HOX clusters of *Latimeria*: Complete characterization provides further evidence for slow evolution of the coelacanth genome

Chris T. Amemiya [*] [†1], Thomas P. Powers [*], Sonja J. Prohaska [†], Jane Grimwood [‡2], Jeremy Schmutz [‡2], Mark Dickson [§], Tsutomu Miyake [*3], Michael A. Schoenborn [*], Richard M. Myers [‡2], Francis H. Ruddle [¶] and Peter F. Stadler [†] [‖1],

[*]Benaroya Research Institute at Virginia Mason, 1201 Ninth Avenue, Seattle, WA 98101 USA,[†]Department of Biology, University of Washington, 106 Kincaid Hall, Seattle, WA 98195 USA,[‡]The Stanford Human Genome Center and the Department of Genetics, Stanford University School of Medicine, Palo Alto, CA 94304, USA,[§]Cardiodx, Inc., Palo Alto, 2500 Faber Place, CA 94303, USA,[¶]Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06520 USA, and [‖]Max Planck Institute for Mathematics in the Science, Inselstraße 22, D-04103 Leipzig, Germany; Fraunhofer Institut für Zelltherapie und Immunologie – IZI Perlickstraße 1, D-04103 Leipzig, Germany, Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria; Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe NM 87501, USA

Preprint

The living coelacanth is a lobe-finned fish that represents an early evolutionary departure from the lineage that led to land vertebrates, and is of extreme interest scientifically. It has changed very little in appearance from fossilized coelacanths of the Cretaceous (150-65 million years ago), and is often referred to as a "living fossil." An important general question is whether long term stasis in morphological evolution is associated with stasis in genome evolution. To this end we have used targeted genome sequencing for acquiring 1,612,752 bp of high-quality finished sequence encompassing the four HOX clusters of the Indonesian coelacanth, *Latimeria menadoensis*. Detailed analyses were carried out on genomic structure, gene and repeat contents, conserved non-coding regions, and relative rates of sequence evolution in both coding and non-coding tracts. Our results demonstrate conclusively that the coelacanth HOX clusters are comparatively slowly evolving and that this taxon should serve as a viable outgroup for interpretating the genomes of tetrapod species.

HOX cluster | *Latimeria menadoensis* | evolution

Abbreviations: BAC, bacterial artifical chromosome; CNCN, conserved non-coding nucleotide; GFP, green fluorescent protein; IGR, intergenic region; PCR, polymerase chain reaction; WGD, whole genome duplication

The sign outside the Toliara Marine Museum in Madagascar shows a large coelacanth together with a depiction of the descent of man with the caption, "Tout le monde evolve sauf moi"[4]. Indeed, the living coelacanth, *Latimeria*, is considered an evolutionary relict that has generated a great deal of intrigue since its discovery in 1938, with interests in its anatomy, physiology, ecology, interrelationships and even politics [1]. Due to its protected status, the best practical approach to its study is from the "inside out", i.e., through comparative genomics. To this end we have constructed a high-representation bacterial artificial chromosome (BAC) library from the Indonesian coelacanth, *Latimeria menadoensis* [2], thus allowing indefinite preservation of its genome. Although genomics *per se* does not provide information as to morphology and function, the information gleaned from the comparative genomics approach can be applied and assayed in other model systems for inferring function [3]. It is using this approach that we are addressing evolutionary and developmental (evo-devo) questions concerning the coelacanth and taxa representative of early lineages of vertebrates.

Much of the interest in *Latimeria* has focused on its unusual morphology, which includes fleshy-lobed fins, a hollow nerve cord, poor ossification of skeleton yet presence of a rigid notochord that persists throughout its lifetime, lack of defined ribs, and a unique bi-lobate caudal region, the structure of which has been maintained in coelacanths since the middle Devonian [4]. While it is largely accepted that the coelacanth represents a *bona fide* outgroup to the tetrapods, the interrelationships of the lungfish, coelacanth and tetrapods (all sarcopterygian taxa) have been very difficult to resolve [5, 6]. In terms of comparative genomics, however, the coelacanth is the only tetrapod outgroup of practical importance, because the lungfishes possess genome sizes that are intractably large for routine genomic analyses [7].

HOX clusters were identified initially in *Drosophila* as gene complexes whose respective members could induce formation of homeotic transformations when mutated [8, 9]. Later, their homology to the vertebrate *Hox* genes was established [10, 11]. The molecular identification of these genes indicated that they all encoded a highly conserved 60 amino acid motif, the homeodomain, that we now know is involved in DNA binding. Mammals were shown to possess four HOX clusters, whose genes are intimately involved in axial patterning and, in vertebrates, a strict relationship exists between respective genes and their expression limits in somitic and neural tissues, the so-called "Hox code" [12]. Due to their intimate involvement in early development, the *Hox* genes have often been implicated as potentiators of evolutionary change and are frequently among the first genes examined in an evolutionary context.

