverse animal species, he noted that there was an apparent stepwise increase in DNA content accompanying the morphological transitions from invertebrates, to primitive chordates, to vertebrates. He suggested that this phenomenon could be explained by two rounds of whole-genome duplication (tetraploidization). Recent estimates of gene numbers showing that modern vertebrates have on the order of 100,000 genes while their close invertebrate relatives possess around 15–20,000 genes (Simmen *et al.* 1998) appear to support this notion, although a wide range of estimates continues to be proposed (Ewing and Green 2000; Liang *et al.* 2000; Roest Crollius *et al.* 2000).

Comparisons based on gene numbers are a better test of genome complexity than those based on DNA content, because the amount of noncoding sequence varies dramatically both within and between taxa, a phenomenon known as the "C-value paradox" (Li 1997). However, an even more precise approach is to compare the number of genes within different gene families present in both vertebrate and invertebrate genomes. In this type of study it is important to sample comprehensively within a family because incomplete data sets will lead to the reconstruction of incomplete phylogenies, making it impossible to calculate the correct number of gene duplication events that have occurred. Even if all the genes within a family are obtained, incorrect inferences regarding the number of duplications can still be made unless correct phylogenetic relationships have been established. For example, a single invertebrate gene may be either closely related to a subset of its vertebrate homologs or equally related to all of them. Clearly these different relationships imply different historic patterns of gene duplication.

Recently, the observation that a single invertebrate locus corresponds to several (sometimes three or four) vertebrate counterparts in a number of gene families served to revive the idea that vertebrate genomes evolved via two rounds of tetraploidization (Holland *et al.* 1994; Sidow 1996). However, with the sole exception of the *Hox* gene clusters, there is no gene family in which all the genes have been isolated from the genomes of both a vertebrate and a basal chordate. This paucity of data has seriously impeded accurate reconstruction of the sequence of gene duplication events that have occurred in the course of vertebrate genome evolution. Consequently, diverse models have been proposed, ranging from several rounds of tetraploidization followed by extensive gene loss, to multiple subchromosomal duplications (reviewed by Skrabanek and Wolfe 1998; Smith *et al.* 1999).

The cephalochordate amphioxus is the closest living

Corresponding author: Jeremy J. Gibson-Brown, Department of Biology, Washington University, 1 Brookings Dr., St. Louis, MO 63130. E-mail: gibbro@biology.wustl.edu

[1] *Present address:* Department of Molecular Biology, Massachusetts General Hospital and Department of Genetics, Harvard Medical School, Boston, MA 02114.

invertebrate relative of the vertebrates (Wada and Satoh 1994) and therefore the best model organism for understanding the composition of the ancestral chordate genome. In an ongoing study of the function of T-box genes during vertebrate embryogenesis and evolution (Gibson-Brown *et al.* 1996, 1998a,b; Ruvinsky *et al.* 1998, 2000), we decided to isolate the amphioxus members of this gene family to investigate their roles during evolution of the vertebrate body plan. T-box genes encode a family of sequence-specific DNA-binding proteins that are known to act as transcription factors during embryogenesis of diverse metazoans ranging from hydra to humans (Papaioannou and Silver 1998; Papaioannou 2000). Since we have found that several T-box gene duplications occurred around the divergence of the vertebrate and invertebrate lineages (Agulnik *et al.* 1996; Ruvinsky and Silver 1997; Ruvinsky *et al.* 2000), we predicted that characterization of the entire gene family in an invertebrate chordate would provide an insight into vertebrate genome evolution. We have therefore undertaken an extensive series of systematic screens for amphioxus T-box genes.

MATERIALS AND METHODS

Tissue samples and cDNA libraries: Adult amphioxus (*Branchiostoma floridae*) were collected off the south shore of Courtney Campbell Causeway in Old Tampa Bay (Tampa, FL) during the spawning season of 1998. Animals were frozen upon collection. Genomic DNA was extracted from a single adult male using a standard phenol-chloroform purification method. Two λZapII amphioxus cDNA libraries were screened for T-box genes. One was constructed from 5- to 24-hr embryos (provided by Jim Langeland of Kalamazoo College, Kalamazoo, MI), the other, from 2- to 4-day larvae (provided by Linda Holland of the Scripps Institution of Oceanography, San Diego, CA).

PCR on genomic DNA: A set of degenerate primers was designed against the following oligopeptide sequences: NSMHKYQ (forward) and VTSYQNHK (reverse). This primer pair amplifies an ~150-nucleotide fragment completely contained within one of the exons of the T-box (Figure 1). A high level of sequence variation within this region allows the unambiguous assignment of a gene to a specific T-box gene subfamily. PCR amplification on genomic DNA was carried out (35 cycles: 95° for 1 min, 50° for 1 min, 72° for 1.5 min) and the products were cloned into the pCR2.1 vector (Invitrogen, San Diego). Thirty-six independent clones were sequenced using an ABI sequencer.

Library screens: Initially, a mixed embryonic stage library was screened at high stringency (hybridized in Church buffer at 65°, washed twice at 65° in 0.1× SSC, 0.1% SDS) with a cocktail of cloned PCR fragments derived from five different amphioxus T-box genes. Positive clones were plaque-purified and excised *in vivo*. Replicate dot-blots were probed with the same five PCR fragments used for screening and led to the discovery of three different genes. Since two anticipated genes were not obtained from this screen, a later-stage larval library was screened under the same conditions with a cocktail of the remaining two PCR probes yielding a single new gene. Finally, the embryonic library was rescreened at moderate stringency (hybridized at 57°, washed twice at 60° in 0.5× SSC, 0.1% SDS) with a probe derived from the zebrafish *tbx16* gene (Ruvinsky *et al.* 1998). Clones corresponding to two more genes were identified. One or more of the longest clones of each gene were sequenced.

Phylogenetic analysis: Amino acid sequences of T-domains from the newly characterized genes were manually aligned with those of other family members using the Wisconsin GCG package (Genetics Computer Group 1996). Unalignable regions were excluded from analysis. A neighbor-joining tree was constructed, and the reliability of its topology was statistically tested, using the METREE program (Rzhetsky and Nei 1994). Appropriate Drosophila and *Caenorhabditis elegans* sequences were included to provide a timescale reference and serve as outgroups.

