# The amphioxus genome and the evolution of the chordate karyotype

Nicholas H. Putnam* [1,8], Thomas Butts [5], David E. K. Ferrier [14], Rebecca F. Furlong [5], Uffe Hellsten [1], Takeshi Kawashima [8], Marc Robinson-Rechavi [9,10], Eiichi Shoguchi [2], Astrid Terry [1], Jr-Kai Yu [4], Èlia Benito-Gutiérrez [15], Inna Dubchak[1], Jordi Garcia-Fernàndez [13], Jeremy J. Gibson-Brown[3], Igor V. Grigoriev [1], Amy C. Horton[3], Pieter J. de Jong [16], Jerzy Jurka[17], Vladimir Kapitonov[17], Yuji Kohara[18], Yoko Kuroki[6], Erika Lindquist [1], Susan Lucas [1], Kazutoyo Osoegawa[16], Len A. Pennacchio [1], Asaf A. Salamov[1], Yutaka Satou [2], Tatjana Sauka-Spengler[4], Jeremy Schmutz[12], Tadasu Shin-I[18], Atsushi Toyoda[6], Marianne Bronner-Fraser[4], Asao Fujiyama [6,11], Linda Z. Holland[7], Peter W. H. Holland [5], Nori Satoh* [2], Daniel S. Rokhsar* [1,8]

(*) Corresponding authors

[1] Department of Energy Joint Genome Institute, Walnut Creek CA 94598
[2] Department of Zoology, Graduate School of Science, Kyoto University, Sakyo-ku, Kyoto, 606-8502, Japan.
[3] Department of Biology, Washington University, St. Louis MO 63130, USA
[4] Division of Biology, California Institute of Technology, Pasadena CA 91125
[5] Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS UK
[6] RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan
[7] Marine Biology Research Division, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA 92093-0202
[8] Center for Integrative Genomics, Department of Molecular and Cell Biology, University of California, Berkeley CA 94720
[9] Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland
[10] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[11] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
[12] JGI Stanford Human Genome Center, 975 California Avenue, Palo Alto, CA 94304
[13] Departament de Genètica, Facultat de Biologia, Universitat de Barcelona
[14] The Gatty Marine Laboratory, University of St Andrews, St Andrews, Fife, KY16 8LB, UK
[15] National Institute for Medical Research, Mill Hill, London NW7 1AA, UK
[16] Children's Hospital of Oakland Research Institute, Oakland, CA 94609
[17] Genetic Information Research Institute, 1925 Landings Drive, Mountain View, CA 94043
[18] National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan

## Abstract

Lancelets ("amphioxus") are the modern survivors of an ancient chordate lineage with a fossil record dating back to the Cambrian. We describe the structure and gene content of the highly polymorphic ~520 million base pair genome of the Florida lancelet *Branchiostoma floridae*, and analyze it in the context of chordate evolution. Whole genome comparisons illuminate the murky relationships among the three chordate groups (tunicates, lancelets, and vertebrates), and allow reconstruction of not only the gene complement of the last common chordate ancestor, but also a partial reconstruction of its genomic organization, as well as a description of two genome-wide

duplications and subsequent reorganizations in the vertebrate lineage. These genome-scale events shaped the vertebrate genome and provided additional genetic variation for exploitation during vertebrate evolution.

## Introduction

Lancelets, or amphioxus, are small worm-like marine animals that spend most of their lives buried in the sea floor, filter-feeding through jawless, ciliated mouths. The vertebrate affinities of these modest creatures were first noted by Costa[1] and Yarrell[2] in the early part of the nineteenth century, and further clarified by the great Russian embryologist and supporter of Darwin, Alexander Kowalevsky[3]. In particular, Kowalevsky observed that, unlike other invertebrates, amphioxus shares key anatomical and developmental features with vertebrates and tunicates (also known as urochordates). These include a hollow dorsal neural tube, a notochord, a perforated pharyngeal region, a segmented body musculature (embryologically derived from somites), and a post-anal tail. Together, the vertebrates, urochordates, and lancelets (also known as cephalochordates) constitute the phylum Chordata, descended from a last common ancestor that lived perhaps 550 million years ago.

Although Kowalevsky, Darwin, and others recognized the evolutionary relationship between chordate groups, the greater morphological, physiological, and neural complexity of vertebrates posed a puzzle: how did the chordate ancestor -- presumably a simple creature that resembled a modern amphioxus or ascidian larva -- make such a transition?

Perhaps the most prevalent hypothesis for the origins of vertebrate complexity is founded on the ideas of Susumu Ohno (1970)[4], who proposed that vertebrate genomes were shaped by a series of ancient genome-wide duplications. In Ohno's original proposal, lancelet and vertebrates genomes were enlarged relative to the basic invertebrate complement by one or two rounds of genome doubling, although subsequent work suggested that these events occurred on the vertebrate stem after divergence of the lancelet lineage [5,6].

Although the sequencing of the human and other vertebrate genomes has shown that the gene number in vertebrates is comparable to, or only modestly greater than, that of invertebrates[7,8], evidence for large-scale segmental or whole genome duplications on the vertebrate stem has mounted, with the parallel realization that most gene duplicates from such events are rapidly lost (reviewed in [9]). The relatively few surviving gene duplicates from the vertebrate stem provide

evidence for ancient paralogous relationships between groups of human chromosomes[10-14] that plausibly arose from multiple rounds of whole genome duplication prior to the emergence of modern vertebrates. Yet the number, timing, and even genomic scale of the duplication events, and their consequences for subsequent genome evolution are poorly understood (for a review see [15]), in part because the tunicate genomes are highly rearranged relative to the un-duplicated early chordate karyotype (see below).

The Florida lancelet *Branchiostoma floridae* (the generic name refers to the characteristic perforated branchial arches) provides a critical point of reference for these studies [16]. This species and its relatives (collectively also known as amphioxus, derived from the Greek amphi+oxus, "sharp at both ends") are widely regarded as living proxies for the chordate ancestor, in part due to the general similarity of modern amphioxus to putative fossil chordates from the early Cambrian Chengjiang fauna (*Yunnanozoon lividum* [17], and the similar *Haikouella lanceolatum* [18]) and the middle Cambrian Burgess Shale (*Pikaia gracilens*[19]), although controversy remains (see, *e.g.*, [20,21]). The study of key developmental genes in amphioxus has shed light on the evolution of such vertebrate organs as the brain, kidney, pancreas, and pituitary, and of the genetic mechanisms of early embryonic patterning in general (reviewed by [22-24]). Amphioxus has also served as a genomic surrogate for the proto-vertebrate ancestor in studies of the Hox cluster[25], in studies of specific genomic regions[26-28], and as an outgroup in numerous gene family studies (reviewed in [29]).

Here we report the draft genome sequence of the Florida lancelet and compare its structure with the genomes of other animals. Robust phylogenetic analysis of gene sequences and exon-intron structures confirms recent proposals that tunicates are the sister group to vertebrates, with lancelets as the most basal chordate subphylum, and that the combined echinoderm-hemichordate clade is sister to chordates. Through comparative analysis we identify 17 ancestral chordate linkage groups that are conserved in the modern amphioxus and vertebrate genomes despite over half a billion years of independent evolution. Over 90% of the human genome is encompassed within these linkage groups, which display a tell-tale four-fold redundancy that is consistent with whole genome quadruplication on the vertebrate stem. Comparison with sequences from the sea squirt, lamprey, elephant shark, and several bony fish constrains the timing of the whole genome events to after the divergence of vertebrates from tunicates and lancelets, but before the split between cartilaginous and bony vertebrates. Within the resolution of our analysis we find evidence for rounds of genome duplication both before and after the split between jawless vertebrates (*e.g.*, lamprey) and jawed vertebrates, although a period of octoploidy encompassing the divergence of jawless and jawed vertebrates remains a possibility. While most duplicate genes from these whole

genome events have been lost, a disproportionate number of genes involved in developmental processes are retained.

## Genome sequence

We sequenced the ~520 Mb amphioxus genome using a whole genome shotgun strategy[30] from approximately 11.5-fold redundant paired-end sequence coverage produced from random-sheared libraries with a range of insert sizes (Supplementary Note 2). Genomic DNA was prepared from the gonads of a single gravid male collected from Tampa Bay, Florida in July 2003, and contained significant allelic variation (3.7% single nucleotide polymorphism, plus 6.8% polymorphic insertion/deletion (Supplementary Note 4)). This is the highest level of sequence variation reported in any individual organism, exceeding that found in the purple sea urchin[31]. Assembly version 1 reports both haplotypes assembled separately, while in assembly version 2 a single haplotype is selected at each locus (Supplemental Figure 63). Assembly version 2 spans 522 Mb, with half of this sequence in 62 scaffolds longer than 2.6 Mb.