Studies of vertebrate HOX cluster genomic organization have shown significant similarities as well as differences among the major taxa. The general conservation of *Hox* gene orthologs appears to be largely maintained, however, overt differences are seen in the number of absolute number of HOX clusters per taxon due to whole genome duplications (WGD) [13, 14]. The WGD events have also led to differences in the number and composition of respective *Hox* genes via differential gene losses. Collectively, the data indicate that the ancestral condition for the gnathostomes (jawed vertebrates) is four HOX clusters (A, B, C, D). These four clusters are thought to have been derived from an archetypal single HOX cluster via two WGDs prior to the emergence of the cartilaginous fishes [14, 15, 16, 17], Fig. 1. The euteleosts (inclusive bony fish clade) have undergone an independent whole genome duplication such that the ancestral euteleost possessed eight HOX clusters [15, 18, 19, 20] although most modern

Author contributions: C.T.A., T.P.P, R.M.M. and F.H.R. designed research; C.T.A., T.P.P., S.J.P., J.G., J.S., M.D., T.M., M.A.S. and P.F.S. performed research; C.T.A., T.P.P., S.J.P., and P.F.S. analyzed data; C.T.A., S.J.P., and P.F.S. wrote the paper.
The authors declare no conflict of interest.
Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. FJ497005-FJ497008)

[1] To whom correspondence may be addressed: E-mail: camemiya@benaroyaresearch.org or studla@bioinf.uni-leipzig.de

[2] Current address: HudsonAlpha Genome Sequencing Center, HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville AL 35806, USA.

[3] Current address: Department of Anatomy, Jikei University School of Medicine, Tokyo Minato-ku, Japan.

[4] Everybody evolves but me.

day representatives (e.g., zebrafish, medaka, pufferfishes, and cichlids) have less than eight due to cluster loss. The zebrafish genome contains 7 HOX clusters, with a remnant of the 8th (HOXDb) cluster having retained only a single microRNA [21]. A recent PCR survey of the mooneye (*Hiodon alosoides*, Osteoglossomorpha) provides evidence for the survival of all eight HOX clusters in the aftermath of the WGD [22]. Within the teleosts, some fishes such as the salmonids (salmons and trouts) have undergone yet an additional genome doubling event such that they possess twice as many HOX clusters as other teleosts [23]. In contrast, basal ray-finned fishes such as bichir, gar and bowfin do not appear to have undergone this extra WGD [24, 25, 26, 22]. The effects of the extra HOX clusters within teleosts are still unclear; some authors have implicated that they may have contributed to the success (speciation) of the teleost fishes [20, 27, 16] though this is an *ad hoc* hypothesis especially when one considers that this increase in cluster number has been accompanied by increases in gene losses [28].

Koh *et al.* [29] used a comprehensive PCR based approach in order to isolate *Hox* genes from the Indonesian coelacanth and to make inferences with regard to the number of HOX clusters and their genomic organizations. In this report we have greatly extended this analysis by completely isolating all of the HOX clusters of the Indonesian coelacanth in BAC clones, thereby allowing the generation of high quality sequences for the entire HOX complement. This enabled us to unequivocally identify all of the respective *Hox* genes. The goals of the project were to: (1) definitively identify all of the *Hox* genes in the four HOX clusters of the coelacanth, and determine their respective genomic organizations; (2) compare and contrast the HOX cluster organization of the coelacanth with that of other gnathostome species; (3) identify potential cis-regulatory elements using a comparative genomics approach; and (4) to measure relative rates of evolution of the coelacanth coding and noncoding sequences in comparison to that of other gnathostomes.

Results

Cluster Organization. We isolated BAC contigs encompassing the four *L. menadoensis* HOX clusters and determined their complete DNA sequence. The complete sequence of the four clusters revealed a high level of conservation. In total, there are 42 *Hox* genes ordered in the same transcriptional orientation throughout respective clusters, as well as two *Evx* paralogs associated with the HOXA and HOXD clusters. Based on our data and that of other taxa [30, 23, 31, 26, 22, 32, 33, 34] we constructed a more complete scenario of the evolutionary history of vertebrate HOX clusters, as shown in Fig. 1. The coelacanth has, in particular, retained *Hox* genes that are frequently lost in other lineages, such as *HoxC1* and *HoxC3*. Compared with cartilaginous fishes, *L. menadoensis* has lost only *HoxD2* and *HoxD13*. On the other hand, the *HoxA14* gene, which is pseudogenized in the horn shark and elephant shark is still intact in the coelacanth (Fig. 1).