RESULTS

Isolation of seven new amphioxus T-box genes: Amplification by PCR from genomic DNA yielded fragments of five distinct amphioxus T-box genes. High stringency screening of two cDNA libraries with these fragments resulted in the isolation of clones corresponding to four different genes. Two additional genes were isolated in a subsequent low stringency screen. No clones corresponding to one of the five PCR fragments were recovered in any of the library screens. Thus we have recovered cDNA clones of six previously uncharacterized genes and a PCR product derived from a seventh gene. Including the two previously reported genes, *AmphiBra1* and *AmphiBra2* (Holland *et al.* 1995; Terazawa and Satoh 1995), this brings the total complement of T-box genes in the amphioxus genome to a minimum of nine genes. We have aligned the newly obtained amphioxus sequences to those of genes from all previously described T-box subfamilies (Figure 1).

In addition, we have identified and included three new human T-box genes based on sequences available in GenBank. The first, *TBX20* (AJ237589; Meins *et al.* 2000), is orthologous to zebrafish *tbx20* (Ahn *et al.* 2000), also known as *hrT* (Griffin *et al.* 2000), Drosophila *H15* (X98766; Brook and Cohen 1996), and *C. elegans tbx-12* (Agulnik *et al.* 1997). The second is *TBX21*, formerly known as *TBLYM* (AF093098; S. Yang, unpublished results) and *T-bet* (AF241243; Szabo *et al.* 2000). The third gene, which we have designated *TBX22* (AL031000) consistent with our previous practice and with the approval of the Human Gene Nomenclature Committee, has been identified through the genome sequencing efforts of the Sanger Centre Chromosome X Mapping Group.

Phylogenetic positions of amphioxus T-box genes: For meaningful comparisons to be made between genes in different species it is essential to distinguish genes that are orthologous (separated due to speciation events) from those that are paralogous (separated due to gene duplication events). To determine orthology/paralogy relationships between the amphioxus and vertebrate genes we conducted a phylogenetic analysis of the entire gene family. In the analysis we included two orthologs

of each known vertebrate T-box gene whenever possible. When selecting which vertebrate species to include, we consistently chose the two most distantly related organisms for which the longest sequences were available. For example, a human/zebrafish gene pair was preferred over a human/chicken gene pair. Because the mouse and human orthologs are nearly identical they can be considered interchangeable.

The sequence of the PCR fragment for which no cDNA clones were obtained was too short to be included in the phylogenetic analysis. However, since this sequence spans the most variable region within the T-box (Figure 1), visual inspection allowed its provisional assignment as an amphioxus ortholog of the vertebrate *Tbx20* gene (within the 34 amino acids compared there were only 7 amino acid replacements, of which 3 are conservative).

The phylogenetic relationships of the rest of the newly obtained amphioxus T-box sequences were determined by a neighbor-joining analysis (Figure 2). Examination of the tree reveals that in no case do we find a 1:4 correspondence between the number of amphioxus and vertebrate genes as predicted by the "two whole-genome duplication" model. Instead, we typically observe a 1:2 or 1:3 correspondence. We consider each subfamily individually below.

Tbx1/10: A single amphioxus gene corresponds to two vertebrate genes, a result consistent with a single genome duplication.

Tbx15/18/22: A single amphioxus gene corresponds to three vertebrate genes. It should be noted that whereas *Tbx15* and *Tbx18* comprise a pair of most closely related paralogs, the branching order of *Tbx22* and *AmphiTbx15/18/22* is only weakly supported and should therefore be considered unresolved. This result is consistent with at least two possible scenarios: two genome duplications followed by a single gene loss, or a single tetraploidization followed by a local gene duplication.

Tbx20: A single amphioxus gene corresponds to a single vertebrate gene. If one genome duplication had occurred after separation of the cephalochordate and vertebrate lineages, only a single gene loss would have to be invoked. More gene losses would have to be postulated if additional genome duplications had occurred. If no genome duplications have occurred, no gene losses would have to be invoked.

Tbx2/3 and Tbx4/5: Genes within these two subfamilies are present in the genome as two cognate, linked pairs (Agulnik *et al.* 1996; Ruvinsky and Silver 1997). Because of their close linkage, *Tbx2* and *Tbx4* should be considered as sampling a single locus, as should *Tbx3* and *Tbx5*. The topology within the *Tbx2/3* subfamily is inconsistent with the well-established phylogenetic relationships of the species: amphioxus is more closely related to vertebrates than is Drosophila. However, the internal branch separating (d-*omb* (*Tbx2*,*Tbx3*)) from *AmphiTbx2/3* receives little statistical support and should thus be considered artifactual. In both of these subfamilies a single amphioxus gene corresponds to two vertebrate genes, consistent with a single genome duplication.

Eomes/Tbr1/Tbx21: Due to the lack of statistical support, the divergence patterns of the basal branches within this subfamily should be considered unresolved. There is therefore an apparent correspondence between a single amphioxus gene and three vertebrate genes. Thus the two possible scenarios outlined above for the *Tbx15/18/22* subfamily apply in this case as well.

Brachyury/Tbx19: The phylogenetic relationships within this subfamily are complicated. The two amphioxus *Brachyury* genes are derivatives of a relatively recent lineage-specific duplication (Holland *et al.* 1995; Figure 2), implying that the ancestral cephalochordate genome contained a single locus. It is possible that, as in the case of the *Tbx20* subfamily, this single ancestral locus corresponds to a single vertebrate gene, implying that an amphioxus counterpart to *Tbx19* either could have been lost or is waiting to be discovered. It is also possible that, despite a high confidence probability value, the nesting of the amphioxus genes with vertebrate *Brachyury* is artifactual. This interpretation would imply that a single ancestral locus gave rise to both the vertebrate *Brachyury* and *Tbx19* genes, subsequent to the divergence of the cephalochordates. Finally, it should be noted that the topology of this subfamily is similar to that of the *Eomes/Tbr1/Tbx21* subfamily. If only a single gene loss had occurred in the latter (*e.g.*, *Tbr1*), the two topologies would become identical. It is formally possible that a recently described *Brachyury*-like gene in Xenopus (*Xbra3*, Hayata *et al.* 1999) represents this "lost" gene. However, it is more closely related to the other Xenopus *Brachyury* gene (*Xbra*) than it is to either *Tbx19* or the *Brachyury* genes from other tetrapods (analyses not shown). Since Xenopus is known to be a tetraploid species (Skrabanek and Wolfe 1998), *Xbra3* is most likely a pseudoallele of *Brachyury*. If an ortholog of this gene were to be found in nontetraploid species such as humans and mice, this interpretation would have to be rejected. The above arguments suggest that any of the scenarios encountered so far (1:1, 1:2, or 1:3) are possible in the case of this subfamily.