Currently there are no physical or genetic maps of amphioxus, so we could not reconstruct the genome as its 2N=38 chromosomes[32]. Nevertheless, since half of the predicted genes are contained in scaffolds containing 138 or more genes, the current draft assembly is sufficiently long-range to permit useful analysis of conserved synteny with other species, as shown below. Comparison of the assembled sequence with open reading frames derived from expressed sequence tags (ESTs, see below) shows that the assembly captures more than 95% of the known protein coding content, and comparison to finished clone sequences demonstrates the base-level and long-range accuracy of the assembly.

## Annotation of protein coding genes and transposable elements

We estimate that the haploid amphioxus genome contains 21,900 protein-coding loci. This gene complement was modeled with standard methods tuned for amphioxus, integrating homology and *ab initio* gene prediction methods with over 480,000 expressed sequence tags (ESTs) derived from a variety of developmental stages[24]. (Supplementary Note 3). Approximately two thirds of the protein coding loci (15,123) are captured in both haplotypes. Transposable elements constitute ~30% of the amphioxus genome assembly (Supplementary Table S5) and belong to >500 families. Based on their bulk contribution to the genome size, DNA transposons are twice more abundant than retrotransposons.

## Polymorphism

The distribution of observed local heterozygosity over short length scales obeys a geometric distribution (Supplementary Figure S1), consistent with the prediction of the random mating model, as observed in *Ciona savignyi*[33], with a population mutation rate $4 \mu N_e = 0.0562$, where $\mu$ is the per generation mutation rate and $N_e$ is the effective population size. High heterozygosity can, in principle, be explained by 1) a large effective population size maintained over many generations, 2) a high mutation rate per generation, or 3) the recent mixing of previously isolated populations; in the latter case, a geometric distribution of local heterozygosity would not be expected. Assuming a typical metazoan mutation rate on the order of one to ten substitutions per gigabase per generation[34,35] this observed heterozygosity between alleles suggests a large but plausible effective breeding population on the order of millions of individuals. The observed heterozygosity shows correlations at short distances that decay on scales greater than ~1kb, indicating extensive recombination in the population (Supplementary Figure S2). An analysis of the ratio Ka/Ks of non-synonymous to synonymous substitutions shows evidence of purifying selection comparable to that found between mammalian species (Supplementary Note 4). Indel polymorphisms are common, as found in other intra- and intergenome comparisons[36] (Supplementary Figure S3). Structural variation between haplotypes also includes local inversions and tandem duplications (Supplementary Figure 63).

## Deuterostome relationships

With the draft amphioxus sequence in hand, we reconsidered the phylogenetic relationships between deuterostome phyla (chordates, echinoderms, and hemichordates). The traditional placement of lancelets as sister to vertebrates, with tunicates as the earliest diverging chordate subphylum, has recently been questioned [37,38]. A preliminary study[39] using 146 gene loci (33,600 aligned amino acid positions) and trace data from the present amphioxus genome project found support for tunicates as the sister to vertebrates. This analysis, however, also suggested (albeit with limited statistical support) that amphioxus is more closely related to echinoderms than to tunicates or vertebrates, which would render chordates a polyphyletic group. A second study with a similarly sized set of genes, but more diverse deuterostome taxa, supported the early branching of the cephalochordate lineage, but not the close relationship of amphioxus and echinoderms[40].

To address the controversial phylogenetic position of amphioxus, we analyzed a much larger set of

1,090 orthologous genes (see Supplementary Note 5). Both Bayesian and maximum likelihood methods support the new chordate phylogeny [38-40] in which cephalochordates represent the most basal extant chordate lineage, with tunicates (represented by both *Ciona* and *Oikopleura* in our analysis) sister to vertebrates but with long branches that indicate higher levels of amino acid substitution. Individual gene trees also lend support to this topology; genes supporting tunicates as sister group to vertebrates outnumber those with amphioxus in this position by a 2:1 ratio. An analysis of intron gain and loss in deuterostomes provides independent support for amphioxus as the basal extant chordate subphylum (see below). We group echinoderms (*i.e.*, the purple sea urchin) and hemichordates (*i.e.*, the acorn worm) together (ambulacrarians) as sister to a monophyletic chordate clade, as in [15,41] but in contrast to the suggestion of [39]. With the exception of the long-branched tunicates, the maximum likelihood tree suggests a roughly constant evolutionary rate of peptide change across the deuterostome tree, although an excess of substitutions is found in the vertebrates relative to the predictions of a simple molecular clock model.

**Intron evolution**

To assess the evolution of gene structure within the deuterostomes and chordates, we compared the position and phase of amphioxus introns to other animals. Amphioxus and human (along with other vertebrates) share a large fraction of their introns (85% in alignable regions), which match precisely in both position and phase (Supplementary Note 6), as also found in the sea anemone *Nematostella vectensis*[42]. We find that the intron-rich gene structures of the eumetazoan ancestor were carried forward to the common chordate ancestor with relatively few gains or losses. The tunicates *Ciona intestinalis* and *Oikopleura dioica*, however, share many fewer introns with vertebrates[43] or amphioxus.

Notably, intron presence or absence carries a significant (as measured by bootstrap values) phylogenetic signal, and Bayesian analysis of the associated character matrix supports the sister relationship between tunicates and vertebrates (Supplementary Figure S8; Methods). This is evidently due to shared gain or loss of introns along the stem group leading to their common ancestor, which remarkably is still detectable despite additional extensive secondary losses, and modest gains, in the tunicate lineages. Thus intron dynamics provide independent support for the new chordate phylogeny.

**Chordate gene families and novelties**

Through comparison of the amphioxus gene set with those of other animals, we identified 8,437 chordate gene families with members in amphioxus and other chordates that each nominally represent the modern descendants of a single gene in the last common chordate ancestor (Supplementary Note 7). That ancestor certainly possessed more genes than this number, but the others are inaccessible to us now due to subsequent sequence divergence and/or gene loss in the living chordates. Through subsequent gene family expansions (via both local and/or genome-wide duplications), these families account for 13,610 amphioxus gene loci, 13,401 human genes, and 7,216 C*iona intestinalis* genes. The dramatically lower number of descendant genes in *Ciona* is due in large part to gene loss [44], with the present analysis identifying 2,251 ancient chordate genes missing in this genome sequence. We found 8 apparent chordate stem gene losses (*i.e.*, genes found in sea urchin and fly or *Nematostella*, but not in vertebrates, amphioxus, or *Ciona intestinalis*). A list of these genes can be found in Supplementary Table S10.

We identified 239 apparent chordate gene novelties, that is, gene families represented in amphioxus and at least one vertebrate or *Ciona*, but without an obvious direct counterpart in non-chordate genomes. These can be characterized[42] as 137 families with no detectable sequence similarity to non-chordate genes (type I novelties), 10 containing one or more chordate-specific domains linked to pre-existing metazoan domains (type II novelties), and 92 with chordate-specific combinations of pre-existing metazoan domains (type III novelties). (See Supplementary Note 7.) These gene families and others of special interest to vertebrate biology are discussed in a separate paper (Holland, 2008 #353).

**Synteny**

We find extensive conservation of gene linkage on the scale of whole chromosomes (macro-synteny) between the amphioxus genome and those of vertebrates (represented in our analysis by human, chicken, and teleost fish), but only limited conservation of local gene order (micro-synteny). Through comparative analysis of these conserved features, we reconstructed the gene complements of seventeen linkage groups (i.e., proto-chromosomes) of the last common chordate ancestor. When vertebrate genomes are analyzed in the light of these putative ancestral chordate chromosomes, a clear pattern of global four-fold conserved macro-synteny is found, demonstrating that two rounds of whole genome duplication occurred on the vertebrate stem.

**Amphioxus-vertebrate synteny and the reconstruction of chordate linkage groups**

To identify ancestral chordate linkage groups, we first noted that many individual amphioxus scaffolds show conserved syntenic associations with human chromosomes, reflecting conserved linkage between the two genomes (Figure 2; see also Oxford Grid in Supplementary File 5. For simplicity, we emphasize the amphioxus-human comparison in the main text, and include similar results for chicken, stickleback, and pufferfish as Supplemental Material.) 96 scaffolds (out of 129 that possess twenty or more independent vertebrate orthologs) have a significant ($p<0.05$, multiple test corrected) concentration of orthologs on one or more human chromosomes. By contrast, in the comparable comparison to human chromosomes, only 12 *Ciona intestinalis* scaffolds (out of 134 that contain twenty or more vertebrate orthologs) show a significant concentration.

Genes on individual amphioxus scaffolds have orthologs that are generally concentrated in specific regions of vertebrate chromosomes (Figure 2). Furthermore, multiple amphioxus scaffolds typically exhibit hits to the same sets of human chromosomal regions. Within each region, only limited conservation of gene order is observed (Methods; Supplementary Note 8). This pattern of conserved synteny shows that genome rearrangements have not erased the imprint of the genome organization of the last common chordate ancestor from the present human and amphioxus genomes. By using this pattern, we identified 135 human chromosomal segments (listed in Supplemental Table S14) that retain relict signals of the ancestral chordate karyotype despite chromosomal rearrangements in each lineage (Methods). These segments span a mean of 170 genes.

We exploited the pattern of amphioxus-human synteny to identify seventeen ancient chordate linkage groups (CLGs) by clustering both amphioxus scaffolds and human chromosomal segments according to their pattern of hits in the other genome (Methods). The resulting “dot plot” (Supplementary Figure 64) shows that orthologs are concentrated in seventeen distinct blocks. Within each block gene order is considerably scrambled. The natural interpretation of these blocks is that each represents an ancient chordate linkage group that evolved into a defined group of chromosomal segments in amphioxus, human, chicken, and teleost fish.

We tested our interpretation of the chordate linkage groups as coherently evolving segments by using fluorescent in situ hybridization to demonstrate that fifteen of sixteen scaffolds from CLG #15 localized to a single amphioxus chromosome. (FISH of the BACs corresponding to the sixteenth scaffold were ambiguous. Supplementary Table S2) Similarly, an independent study of amphioxus cosmids containing NK group homeobox genes in CLG #7 found that they localize to several distant regions of a single chromosome in amphioxus [27,28]. These data support the claim that the 17

putative ancestral chordate linkage groups have been maintained in modern amphioxus as coherent chromosomal segments. The 19 pairs of modern amphioxus chromosomes, however, imply at least (see below) two subsequent fissions in the amphioxus lineage.

Within these segments, nearly 60% of the human genes that possess amphioxus orthologs participate in the conserved linkage groups. This represents a lower bound, since short amphioxus scaffolds are less likely to be assigned to CLGs. Conversely, 88% of amphioxus gene models on scaffolds assigned to a CLG have their human ortholog in conserved position (*i.e.*, in the same CLG). Remarkably, to the resolution of our analysis, some entire chromosomes (*e.g.* human 18 and 21; chicken 7, 12, 15, 19, 21, 24, 27), and chromosome arms (including human 5p) appear to have maintained their integrity (with local scrambling and some gene gain and loss) since the last common chordate ancestor. The CLG's defined by comparing amphioxus and vertebrate genomes also provide a new perspective on tunicate genome evolution, since it appears that *C. intestinalis* chromosomes 10, 12, and 14 are each relicts of a single CLG (11, 5, and 8, respectively), and other conserved linkages are evident (Supplemental Figure S14).

**Whole genome duplication and quadruple conserved synteny.**

We can trace the evolution of chordate genomes through time using two additional types of evidence. First, we can constrain the timing of specific chromosome breaks through parsimony analysis of conserved synteny across human, chicken, and teleost genomes. Second, we can use the presence (or absence) of paralogous gene pairs to identify segments derived from the same chordate proto-chromosome by duplication (or fission).

For example, five human segments from chromosomes 1, 5, 9 and 19 (segments 1.5/7, 5.1, 9.1/3, 19.1/3, and 19.2) cluster together in CLG #1 of Supplementary Figure 64. Segment 1.5/7 is related to each of the others by a significant concentration of ancient gene paralogs (17 to 31 pairs, $p < 1e-10$), indicating that it is related to the other segments by duplication. In contrast, only a single pair of ancient paralogs relate segments 5.1 and 9.1/3, and orthologs of the genes in these segments occur predominantly on the same chromosomes of both pufferfish and stickleback. Thus 5.1 and 9.1/3 were likely created by breakage of a single ancestral segment of the bony vertebrate ancestor. If 5.1 and 9.1/3 are virtually merged, then all remaining pairings of human segments from CLG #1 show a significant excess of ancient paralogs, consistent with amplification to four through two successive duplications.

To obtain a genome-wide view of the history of chromosomal evolution on the vertebrate stem, we applied a similar analysis systematically to the 17 CLGs by exhaustively evaluating all partitionings of human genome segments, and using a parsimony criterion to identify the most likely

reconstruction. The most parsimonious partitionings of human segments into paralogy groups is summarized in Supplemental Table 1, and diagrammed in Figure 3. This analysis shows that the vast majority of the human genome (112 segments spanning 2.68 Gbp, or 95% of the euchromatic genome) was affected by large scale duplication events on the vertebrate stem prior to the bony-vertebrate radiation (*i.e.*, the teleost/tetrapod split), and that nearly all the ancient chordate chromosomes were expanded four-fold. (Supplemental Figure S9)

This pattern of genome-wide quadruple conserved synteny[15] definitively demonstrates the occurrence of two rounds of whole genome duplication (2R) and provides a comprehensive reconstruction of the evolutionary origin of the human chromosomes (and those of other jawed vertebrates) through these duplications on the vertebrate stem. This characterization extends previous lines of evidence for whole genome duplication events based on comparative studies of specific regions of interest across chordate genomes (e.g., the Hox cluster [25] and MHC-region [28,45]) and the analysis of vertebrate gene families (reviewed in [29], as well as the identification of paralogous segments and chromosomal relationships within the human genome [10,13,14,46]. A manual, phylogeny-based analysis of the four scaffolds making up the NK-containing paralogon was in agreement with these results (Methods).

**Timing of events on the vertebrate stem**

The amphioxus-human synteny analysis presented above demonstrates that two rounds of whole genome duplication occurred on the vertebrate stem after the divergence of cephalochordates but before the split of teleosts and tetrapods. The next question is whether these two genome-scale duplications happened in rapid succession or even effectively simultaneously, or were separated in time [15]. We sought to resolve the 2R events relative to the divergence of cartilaginous fish, urochordates, and jawless vertebrates (*e.g.*, lamprey).

Sample sequencing from the elephant shark *Callorhynchus milii*, for example, demonstrates significant conserved macrosynteny between cartilaginous fish and humans, as pairs of genes that are ~35-40 kb apart in the elephant shark genome are also linked on the human genome [47]. These links occur predominantly within the human segments defined above, indicating that the orthologous chromosome segments are also found in the elephant shark genome (Supplementary Note 9). Furthermore, previous analysis of phylogenetic topologies dated all duplications prior to the split between cartilaginous and bony vertebrates [48]. 2R therefore occurred prior to this split [47]. Similarly, the preservation of several CLGs as intact single chromosomes in *Ciona intestinalis* (Supplementary Figure 14) implies that both rounds of duplication occurred after the divergence of the urochordate lineage.

Sequencing of the repeat-rich lamprey genome has not generated enough long scaffolds to permit

large-scale analysis of synteny [49]. To infer the timing of 2R relative to the divergence of the lamprey lineage, we generated a set of ~50,000 expressed sequence tags (ESTs) from the sea lamprey *Petromyzon marinus* (Supplementary Note 3) and analyzed the phylogenetic topology of 358 gene families that include pairs of synteny-confirmed human paralogs produced during 2R. The results are summarized in Table 1, along with a parallel analysis using *Ciona* and *Fugu* for comparison. We find that ~58% of the resolved four-gene phylogenies involving a lamprey, amphioxus, and two randomly chosen human paralogs arising from 2R place the lamprey gene closer to one of the human paralogs, similar to the results of [50] but analysing a ten-fold larger set of gene families (Figure 4). This result is clearly distinct from that expected if the lamprey lineage diverged either well before (cf. *Ciona*) or well after (cf. *Fugu*) 2R. The remaining scenarios are that either (a) the jawed vertebrate and lamprey lineages diverged in the period between two well-separated whole genome duplications, or (b) one, or both, of the 2R whole genome events occurred nearly coincident with the lamprey lineage divergence. The time interval that distinguishes "nearly coincident" from "well-separated" is determined by that the process of rediploidization, during which most gene duplicates are lost and the sequences of the surviving paralogs diverge.[9]

**Subsequent karyotypic changes in the vertebrate and tunicate lineages**

From the 17 ancestral chordate linkage groups, 2R nominally produced 17 x 4 = 68 proto-vertebrate segments, although this naive inference assumes that (a) all duplicated segments were retained and (b) no fusions, fissions, or additional segmental duplications occurred during 2R. Some 2R-produced segments, however, are consistently linked in contemporary bony vertebrate genomes (for example 12b, 1d which co-occur on human chromosome 1, chicken chromosome 8 and stickleback linkage groups III and VIII), indicating a fusion prior to the bony vertebrate (osteichthyan) ancestor. We find evidence for at least 20 such fusions (Supplementary Note 8). Allowing for a range of nearly parsimonious reconstructions of 2R, we estimate that the bony vertebrate ancestor had between 37 and 49 chromosomes. Additional fusions on the teleost stem reduced this number to 12 [51-55] prior to the teleost-specific genome duplication. On the tetrapod stem, the chicken and human genomes share 4 fusions of bony vertebrate segments, suggesting 33-45 chromosomes. These are consistent with recent estimates based on intra-vertebrate comparisons [46,51-55].