Tbx6/Tbx16: Previous analyses have demonstrated that orthology assignments within the vertebrate *Tbx6/Tbx16* subfamily are complicated. For example, despite almost identical expression patterns (Chapman *et al.* 1996; Hug *et al.* 1997), the mouse and zebrafish *Tbx6* genes are apparently not orthologous (Ruvinsky *et al.* 1998). Furthermore, orthologs of *Tbx16* have been described in zebrafish (*tbx16*), Xenopus (variously named *Antipodean*, *Brat*, *VegT*, and *Xombi*), and chicken (*Tbx6L*), but not in mouse or human, the two species from which the largest number of T-box genes are known and in which the most intensive screens for new genes have been undertaken. Several possible explanations can ac-

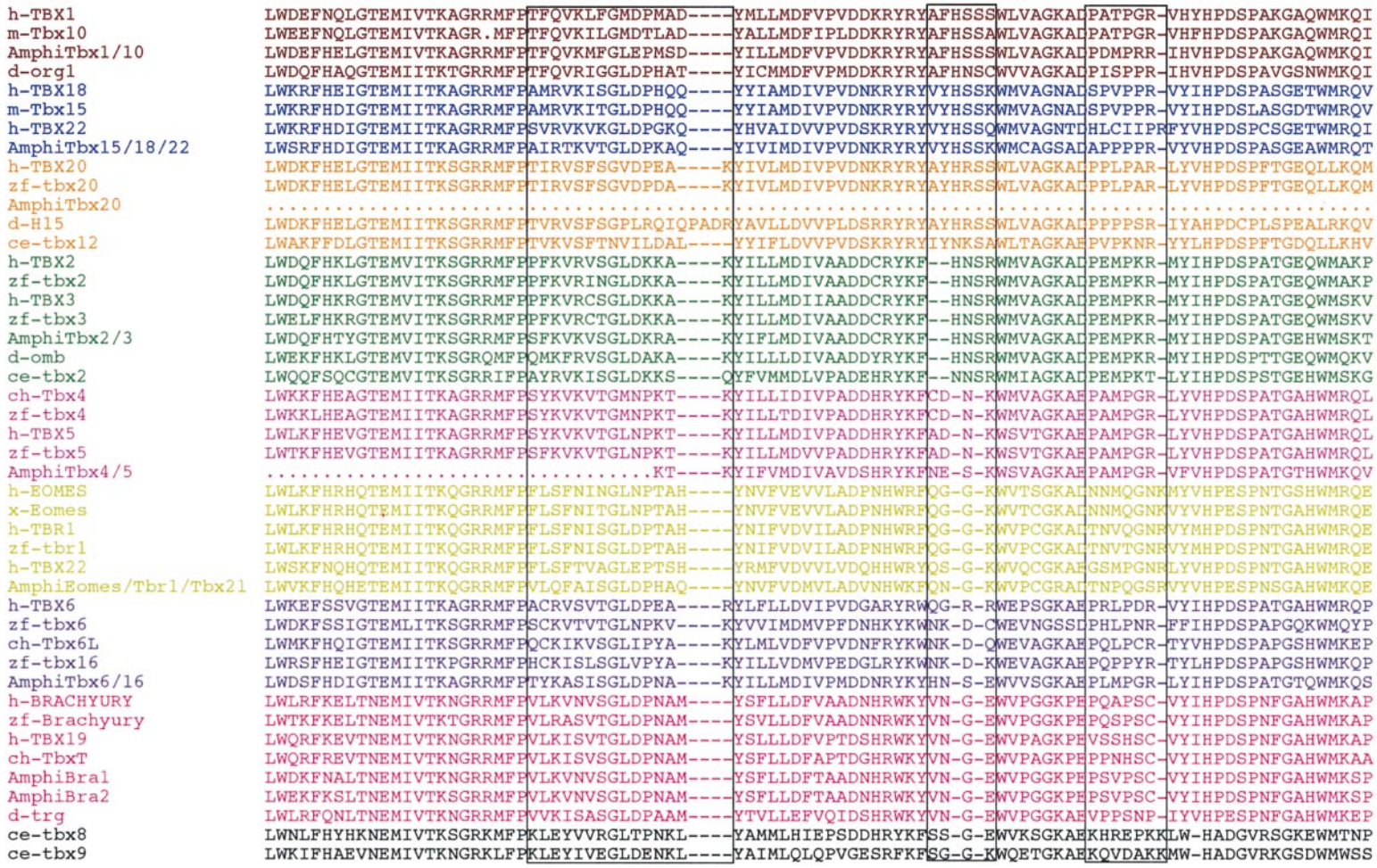

FIGURE 1.—Alignment of the amino acid sequences of amphioxus T-domains to those of other species. Members of the same subfamily are shown in the same color. Boxed regions were unalignable and were excluded from the phylogenetic analysis. Arrows indicate positions of the PCR primers. Abbreviations: h, human; m, mouse; ch, chicken; x, Xenopus; zf, zebrafish; d, Drosophila; ce, *C. elegans*. All sequences reported here have been deposited in GenBank under the following accession numbers: *AmphiTbx1/10* (AF262562), *AmphiTbx2/3* (AF262563), *AmphiTbx4/5* (AF262564), *AmphiTbx6/16* (AF262565), *AmphiTbx15/18/22* (AF262566), *AmphiTbx20* (AF262567), and *AmphiEomes/Tbr1/Tbx21* (AF262568).