***Ancient developmental gene linkages***

The amphioxus genome has also retained ancient local gene linkages (micro-synteny) in addition to conserved macro-synteny. In some cases local linkages are even older than the chordates, and date back to the bilaterian ancestor or earlier. As an example, we considered gene families that expanded by tandem duplication early in animal evolution, specifically, the ANTP and PRD classes

of homeobox genes and the Wnt gene family. We examined how frequently these genes are still neighbors in the amphioxus genome, and discovered five new examples of ancient pairs or clusters of ANTP or PRD genes: Otx/goosecoid, Mnx/ro, Nkx2-1/Nkx2-2, Nkx6/Nkx7/Lbx/Tlx, En/Nedxa/Nedxb/Dll (Supplementary Table S3). These gene pairs or clusters (along with the well-known Hox, ParaHox and NK linkages) originated by tandem duplication before the divergence of bilaterians, yet their tight linkage has not been disrupted by genome rearrangement. The Nkx2-1/Nkx2-2 gene pair has been retained in vertebrates (and is duplicated), but in every other example the tight linkage (clustering) has been lost in the human genome. None of the five newly described examples are retained in the *Drosophila melanogaster* genome. The situation in the Wnt gene family is a little different, since both amphioxus and *Drosophila* have retained tight linkages that have been disrupted in the human lineage due to genome duplication followed by differential gene loss (Figure 4). These results underscore the fact that the amphioxus genome has undergone less genomic rearrangement than the human and other vertebrate genomes since their shared ancestor over half a billion years ago.

**Impact of whole genome duplications on the modern vertebrate gene repertoire**

How many duplicate genes survive in modern vertebrate genomes from the two genome-wide events? 2,131 (25%) of the ancestral chordate gene families (out of the 8,437 indicated above) have two or more ancient vertebrate paralogs ("Ohnologs") that were evidently produced by ancient gene duplication(s) after the divergence of amphioxus. (See also Supplementary Fig 3.) Of these, 1,489 (70%) are embedded within paralogous segments from our reconstruction of 2R, as portrayed in Figure 3, and were plausibly created through 2R. These retention rates for 2R-duplicated genes are comparable to other estimates based on large scale gene phylogenies [10,14,56,57]. Similar retention rates are found for the ~350 My old teleost-fish-specific duplication [54,58-60] and estimated for the ~40 My old *Xenopus* (frog) -specific duplication[61].

Gene duplicates from 2R that have been retained in modern genomes are significantly enriched for functions associated with signal transduction, transcriptional regulation, neuronal activities, and developmental processes (Supplementary Table 18, Methods). For example, genes implicated in signal transduction are more than twice as likely to be retained in two or more copies from 2R compared to the overall retention rate. These results are consistent with the hypothesis that paralogs created by whole genome duplication were recruited for roles in the development of novel features of vertebrate biology, and with similar biased retention in teleost fishes[60]. Whole genome duplications, however, may have allowed entire molecular pathways to be duplicated and sub-

functionalized coincidently (reviewed in: [62]). While similar numbers of gene duplicates are found in amphioxus relative to the chordate ancestor, different gene classes have been expanded, and the mechanism of gene duplication is different. (Supplementary Note 7).

**Conserved non-coding sequences**

Inspired by the extensive conserved synteny between amphioxus and vertebrates, we searched for conserved non-coding sequences that might reflect ancient chordate cis-regulatory elements. While genome-scale comparisons between mammals and teleost fish have revealed up to 3,100 conserved non-coding sequences, the majority of which function as tissue-specific enhancers.[63,64] At greater phylogenetic distances no conservation outside of coding sequences and conserved microRNAs [65] has so far been identified. By aligning the amphioxus and human genomes (Supplementary Note 10), 77 putative chordate conserved non-coding elements were identified (>60% identity over >50 bp), after excluding transcribed or repetitive sequences and requiring conservation in at least one other vertebrate. Of these, 16 overlap with or are immediately adjacent to the 3' or 5' untranslated regions (UTRs) of human genes, and are likely conserved UTR elements. Four are adjacent to exons and represent likely conserved splicing enhancers. A single conserved noncoding element overlapped a highly conserved microRNA (mir-10b adjacent to the human HOXD4 gene). The remaining 56 elements are of unknown function, but can be tested experimentally for enhancer activity [66].

**Conclusions**

The amphioxus sequence reveals key features of the genome of the last common ancestor of all chordates through comparison with the genomes of other animals. This ancestor likely lived before the Cambrian, and gave rise to the chordate lineage that is represented today by modern cephalochordates like amphioxus, as well as urochordates and vertebrates. Of the living lineages, the cephalochordates diverged first, prior to the split between the morphologically diverse urochordates and vertebrates. To a remarkable extent, the amphioxus genome appears to be a good surrogate for the ancestral chordate genome with respect to gene content, exon-intron gene structure, and even chromosomal organization. The sequences of model ascidians with small genomes like *Ciona* and *Oikopleura* are by comparison simplified by gene loss, intron loss, and genome rearrangement. Remarkably, modest levels of non-coding sequence have been conserved between amphioxus and human -- the oldest conserved non-coding regions yet detected through direct sequence alignment -- and may provide a tantalizing glimpse of the gene regulatory systems of the chordate ancestor.

Extensive conserved synteny between the genomes of amphioxus and various vertebrates lends unprecedented clarity and coherence to the history of genome-scale events in vertebrate evolution.

The human and other jawed vertebrate genomes show widespread quadruple-conserved synteny relative to the amphioxus sequence, which extends earlier regional studies and provides a unified explanation for paralogous chromosomal regions in vertebrates. Our analysis thus provides conclusive evidence for two rounds of complete genome duplication on the jawed vertebrate stem. These genome duplications occurred after the divergence of urochordates but before the split between cartilaginous fish and bony vertebrates. The jawless vertebrates (*e.g.*, lamprey and hagfish) represent the only other chordate lineages that survive from this period, and at least the lamprey appears to have diverged between the two rounds of duplication, although the data still allow for an octoploid population as the progenitor of the jawless and jawed vertebrates. The detailed mechanism of these events -- in particular, whether they occurred by allo- and/or autotetraploidizations, how closely spaced in time they were, and the precise nature of the rediploidization process -- remain unknown. While it is tempting to relate the genome duplications to specific morphological radiations in vertebrate evolution, the fossil record reflects a relatively steady diversification rather than a dramatic discontinuity of stem-group vertebrate forms[67]. The genomic features that are associated with organismal complexity, if such generic features exist at all, remain unknown. [68] It is tempting to speculate, however, that the creation of the proto-jawed-vertebrate genome by two rounds of genome duplication was a formative event in the early history of vertebrates that provided genomic flexibility through the duplication of coding and cis-regulatory sequences for the emergence of familiar developmental, morphological, and physiological novelties such as chondrogenic and skeletogenic neural crest cells, the sclerotome (vertebral) compartment of the somites, elaborate hindbrain patterning, finely graded nervous system organization, and the elaborated endocrine system of vertebrates. Indeed, we find that genes involved in developmental signaling and gene regulation are significantly more likely to be retained in multiple copies in living species than genes overall, suggesting that diversified developmental regulation is correlated with the evolution of vertebrate novelties. This begs the question, dating back to Ohno, of how such duplicated genes became integrated into the biochemical and genetic networks of vertebrates. In a separate paper [66], we examine vertebrate biology in the light of the amphioxus genome data and the genome-scale duplication events on the vertebrate stem.

## Methods Summary

**Genome Sequencing, Assembly and Annotation:** 7.3 million high-quality sequence Sanger reads were generated and assembled using JAZZ[69]. Protein coding genes were annotated using expressed sequence tag, homology and ab initio methods as previously described [42,70].

**Deuterostome relationships** Orthologous gene alignments were created using CLUSTALW[71], Gblocks[72], and analyzed with Bayesian and maximum likelihood methods.

**Intron evolution** The presence and absence of an intron at each of 5,337 orthologous coding positions was treated as a binary character in parsimony and Bayesian analysis.

**Construction of Chordate Linkage Groups** Human chromosome segments and amphioxus scaffolds were clustered by ortholog distribution profile. The null hypothesis (orthologous genes random distributed across the two genomes) was evaluated using Fishers exact test, with a Bonferroni correction for the total number of pairwise tests.

**Decomposition of CLGs into independent products of duplication** The most parsimonious partitioning of the human chromosomal segments assigned to the CLG was obtained using a scoring system that included shared ohnologs and position in the multi-species synteny comparison.

**NK quadruple conserved synteny** In addition to the genome-wide synteny analysis, a detailed manual curation was carried out on four v1 scaffolds (56, 124, 185, and 294) that make up the NK homeobox cluster in amphioxus. The 82 amphioxus genes correspond to 111 human genes enriched on chromosomes 4, 5, 8 and 10 (chi-squared test p«0001), in agreement with the genome-wide analysis of CLG #7. (Supplementary Note 11)

**Ancient Developmental Gene Linkages** Orthology of homeodomain and Wnt containing genes was assigned from phylogenetic tree reconstruction using neighbor-joining and maximum likelihood approaches, supported by high bootstrap values.

**Tables**

**Table 1. Timing of whole genome duplications relative to the divergence of chordate groups.**

| X | N attempted | N resolved | N "in" | % in |
|---|---|---|---|---|
| Ciona | 736 | 273 | 34 | 12 ± 2 % |
| Lamprey | 358 | 159 | 93 | 58 ± 6 % |
| Fugu | 1009 | 389 | 351 | 90 ± 5 % |

**Figure Captions:**

**FIGURE 1. Deuterostome phylogeny.** Bayesian phylogenetic tree of deuterostome relationships with branch length proportional to the number of expected substitutions per amino acid position, using a concatenated alignment of 1,090 genes. The scale bar represents 0.05 expected substitution per site in the aligned regions. Long branches for *Ciona* and *Oikopleura* indicate high levels of amino acid substitution. This tree topology was observed in 100% of sampled trees (See Supplemental Note 5). Numbers in red indicate bootstrap support under maximum likelihood. Unlabeled nodes were constrained.

**FIGURE 2. Amphioxus-human synteny.** Four amphioxus scaffolds from the non-redundant version 2 assembly with synteny to human chromosome 17. Note that orthologous genes from these scaffolds are concentrated in specific regions of the chromosome, and that several scaffolds (*e.g.*, 18 and 162, or 149 and 207) have a high density of hits to the same segments of the chromosome, which enables a partitioning of the human genome into 135 ancient segments. Supplementary File 5 contains an Oxford grid tabulating the number of orthologs for each scaffold-segment pair.

**Figure 3. Quadruple conserved synteny.** Partition of the human chromosomes into segments with defined patterns of conserved synteny to amphioxus scaffolds.

**Figure 4. Ancient Developmental Gene Linkages.** In the amphioxus genome, Wnt6, Wnt1, and Wnt9 form a compact gene cluster, 2.5 Mbp from Wnt10 and Wnt3 but all on scaffold 12. Orthologs of four of these genes are also clustered in *Drosophila* (not shown), although Wnt3 has been lost, inferred from its presence in cnidarians. The human orthologs are on four chromosomes, and show disruption of gene clustering through duplication followed by gene loss. Linkages of Hox clusters to three of the human loci gives additional support for the large-scale duplication events involved. The three clusters which are linked to Hox clusters (as well as the four Hox clusters) fall in chromosome segments grouped in CLG #16, as well as the Hox-bearing amphioxus scaffold. Genes drawn as boxes above the lines are transcribed from left to right; genes depicted below lines are transcribed from right to left.

**Acknowledgments**

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396. The Center for Integrative Genomics is supported by a grant from the Gordon and Betty Moore Foundation. D.S.R. acknowledges support from R.A. Melmon. This work was funded by grants from MEXT, Japan (N.S., A.F., Y.K., H.W.), the 21th Century and Global COEs at Kyoto University (N.S.), BBSRC (T.B & D.E.K.F.), and the Wellcome Trust (P.W.H.H.).

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**References**

1. Costa, O. G. Cenni zoologici, ossia descrizione delle specie nuove di animali discoperti in diverse coutrade del regno nell' anno 1834. *Annuario Zoologico* **12**, 49 (1835).
2. Yarrell, W. in *A history of British fishes.* 468-472 (J. Van Voorst, London, 1836).
3. Kowalevsky, A. Entwicklungsgeschichte des Amphioxus lanceolatus. *Memoires de l'Academie des Science de St Petersbourg* **(7)11**, 1-17 (1866).
4. Ohno, S. *Evolution by gene duplication* (Allen & Unwin; Springer-Verlag, London, New York,, 1970).
5. Schmidtke, J., Weiler, C., Kunz, B. & Engel, W. Isozymes of a tunicate and a cephalochordate as a test of polyploidisation in chordate evolution. *Nature* **266**, 532-3 (1977).
6. Holland, P. W., Garcia-Fernandez, J., Williams, N. A. & Sidow, A. Gene duplications and the origins of vertebrate development. *Dev Suppl*, 125-33 (1994).
7. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
8. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304-51 (2001).
9. Wolfe, K. H. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* **2**, 333-41 (2001).
10. Popovici, C., Leveugle, M., Birnbaum, D. & Coulier, F. Coparalogy: physical and functional clusterings in the human genome. *Biochem Biophys Res Commun* **288**, 362-70 (2001).
11. Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P. & Inoko, H. Evidence of en bloc duplication in vertebrate genomes. *Nat Genet* **31**, 100-5 (2002).
12. McLysaght, A., Hokamp, K. & Wolfe, K. H. Extensive genomic duplication during early chordate evolution. *Nat Genet* **31**, 200-4 (2002).
13. Lundin, L. G., Larhammar, D. & Hallbook, F. Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. *J Struct Funct Genomics* **3**, 53-63 (2003).
14. Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3**, e314 (2005).
15. Furlong, R. F. & Holland, P. W. Were vertebrates octoploid? *Philos Trans R Soc Lond B Biol Sci* **357**, 531-44 (2002).
16. Stokes, M. D. & Holland, N. D. The lancelet. *American Scientist* **86**, 552-560 (1998).
17. Chen, J.-Y., Dzik, J., Edgecombe, G. D., Ramskold, L. & Zhou, G.-Q. A possible Early Cambrian

chordate. *Nature* **377**, 720-2 (1995).
18. Chen, J.-Y., D-Y, H. & C-W, L. An early Cambrian craniate-like chordate. *Nature* **402**, 518-522 (1999).
19. Conway Morris, S. & Whittington, H. B. The Animals of the Burgess Shale. *Sci Am* **240**, 122-133 (1979).
20. Shu, D., Zhang, X. & Chen, L. Reinterpretation of *Yunnanozoon* as the earliest known hemichordate. *Nature* **380**, 428-430 (1996).
21. Mallatt, J. & Chen, J. Y. Fossil sister group of craniates: predicted and found. *J Morphol* **258**, 1-31 (2003).
22. Holland, L. Z. & Holland, N. D. Chordate origins of the vertebrate central nervous system. *Current opinion in neurobiology* **9**, 596-602 (1999).
23. Holland, N. D. & Chen, J. Origin and early evolution of the vertebrates: new insights from advances in molecular biology, anatomy, and palaeontology. *Bioessays* **23**, 142-51 (2001).
24. Yu, J. K. et al. Axial patterning in cephalochordates and the evolution of the organizer. *Nature* **445**, 613-7 (2007).
25. Garcia-Fernandez, J. & Holland, P. W. Archetypal organization of the amphioxus Hox gene cluster. *Nature* **370**, 563-6 (1994).
26. Castro, L. F. & Holland, P. W. Chromosomal mapping of ANTP class homeobox genes in amphioxus: piecing together ancestral genomes. *Evol Dev* **5**, 459-65 (2003).
27. Luke, G. N. et al. Dispersal of NK homeobox gene clusters in amphioxus and humans. *Proc Natl Acad Sci U S A* **100**, 5292-5 (2003).
28. Castro, L. F., Furlong, R. F. & Holland, P. W. An antecedent of the MHC-linked genomic region in amphioxus. *Immunogenetics* **55**, 782-4 (2004).
29. Panopoulou, G. & Poustka, A. J. Timing and mechanism of ancient vertebrate genome duplications -- the adventure of a hypothesis. *Trends Genet* **21**, 559-67 (2005).
30. Weber, J. L. & Myers, E. W. Human whole-genome shotgun sequencing. *Genome Res* **7**, 401-9 (1997).
31. Sodergren, E. et al. The genome of the sea urchin Strongylocentrotus purpuratus. *Science* **314**, 941-52 (2006).
32. Howell, W. M. & Boschung, H. T., Jr. Chromosomes of the lancelet, Branchiostoma floridae (order Amphioxi). *Experientia* **27**, 1495-6 (1971).
33. Small, K. S., Brudno, M., Hill, M. M. & Sidow, A. Extreme genomic variation in a natural population. *Proc Natl Acad Sci U S A* **104**, 5698-703 (2007).
34. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297-304 (2000).
35. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc Natl Acad Sci U S A* **99**, 803-8 (2002).
36. Britten, R. J., Rowen, L., Williams, J. & Cameron, R. A. Majority of divergence between closely related DNA samples is due to indels. *Proc Natl Acad Sci U S A* **100**, 4661-5 (2003).
37. Ivanova-Kazas, O. M. An essay on the phylogeny of lower chordates. *Proceedings of the St. Petersburg society of naturalists* **84** (1995).
38. Blair, J. E. & Hedges, S. B. Molecular phylogeny and divergence times of deuterostome animals. *Mol Biol Evol* **22**, 2275-84 (2005).
39. Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**, 965-8 (2006).
40. Bourlat, S. J. et al. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* **444**, 85-8 (2006).
41. Cameron, C. B., Garey, J. R. & Swalla, B. J. Evolution of the chordate body plan: new insights from phylogenetic analyses of deuterostome phyla. *Proc Natl Acad Sci U S A* **97**, 4469-74 (2000).
42. Putnam, N. H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86-94 (2007).
43. Edvardsen, R. B. et al. Hypervariable and highly divergent intron-exon organizations in the chordate