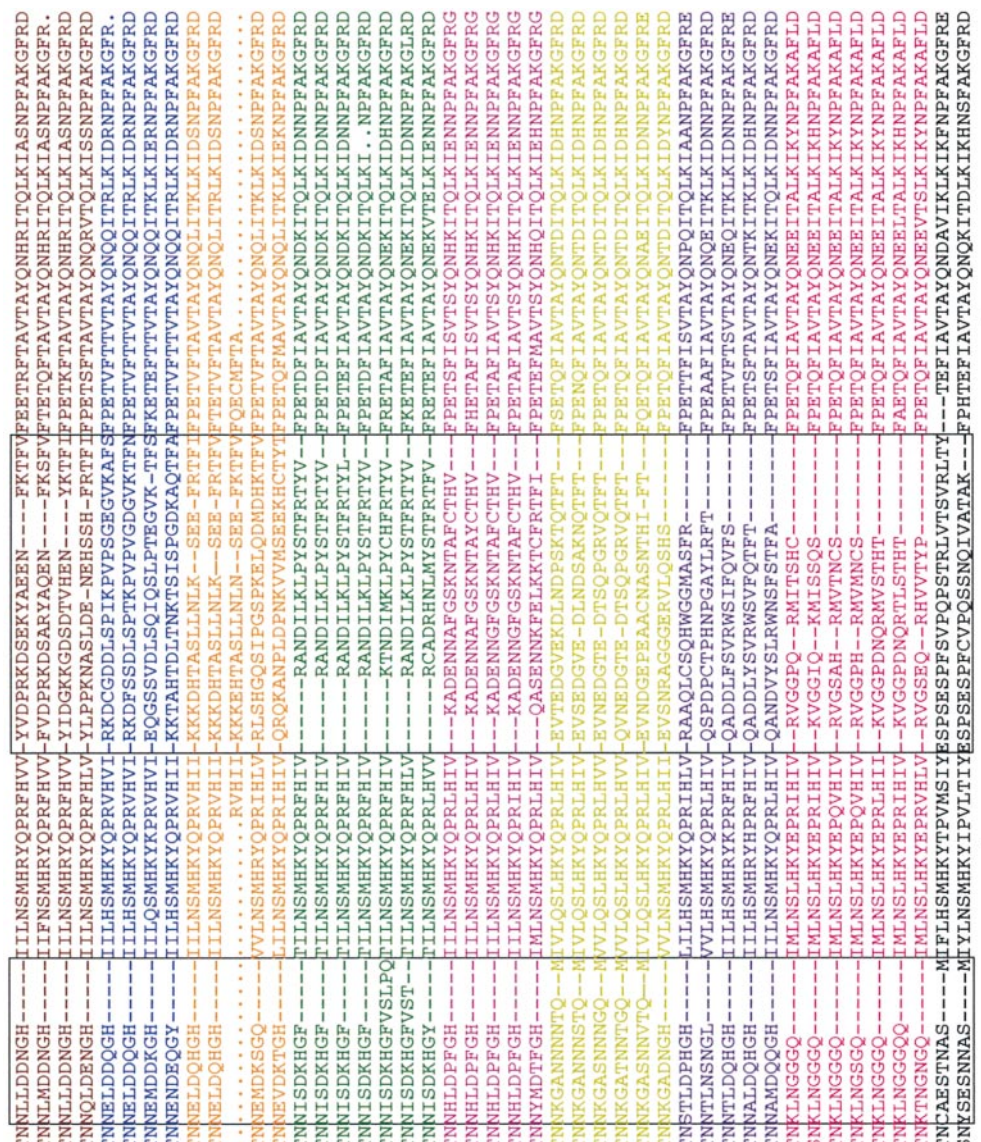

FIGURE 1.—*Continued.*

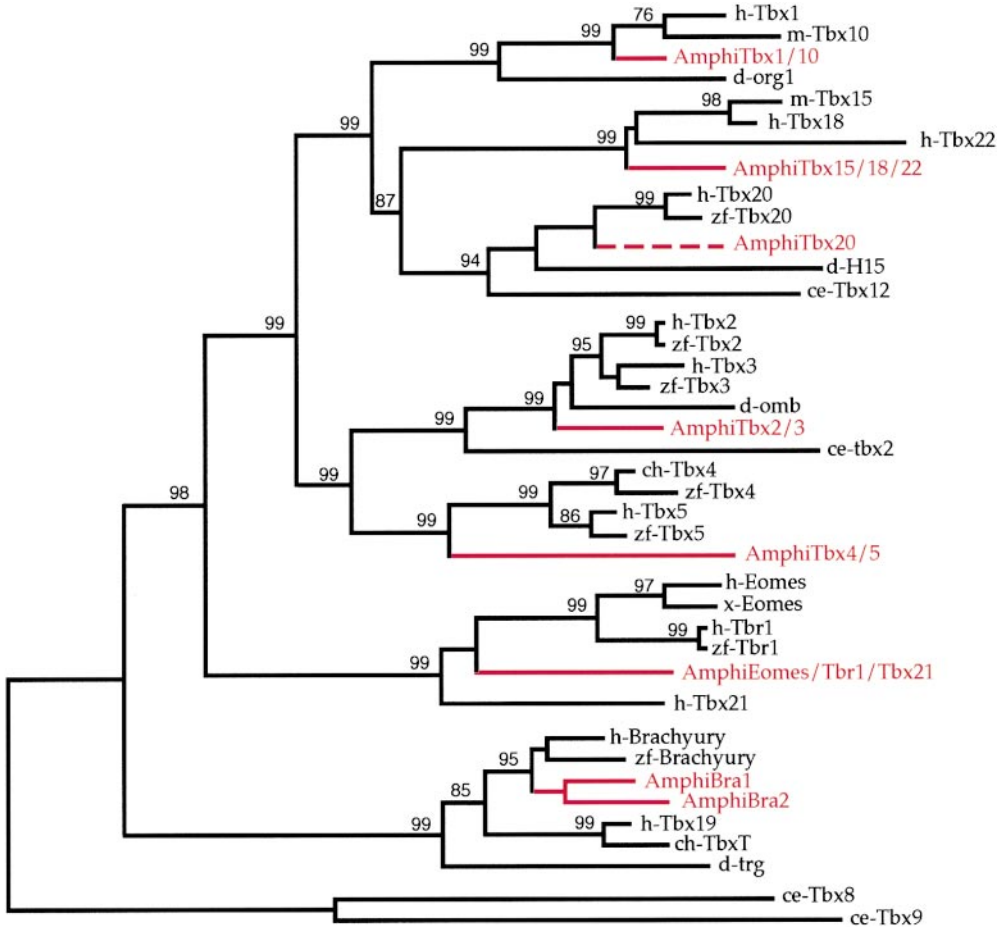

FIGURE 2.—Evolution of the T-box gene family. Phylogenetic positions of amphioxus genes (red) as revealed by a neighbor-joining algorithm. Dashed line indicates provisional placement of *AmphiTbx20* based on visual comparison of the sequence of its short PCR fragment to other family members. Confidence probability values >75% (shown) indicate reliable nodes; others should be deemed unreliable. *ce-tbx8* and *ce-tbx9* were used as an outgroup. Abbreviations are as in Figure 1.

count for this phenomenon. First, genes of this subfamily appear to be evolving at a faster rate than those of other subfamilies, thus complicating the phylogenetic analysis (LI 1997). Second, there may have been one or more instances of gene evolution by a birth-and-death mechanism, whereby different paralogs are eliminated in different lineages (NEI *et al.* 1997). Third, a relatively recent gene conversion event between paralogous T-box genes could have been responsible for the origin of substantial sequence differences between genuine orthologs (LI 1997). For these reasons, and since inclusion of genes of the *Tbx6/Tbx16* subfamily disrupts the overall topology of the T-box family tree (analysis not shown), we excluded them from the phylogenetic analysis. However, as with *AmphiTbx20*, we were able to assign one of the amphioxus cDNA clones to this putative subfamily on the basis of visual comparison of its sequence to those of other T-box genes within the highly variant region of the T-domain (Figure 1, region between the PCR primers). As in the case of the *Brachyury/Tbx19* subfamily, it is not possible at present to determine the true ratio between the vertebrate and amphioxus paralogs. Additional work will be required to resolve the enigmatic phylogenetic relationships within this putative subfamily.