Oikopleura dioica. *J Mol Evol* **59**, 448-57 (2004).
44. Hughes, A. L. & Friedman, R. Loss of ancestral genes in the genomic evolution of Ciona intestinalis. *Evol Dev* **7**, 196-200 (2005).
45. Danchin, E. G. & Pontarotti, P. Towards the reconstruction of the bilaterian ancestral pre-MHC region. *Trends Genet* **20**, 587-91 (2004).
46. Nakatani, Y., Takeda, H., Kohara, Y. & Morishita, S. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**, 1254-65 (2007).
47. Venkatesh, B. et al. Survey Sequencing and Comparative Analysis of the Elephant Shark (Callorhinchus milii) Genome. *PLoS Biol* **5**, e101 (2007).
48. Robinson-Rechavi, M., Boussau, B. & Laudet, V. Phylogenetic dating and characterization of gene duplications in vertebrates: the cartilaginous fish reference. *Mol Biol Evol* **21**, 580-6 (2004).
49. WashU. (http://genome.wustl.edu/genome.cgi?GENOME=Petromyzon%20marinus, 2007).
50. Escriva, H., Manzon, L., Youson, J. & Laudet, V. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. *Mol Biol Evol* **19**, 1440-50 (2002).
51. Kohn, M. et al. Reconstruction of a 450-My-old ancestral vertebrate protokaryotype. *Trends Genet* **22**, 203-10 (2006).
52. Naruse, K. et al. A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Res* **14**, 820-8 (2004).
53. Woods, I. G. et al. The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res* **15**, 1307-14 (2005).
54. Jaillon, O. et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431**, 946-57 (2004).
55. Postlethwait, J. H. et al. Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res* **10**, 1890-902 (2000).
56. Panopoulou, G. et al. New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res* **13**, 1056-66 (2003).
57. Blomme, T. et al. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol* **7**, R43 (2006).
58. Christoffels, A. et al. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol* **21**, 1146-51 (2004).
59. Hoegg, S., Brinkmann, H., Taylor, J. S. & Meyer, A. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J Mol Evol* **59**, 190-203 (2004).
60. Brunet, F. G. et al. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* **23**, 1808-16 (2006).
61. Hellsten, U. et al. Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog Xenopus laevis. *BMC Biol* **5**, 31 (2007).
62. De Bodt, S., Maere, S. & Van de Peer, Y. Genome duplications and the origin of angiosperms. *Trends Ecol Evol* **20**, 591-597 (2005).
63. Woolfe, A. et al. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**, e7 (2005).
64. Pennacchio, L. A. et al. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499-502 (2006).
65. Prochnik, S. E., Rokhsar, D. S. & Aboobaker, A. A. Evidence for a microRNA expansion in the bilaterian ancestor. *Dev Genes Evol* **217**, 73-7 (2007).
66. Holland, L. Z. *Genome Res* **in press** (2008).
67. Donoghue, P. C. & Purnell, M. A. Genome duplication, extinction and vertebrate evolution. *Trends Ecol Evol* **20**, 312-9 (2005).
68. Valentine, J. W. Two genomic paths to the evolution of complexity in bodyplans. *Paleobiology* **26**, 513-519 (2000).
69. Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes.

*Science* **297**, 1301-10 (2002).
70. Tuskan, G. A. et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313**, 1596-604 (2006).
71. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).
72. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-52 (2000).

## Methods

**Genome Sequence and assembly**. 7.3 million high-quality sequence Sanger reads were generated and assembled using JAZZ[69,73]. See Supplementary Note 2 for details of the genomic libraries, assembly methods, and validation.

**Annotation of protein coding genes**. Protein coding genes were annotated by the JGI annotation pipeline as previously described [42,70]. See Supplementary Note 3 for a description of amphioxus-specific details. Gene models representing allelic pairs were identified using a combination of similarity of predicted peptide sequence and gene neighborhood context (Supplementary Note 3).

**Deuterostome relationships**. Sets of orthologous genes were collected by grouping together mutual-best BLAST[74] hits between Nematostella vectensis (sea anemone) and gene sets from other published deuterostome genomes plus expressed sequence data from the acorn worm Saccoglossus kowalevskii [75] and sea lamprey Petromyzon marinus (Supplementary Note 5);. Individual multiple alignments were created with CLUSTALW[71], manually reviewed, trimmed with Gblocks[72], and concatenated. 364 ortholog sets had representation from all genomes (alignment 1), and 1,090 had up to one missing (alignment 2). The alignments were analyzed by Bayesian and maximum likelihood methods using mrbayes[76,77] and PHYML[78]. Supplementary Note 5 contains more details of the data sources, data compilation and analysis.

**Intron evolution**. A collection of 5,337 orthologous well-aligned coding sequence positions which contain an intron in at least one genomes was collected and analyzed by weighted parsimony analysis with PAUP[79] and Bayesian analysis with mrbayes[76,77]. (Supplementary Note 6)

**Chordate gene families**. Using predicted proteomes for human, chicken, stickleback, puffer fish, sea squirt, amphioxus, sea urchin, fruit fly and sea anemone (Supplementary Note 5), families

("clusters") of orthologous genes were constructed to represent the ancestral gene complements of the tetrapod, teleost, jawed vertebrate, "olfactores" (i.e., vertebrates plus urochordate), chordate, and deuterostome ancestors, as previously described [42], with modifications described in Supplementary Note 7.

**Chromosome segmentation**. The human, chicken and stickleback chromosomes were segmented iteratively by comparison to one another and to the scaffolds of the fugu genome assembly. See Supplementaty Note 8 for complete details.

**Construction of Chordate Linkage Groups**. For whole-genome synteny analysis, orthology between genomes was based on c-score clustering as previously described [42], with c-score threshold of 0.75 when comparing human and amphioxus, and 0.95 when comparing human to other vertebrates. To define initial CLGs (Supplementary Figure 64), human chromosome segments and amphioxus scaffolds were clustered using the same method as for chromosome segmentation, with a correlation threshold of 0.25. Statistical significance of ortholog concentration between regions of one genome and another was computed with Fishers Exact with a Bonferroni correction for the total number of pairwise tests. (See Supplementary Note 8 for additional details.)

**Fluorescent in situ hybridization (FISH)**. Chromosome preparation was performed as previously described[80] with modifications described in Supplementary Note 8.

**Multi-species synteny comparison**. The clustering protocol described above for human-amphioxus was repeated for human-fugu, human-stickleback and human-*Nematostella* to define clusters of scaffolds or chromosome segments (a "cluster set") for each genome based on comparison to human. All pairs of human chromosome segments were compared to each cluster set, and for each set classified as having conserved synteny to the same cluster (coded "1"), having conserved synteny (only) to different clusters (coded with "0"), or having indeterminate conserved synteny if one or both human segments lack significant conserved synteny to any cluster in the cluster set (coded with "?"). The complete results are represented as a color-coded matrix in Supplementary Figure S9.

**Identification of Ohnologs.** Operationally, we define a pair of human genes as Ohnologs (paralogues descendent from 2R in vertebrate evolution) if they (a) are found in the same chordate gene family (excluding large gene families with >10 members) and (b) are ancient paralogs differing by more than 0.2 transversions/site at synonymous positions. (We use transversions

rather than the more common total substitutions because transversions occur more slowly and therefore show less saturation at the time scales of interest.)

**Decomposition of CLGs into independent products of duplication.** For each CLG, all partitionings of the human chromosomal segments assigned to the CLG were tabulated. Each partitioning was assigned a score as follows: +1 for each pair of segments from different partitions with a significant number of predicted ohnologs between them. -1 for each pair of segments from different partitions without a significant number of ohnologs between them. Among the partitionings with the maximum score, ties were broken by using the multi-species synteny comparison results: a score of +epsilon for each pair of segments from different partitions colored red or orange, and a score of -epsilon for each pair of segments from different partitions where multi-synteny comparison indicates the two segments were one segment in the jawed vertebrate ancestor (colored blue or purple in Supplementary Figures S9 and S10), where epsilon is a positive number much less than 1.

**Timing of genome duplications.** Genes from fugu, lamprey, and *Ciona intestinalis* ("X") were aligned to pairs of human Ohnologs ("Hs1" and "Hs2") and their orthologous amphioxus gene "Bf", and phylogenetic position considered resolved if it has at least 50% maximum likelihood bootstrap support. (Supplementary Note 9; Supplementary Figure S11).

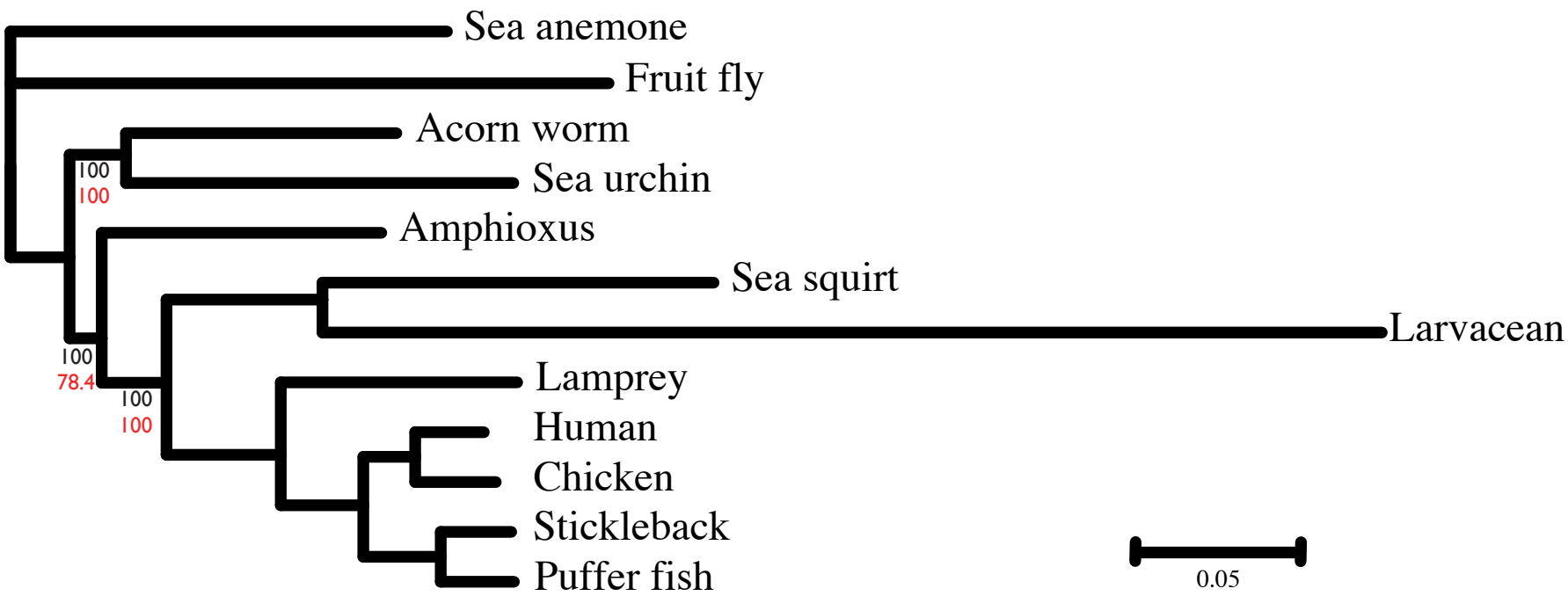
**Ancient Developmental Gene Linkages.** Genes were identified by tBLASTn against version 1.0 *Branchiostoma floridae* genome assembly with vertebrate and invertebrate homeodomain and Wnt sequences. Orthology was assigned from phylogenetic tree reconstruction using neighbor-joining and maximum likelihood approaches. Support for nodes was assessed by bootstrapping; all gene families were recovered with high support. Human data from Ensembl release 47 was used. Supplementary Table S3 lists the genes and gene models examined.

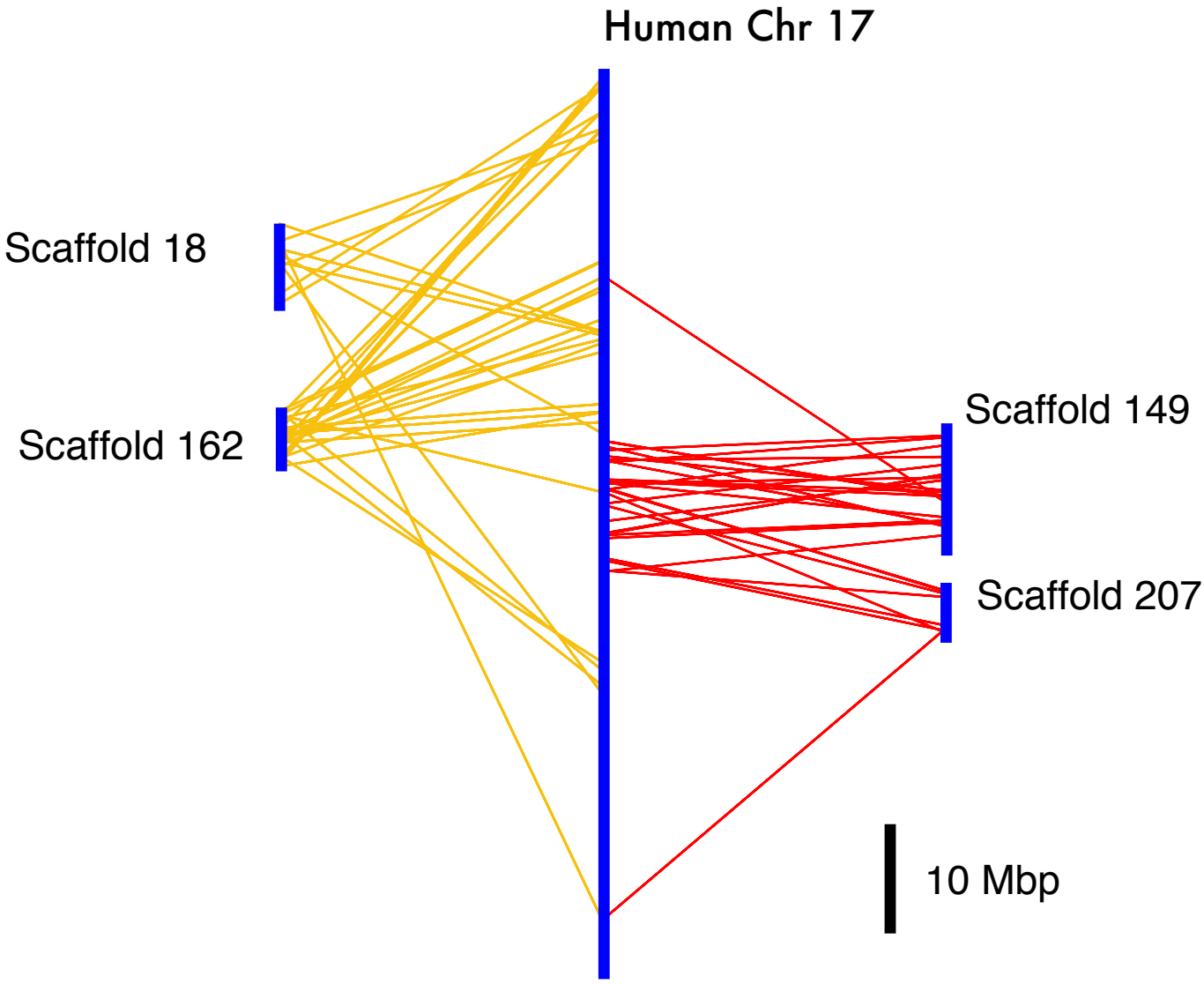
**Functional categorization of retained duplicate genes**. PANTHER functional annotations were mapped to inferred ancestral chordate genes, and subsets of these genes were analyzed for enrichment in functional categories by methods previously described for the analysis of ancestral eumetazoan genes.[42] Since functional annotations overlap, the category of "developmental processes" is itself dominated by genes associated with signal transduction and transcriptional regulation.