DISCUSSION

A tentative interpretation of the relationships between the amphioxus and vertebrate T-box genes, based on the phylogenetic tree and the above arguments, is represented schematically in Figure 3. Examination of this diagram reveals three clear cases of a 1:2 correspondence between the number of cephalochordate and vertebrate genes (*Tbx1/10*, *Tbx2/3*, and *Tbx4/5*). Since *Tbx2* and *Tbx4*, as well as *Tbx3* and *Tbx5*, are organized in two tightly linked clusters (AGULNIK *et al.* 1996; RUVINSKY and SILVER 1997; WATTLER *et al.* 1998), *Tbx2/3* and *Tbx4/5* were linked in the preduplication condition. The amphioxus genes should therefore be considered as representing a single locus. There are two cases of an apparent 1:3 correspondence (*Tbx15/18/22* and *Eomes/Tbr1/Tbx21*) and one instance of a 1:1 correspondence (*Tbx20*). Finally, in the case of the last two subfamilies (*Tbx6/16* and *Brachyury/Tbx19*), where the relation-

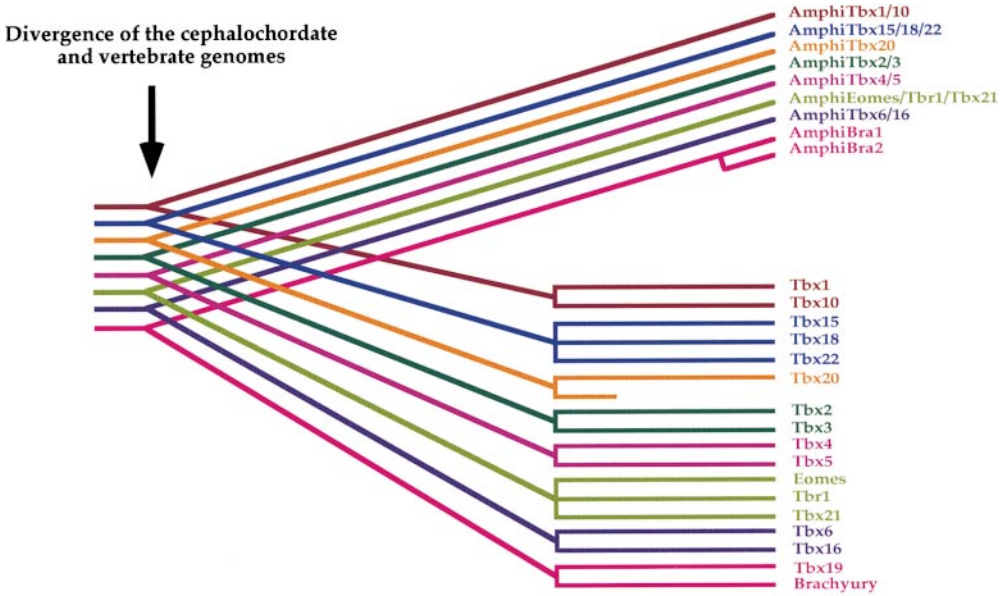

FIGURE 3.—A tentative interpretation of the relationships between the amphioxus and vertebrate T-box genes. Truncated terminal branch indicates inferred gene loss. Branch bifurcations within the vertebrate lineage should not be interpreted as necessarily representing simultaneous events. Unresolved trichotomies do not imply simultaneous gene birth.

ships are far from clear, two vertebrate genes appear to correspond to a single cephalochordate gene.

Comprehensive sampling of a gene family is essential for determining correct orthology/paralogy relationships. Incomplete data sets are bound to give incorrect estimates of the number and pattern of gene duplication events during evolution of the family, undermining their utility for the understanding of genome evolution.

Our data represent the most extensive sampling of an amphioxus gene family to date: nine loci were analyzed, of which seven can be considered independent data points for the analysis of genome evolution, as they are dispersed throughout the genome (BOLLAG *et al.* 1994; AGULNIK *et al.* 1996, 1998; HANCOCK *et al.* 1999; YI *et al.* 1999). To assess the completeness of our vertebrate T-box gene data set, we searched GenBank to see how many of the known T-box genes have been identified through the "random" sequencing efforts of the Human Genome Project. Because, in the three-fourths of the human genome sequenced to date (press release dated 04/15/2000; http://www.ncbi.nlm.nih.gov/genome/seq/), 13 of the 17 known human T-box genes have been found, it is unlikely that many, if any, more genes remain to be discovered. Because we were able to isolate amphioxus cognates of all known vertebrate T-box genes and because no amphioxus genes without a vertebrate counterpart were recovered, we can be confident that we have obtained a comprehensive data set.

The overall topology of the phylogenetic tree presented in Figure 2 immediately suggests a framework for a revised, rational nomenclature of the T-box gene family. In particular we note that, in accordance with the earlier proposals of AGULNIK *et al.* (1996) and PAPAIOANNOU and SILVER (1998), the family can be subdivided into a number of subfamilies. Once the complete sequence of the human genome is available, it would be an opportune time to rationalize the nomenclature taking into consideration the phylogenetic relationships within the entire family. The purpose of such a scheme would be to allow the unambiguous placement and appropriate naming, of any newly discovered gene, from any metazoan, within a preestablished framework. This would prevent the unfortunate practice of the inconsistent naming of new genes, benefiting the community as a whole and especially those engaged in comparative studies of T-box genes in different species.

The widely accepted notion that there have been two rounds of whole-genome duplication at the base of the vertebrate lineage derives, in large part, from the fact that amphioxus possesses a single *Hox* cluster, whereas the inferred ancestral condition for jawed vertebrates is four *Hox* clusters (GARCIA-FERNANDEZ and HOLLAND 1994). The recent discovery of at least seven *Hox* clusters in zebrafish (AMORES *et al.* 1998; PRINCE *et al.* 1998) and medaka (NARUSE *et al.* 2000) represents a derived condition within the teleost fish lineage and does not alter this interpretation. There are two distinct problems in inferring the pattern of evolution of the entire genome from the *Hox* data set. First, despite the fact that there are as many as 13 genes in each cluster, since they are tightly linked, each cluster can only be considered as sampling a single locus. Thus a phylogenetic analysis based on *Hox* clusters can reveal the evolutionary history of only a very small portion of the genome. Confident reconstructions of genome history should be based on the examination of a large number of independent, unlinked loci. Thus our data set of seven independent loci provides a much more extensive coverage of the genome. Second, if four genes (1, 2, 3, and 4) are the products of two successive rounds of whole-genome duplication, their phylogenetic relationship must be

((1,2)(3,4)), yet the topology reconstructed for the *Hox* clusters (Zhang and Nei 1996; Bailey *et al.* 1997) actually appears to be (1(2(3,4))). This can be interpreted as evidence for a three-step sequential origin of four *Hox* clusters, contradicting the two whole-genome duplication model (Bailey *et al.* 1997). Other studies (Skrabanek and Wolfe 1998; Hughes 1999; Martin 1999) also demonstrate that, despite perceptions to the contrary, existing data do not currently support the view that vertebrate genome evolution has proceeded via two rounds of tetraploidization.

What can be concluded about the evolution of the vertebrate genome on the basis of our data? When drawing inferences about the distant evolutionary past of complex genetic systems, as in other areas of science, one can never prove a conjecture, but can merely gather the evidence required to reject a specific hypothesis. Additional complications arise in this case because there is no single history of "the vertebrate genome," since different gene families have evolved along different routes in different lineages. This is not to say that no progress can be made.

Clearly, there has been a dramatic increase in the number of genes within the vertebrate lineage following its separation from the cephalochordates, rejecting the concept of a "static genome." This increase in gene number could have been due to either numerous small-scale duplications or a few genome-wide duplications, or perhaps a combination of the two.

If the vertebrate genome was assembled in a piecemeal manner, this would imply two distinct phases in the rate of genome evolution. In the early phase, between the divergence of cephalochordates and the origin of jawed vertebrates, a high rate of local gene duplications would have to be postulated. Subsequently, the rate of duplications must have slowed considerably, or almost stopped, because all jawed vertebrates have a very similar gene complement (teleost- and Xenopus-specific tetraploidizations notwithstanding). Both molecular and paleontological data indicate that the first phase was considerably shorter than the second (Kumar and Hedges 1998; Conway Morris 2000). Moreover, it is known that tetraploidizations do occur and produce viable organisms. Thus it seems more plausible to suggest that at least one whole-genome duplication was involved in the elaboration of vertebrate gene families. The identification in vertebrate genomes of large paralogous chromosomal regions (*e.g.*, Lundin 1993; Bailey *et al.* 1997; Ruvinsky and Silver 1997), in which genes appear to have duplicated at the same time, further supports the whole-genome duplication hypothesis.

Conventionally, considerations of parsimony require that, unless compelling evidence is presented to the contrary, the interpretation requiring the minimum number of events is accepted as the most likely explanation. It is formally possible that the vertebrate genome has undergone many rounds of tetraploidization followed by extensive gene loss. Indeed, gene loss is known to be extensive in some lineages and can be responsible for determining the size of the genome (Petrov *et al.* 1996, 2000). Despite this, our data provide no evidence to suggest that there have been more than two whole-genome duplications.

We conclude that at least one but no more than two whole-genome duplications occurred in the vertebrate lineage, after divergence of the cephalochordates, but before the radiation of extant jawed vertebrates. The origin of additional paralogs evident in this and other gene families could be the result of subsequent, smaller-scale chromosomal duplications.