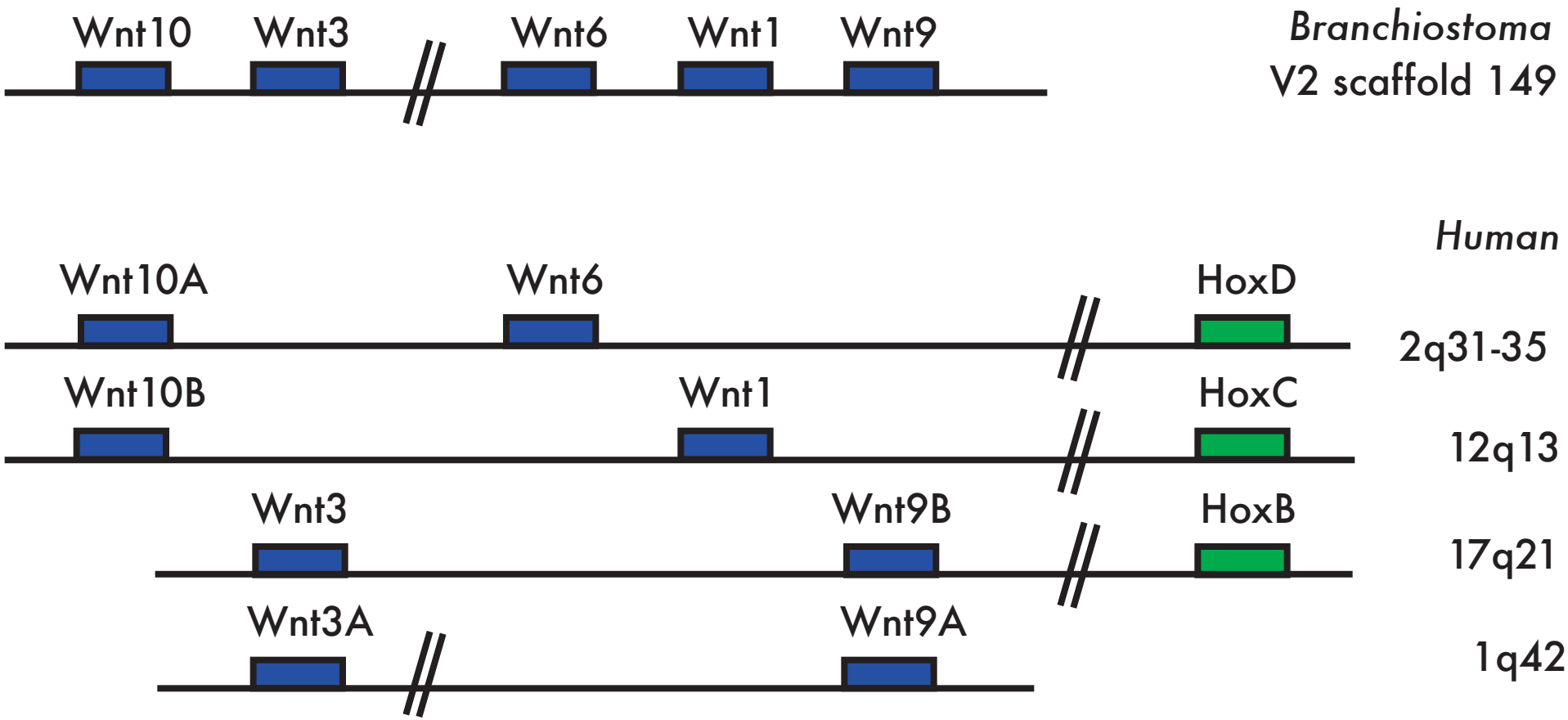
**Conserved non-coding element and Expression analysis**

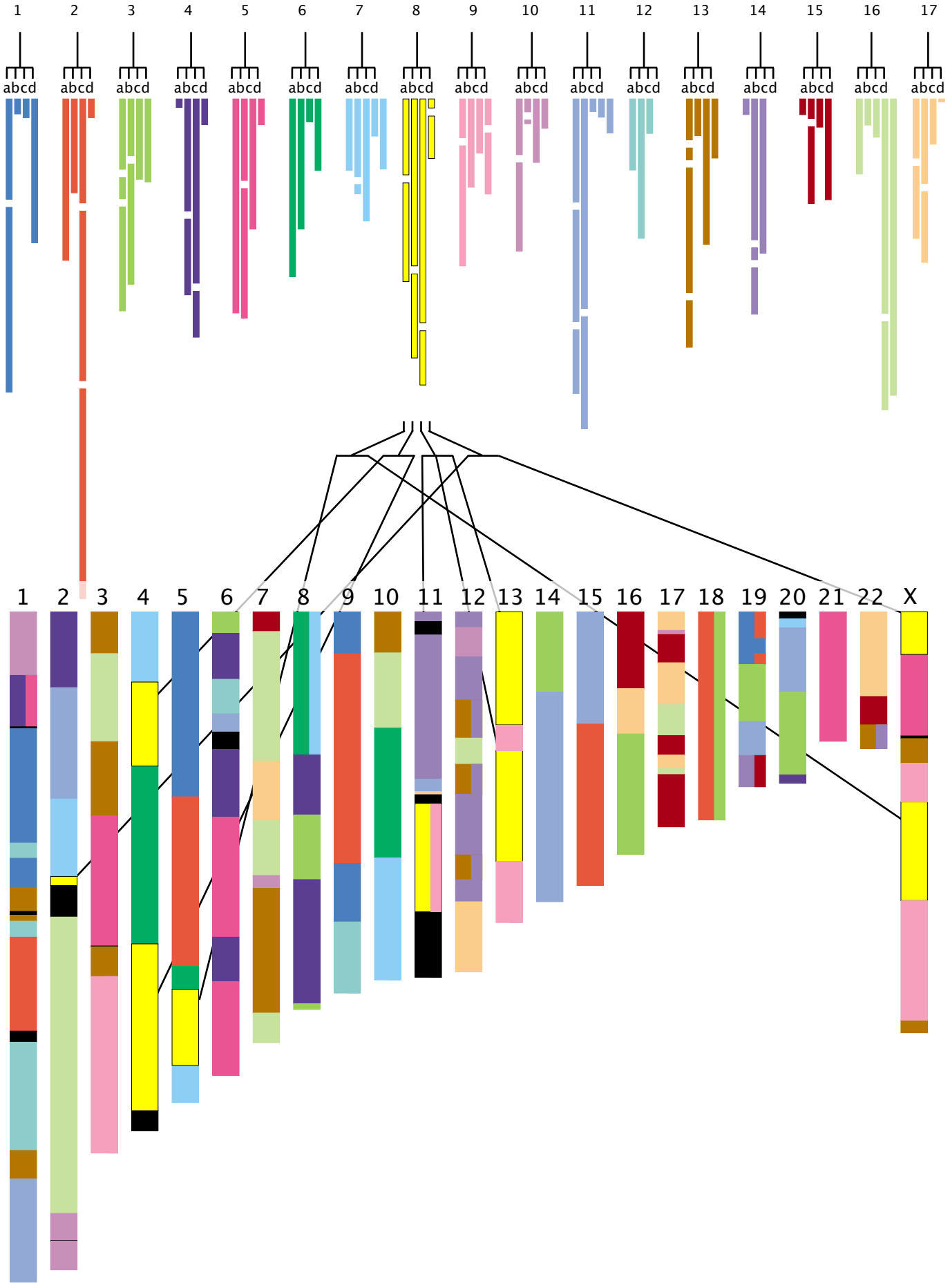
See Supplementary Note 10.

73. Dehal, P. et al. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science* **298**, 2157-67 (2002).
74. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402 (1997).
75. Lowe, C. J. et al. Anteroposterior patterning in hemichordates and the origins of the chordate nervous system. *Cell* **113**, 853-65 (2003).
76. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-5 (2001).
77. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-4 (2003).
78. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696-704 (2003).
79. Swofford, D. L. (Sinauer Associates, Sinderland, Massachusetts, 2003).
80. Castro, L. F. & Holland, P. W. Fluorescent in situ hybridisation to amphioxus chromosomes. *Zoolog Sci* **19**, 1349-53 (2002).


Sea anemone
Fruit fly
Acorn worm
100
100
Sea urchin
Amphioxus
Sea squirt
Larvacean
100
78.4
100
100
Lamprey
Human
Chicken
Stickleback
Puffer fish
0.05


Human Chr 17
Scaffold 18
Scaffold 162
Scaffold 149
Scaffold 207
10 Mbp


Wnt10
Wnt3
Wnt6
Wnt1
Wnt9
Branchiostoma
V2 scaffold 149
Human
Wnt10A
Wnt6
HoxD
2q31-35
Wnt10B
Wnt1
HoxC
12q13
Wnt3
Wnt9B
HoxB
17q21
Wnt3A
Wnt9A
1q42


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
abcd
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X