To infer the steps through which the vertebrate genome has evolved it is ultimately desirable to compare the full complement of genes from the genomes of a basal chordate and a crown-group vertebrate. Completion of the Human Genome Project in the near future will provide a complete data set for the latter. Currently, the fully sequenced genomes of Drosophila and *C. elegans* provide the only source of information for comparative genome analyses in metazoans. The present study highlights the utility of amphioxus as a more appropriate organism for understanding the ancestral composition of the chordate genome. If complete data sets for a large number of amphioxus gene families were to become available, they could be subjected to the type of phylogenetic analysis presented here. This large number of independent data sets would provide an invaluable resource for the understanding of vertebrate genome evolution.

We thank Nick and Linda Holland for instruction in the collection of amphioxus, Jim Langeland and Linda Holland for their generous gifts of the cDNA libraries used in these studies, John Lawrence for kindly providing laboratory space at the University of South Florida, Valery Kanevsky for statistical advice, and Ginny Papaioannou and Maurice Eash for critical reading of the manuscript. This work was supported by National Institutes of Health grant HD-20275 (L.M.S.), National Science Foundation grant DEB-9901943 (I.R. and L.M.S.), and a *Development* Travelling Fellowship from The Company of Biologists (J.J.G.-B.).

Note added in proof: Since acceptance of the manuscript, the draft sequence of the human genome has been released. By searching GenBank we have found one additional human T-box gene, which we have designated *TBX23* with the approval of the Human Gene Nomenclature Committee (accession no. AL157899), that was not included in our original analysis. *TBX23* is closely related to the human *T* and *TBX19* genes, but only distantly related to the genes from other subfamilies. This increases to three the number of T-box subfamilies in which there is an apparent 1:3 correspondence between the number of cephalochordate and vertebrate genes.

LITERATURE CITED

Agulnik, S. I., N. Garvey, S. Hancock, I. Ruvinsky, D. L. Chapman *et al.*, 1996 Evolution of mouse T-box genes by tandem duplication and cluster dispersion. Genetics **144:** 249–254.

Agulnik, S. I., I. Ruvinsky and L. M. Silver, 1997 Three novel T-box genes in *Caenorhabditis elegans*. Genome **40:** 458–464.

Agulnik, S. I., V. E. Papaioannou and L. M. Silver, 1998 Cloning, mapping, and expression analysis of *TBX15*, a new member of the T-box gene family. Genomics **51:** 68–75.

Ahn, D., I. Ruvinsky, A. C. Oates, L. M. Silver and R. K. Ho, 2000 *tbx20*, a new vertebrate T-box gene expressed in the cranial motor neurons and developing cardiovascular structures in zebrafish. Mech. Dev. **95:** 253–258.

Amores, A., A. Force, Y.-L. Yan, L. Joly, C. Amemiya *et al.*, 1998 Zebrafish *hox* clusters and vertebrate genome evolution. Science **282:** 1711–1714.

Bailey, W. J., J. Kim, G. P. Wagner and F. H. Ruddle, 1997 Phylogenetic reconstruction of vertebrate *Hox* cluster duplications. Mol. Biol. Evol. **14:** 843–853.

Bollag, R. J., Z. Siegfried, J. A. Cebra-Thomas, N. Garvey, E. M. Davidson *et al.*, 1994 An ancient family of embryonically expressed mouse genes sharing a conserved protein motif with the *T* locus. Nat. Genet. **7:** 383–389.

Brook, W. J., and S. M. Cohen, 1996 Antagonistic interactions between *wingless* and *decapentaplegic* responsible for dorsal-ventral pattern in the *Drosophila* leg. Science **273:** 1373–1377.

Chapman, D. L., I. Agulnik, S. Hancock, L. M. Silver and V. E. Papaioannou, 1996 *Tbx6*, a mouse T-box gene implicated in paraxial mesoderm formation at gastrulation. Dev. Biol. **180:** 534–542.

Conway Morris, S., 2000 The Cambrian "explosion": slow-fuse or megatonnage? Proc. Natl. Acad. Sci. USA **97:** 4426–4429.

Ewing, B., and P. Green, 2000 Analysis of expressed sequence tags indicates 35,000 human genes. Nat. Genet. **25:** 232–234.

Garcia-Fernandez, J., and P. W. H. Holland, 1994 Archetypal organization of the amphioxus *Hox* gene cluster. Nature **370:** 563–566.

Genetics Computer Group, 1996 *The Wisconsin GCG Package: Version 9*. G.C.G., Inc., Madison, WI.

Gibson-Brown, J. J., S. I. Agulnik, D. L. Chapman, M. Alexiou, N. Garvey *et al.*, 1996 Evidence of a role for T-box genes in the evolution of limb morphogenesis and the specification of forelimb/hindlimb identity. Mech. Dev. **56**: 93–101.

Gibson-Brown, J. J., S. I. Agulnik, L. M. Silver, L. Niswander and V. E. Papaioannou, 1998a Involvement of T-box genes *Tbx2-Tbx5* in vertebrate limb specification and development. Development **125:** 2499–2509.

Gibson-Brown, J. J., S. I. Agulnik, L. M. Silver and V. E. Papaioannou, 1998b Expression of T-box genes *Tbx2-Tbx5* during chick organogenesis. Mech. Dev. **74:** 165–169.

Griffin, K. J., J. Stoller, M. Gibson, S. Chen, D. Yelon *et al.*, 2000 A conserved role for *H15*-related T-box transcription factors in zebrafish and *Drosophila* heart formation. Dev. Biol. **218:** 235–247.

Haldane, J. B. S., 1932 *The Causes of Evolution*. Harper Bros., London.

Hancock, S. N., S. I. Agulnik, L. M. Silver and V. E. Papaioannou, 1999 Mapping and expression analysis of the mouse ortholog of *Xenopus Eomesodermin*. Mech. Dev. **81:** 205–208.

Hayata, T., A. Eisaki, H. Kuroda and M. Asashima, 1999 Expression of *Brachyury*-like T-box transcription factor, *Xbra3* in *Xenopus* embryo. Dev. Genes Evol. **209:** 560–563.

Holland, P. W. H., J. Garcia-Fernandez, N. A. Williams and A. Sidow, 1994 Gene duplications and the origins of vertebrate development. Development **120** (Suppl.): 125–133.

Holland, P. W. H., B. Koschorz, L. Z. Holland and B. G. Herrmann, 1995 Conservation of *Brachyury* (*T*) genes in amphioxus and vertebrates: developmental and evolutionary implications. Development **121:** 4283–4291.

Hug, B., V. Walter and D. J. Grunwald, 1997 *tbx6*, a *Brachyury*-related gene expressed by ventral mesendodermal precursors in the zebrafish embryo. Dev. Biol. **183:** 61–73.

Hughes, A. L., 1999 Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. J. Mol. Evol. **48:** 565–576.

Kumar, S., and S. B. Hedges, 1998 A molecular timescale for vertebrate evolution. Nature **392:** 917–920.

Li, W.-H., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Liang, F., I. Holt, G. Pertea, S. Karamycheva, S. L. Salzberg *et al.*, 2000 Gene index analysis of the human genome estimates approximately 120,000 genes. Nat. Genet. **25:** 239–240.

Lundin, L. G., 1993 Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics **16:** 1–19.

Martin, A. P., 1999 Increasing genomic complexity by gene duplication and the origin of vertebrates. Am. Nat. **154:** 111–128.

Meins, M., D. J. Henderson, S. S. Bhattacharya and J. C. Sowden, 2000 Characterization of the human *TBX20* gene, a new member of the T-box gene family closely related to the Drosophila H15 gene. Genomics **67:** 317–332.

Naruse, K., S. Fukamachi, H. Mitani, M. Kondo, T. Matsuoka *et al.*, 2000 A detailed linkage map of medaka, *Oryzias latipes*: comparative genomics and genome evolution. Genetics **154:** 1773–1784.

Nei, M., 1969 Gene duplication and nucleotide substitution in evolution. Nature **221:** 40–42.

Nei, M., X. Gu and T. Sitnikova, 1997 Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. USA **94:** 7799–7806.

Ohno, S., 1967 *Sex Chromosomes and Sex-Linked Genes*. Springer-Verlag, Berlin.

Ohno, S., 1970 *Evolution by Gene Duplication*. Springer-Verlag, New York.

Papaioannou, V. E., 2000 T-box genes in development: from hydra to humans. Intl. Rev. Cytol. (in press).

Papaioannou, V. E., and L. M. Silver, 1998 The T-box gene family. Bioessays **20:** 9–19.

Petrov, D. A., E. R. Lozovskaya and D. L. Hartl, 1996 High intrinsic rate of DNA loss in *Drosophila*. Nature **384:** 346–349.

Petrov, D. A., T. A. Sangster, J. S. Johnston, D. L. Hartl and K. L. Shaw, 2000 Evidence for DNA loss as a determinant of genome size. Science **287:** 1060–1062.

Prince, V. E., L. Joly, M. Ekker and R. K. Ho, 1998 Zebrafish *hox* genes: genomic organization and modified colinear expression patterns in the trunk. Development **125:** 407–420.

Roest Crollius, H., O. Jaillon, A. Bernot, C. Dasilva, L. Bouneau *et al.*, 2000 Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. Nat. Genet. **25:** 235–238.

Ruvinsky, I., and L. M. Silver, 1997 Newly identified paralogous groups on mouse chromosomes 5 and 11 reveal the age of a T-box cluster duplication. Genomics **40:** 262–266.

Ruvinsky, I., L. M. Silver and R. K. Ho, 1998 Characterization of the zebrafish *tbx16* gene and evolution of the vertebrate T-box family. Dev. Genes Evol. **208:** 94–99.

Ruvinsky, I., A. C. Oates, L. M. Silver and R. K. Ho, 2000 The evolution of paired appendages in vertebrates: T-box genes in the zebrafish. Dev. Genes Evol. **210:** 82–91.

Rzhetsky, A., and M. Nei, 1994 METREE: a program package for inferring and testing minimum-evolution trees. Comput. Appl. Biosci. **10:** 409–412.

Sidow, A., 1996 Gen(om)e duplications in the evolution of early vertebrates. Curr. Opin. Genet. Dev. **6:** 715–722.

Simmen, M. W., S. Leitgeb, V. H. Clark, S. J. M. Jones and A. Bird, 1998 Gene number in an invertebrate chordate, *Ciona intestinalis*. Proc. Natl. Acad. Sci. USA **95:** 4437–4440.

Skrabanek, L., and K. H. Wolfe, 1998 Eukaryote genome duplication—where's the evidence? Curr. Opin. Genet. Dev. **8:** 694–700.

Smith, N. G. C., R. Knight and L. D. Hurst, 1999 Vertebrate genome evolution: a slow shuffle or a big bang? Bioessays **21:** 697–703.

Szabo, S. J., S. T. Kim, G. L. Costa, X. Zhang, C. G. Fathman *et al.*, 2000 A novel transcription factor, *T-bet*, directs Th1 lineage commitment. Cell **100:** 655–669.

Terazawa, K., and N. Satoh, 1995 Spatial expression of the amphioxus homologue of *Brachyury* (*T*) gene during early embryogenesis of *Branchiostoma belcheri*. Dev. Growth Differ. **37:** 395–401.

Wada, H., and N. Satoh, 1994 Details of the evolutionary history from invertebrates to vertebrates, as deduced from the sequences of 18S rDNA. Proc. Natl. Acad. Sci. USA **91:** 1801–1804.

Wattler, S., A. Russ, M. Evans and M. Nehls, 1998 A combined analysis of genomic and primary protein structure defines the phylogenetic relationship of new members of the T-box family. Genomics **48:** 24–33.

Yi, C. H., J. A. Terrett, Q. Y. Li, K. Ellington, E. A. Packham *et al.*, 1999 Identification, mapping, and phylogenomic analysis of four new human members of the T-box gene family: *EOMES*, *TBX6*, *TBX18*, and *TBX19*. Genomics **55:** 10–20.

Zhang, J., and M. Nei, 1996 Evolution of Antennapedia-class homeobox genes. Genetics **142:** 295–303.

Communicating editor: C.-I Wu