Gene distances are largely conserved between coelacanth and human, as shown by the scale maps of the four clusters in Fig. 2 and in the graphic illustration in Fig.S1. Differences are visible mostly in the regions where *Hox* genes have been deleted (*HoxA14*). Interestingly, *HoxB10* has been removed from the human HOXB cluster without significant changes in the distance between *HoxB9* and *HoxB13*. The largest differences between human and coelacanth are an increase of the distances between *HoxD12* and *Exv2* that may be associated with the loss of *HoxD13* in the coelacanth, and an expansion of the intergenic region between *HoxD10* and *HoxD9*. Comparisons of HOX cluster structure among various vertebrate species are given in Fig.S2.

The *Latimeria menadoensis* HOX clusters harbour six microRNA genes, three of each of the two HOX associated families *mir-10* and *mir-196*. The genomic locations of the microRNAs in the *Hox10-Hox9* and the *Hox5-Hox4* intergenic regions, respectively, are the same as in other vertebrates [35]. The location of *mir-10* upstream of *Hox4* is also conserved in the cephalochordate *Branchiostoma floridae* [36] and in invertebrates including *Drosophila* [37].

Non-coding sequences. Global alignment-based identification of conserved non-coding sequences using mVISTA was carried out for the four coelacanth HOX clusters and clusters of various other vertebrates (see Supplement). This method has been shown to be effective at identifying and visualizing overtly conserved non-coding elements, including many that had been identified functionally such as the HoxC8 early enhancer [3] and for *Evx* [38], see Fig. S3. A much more inclusive and comprehensive means for identifying conserved non-coding nucleotides (CNCNs) utilizes the `tracker` program [39]. Fig. 3 summarizes the distribution of CNCNs as determined by the combination of `tracker` and `dialign` for the four *Latimeria* HOX clusters. A detailed list of the 875 individual phylogenetic footprints comprising 33,343 nt of CNCNs can be found at the Supplement website. The fraction of the intergenic regions (IGRs) between *Hox* genes contains nearly an order of magnitude more CNCNs than the surrounding genomic regions. This increase in non-coding sequence conservation was previously observed for the HOX clusters of many other vertebrates [40, 24, 39, 41, 42]. Due to the differences in the number and phylogenetic distribution of available HOX sequences for the 4 paralogons, differences in the sensitivity of the footprinting procedure are inevitable, so that the data are not comparable across different clusters. The data also reflect the expected increase in the density of CNCNs in the anterior part of the clusters [42, 36]

Repetitive Elements. As demonstrated for other vertebrate HOX clusters [43], repetitive elements are strongly excluded from the clusters. Repetitive DNA that appears more than once in the same HOX cluster sequence is located predominantly in the regions flanking the HOX cluster, while such repeats are rare in most of the intergenic regions between *Hox* genes (Fig.S4). The same pattern arises by measuring the fraction of interspersed repeats as illustrated in Fig. 4. The search for tRNAs resulted in several tRNA pseudogenes with unassigned anticodon. A `blastn` search against 24 fragments of genomic DNA with a length of more 100,000 nt showed that these sequences are relatively frequent in the *Latimeria* genome. Alignments with the complete set of human tRNAs showed that they fall into just two clusters with related sequences, identifying two related families of repeats. The consensus sequences of the two groups are provided in the Electronic Supplement. Consistent with the strong exclusion of repetitive elements from the HOX clusters, only a single copy was found inside a HOX cluster (between *HoxC3* and *HoxC1*).

Rates of Evolution. Relative rate tests of protein coding sequences demonstrate the reduced rate of evolution in the coelacanth relative to other vertebrate species. The differences are substantial so that Tajima tests on the well-conserved parts of individual protein coding sequences are already significant, Fig. 5a,b (see supplement for individual relative rate tests). Both human and zebrafish proteins evolve significantly faster than those of the coelacanth. The situation is reversed only for a single *Hox* gene, *HoxD10*, which is marginally faster in *Latimeria* than in human.

Rate differences in the evolution of non-coding sequences are harder to measure, since only local alignments are available. One possibility is to consider only sites that are conserved between *two* outgroups. Rate differences can be measured by differential rates in the loss of this ancestral state [44]. The corresponding statistical test be applied directly to the (concatenated) alignments of blocks of CNCNs described in the previous section. The requirement of two outgroups,

however, limits analysis to the A cluster, because appropriate data sets are only available for bichir and shark HOXA and not for other clusters. The duplicated, substantially derived HOX clusters of teleosts are not suitable for this kind of analysis due to the dramatic loss of CNCN in the wake of the teleost-specific genome duplication [39]. The data in Fig. 5c show that CNCNs evolve consistently slower in the HOX cluster than in any of the investigated tetrapod clusters. The fact that we observe larger absolute values of z' under the assumption that *Latimeria* CNCNs evolve at the same rate as the two outgroups implies a consistently accelerated rate in tetrapods relative to the other major gnathostome lineages.

Functionality of Hox14. In order to access whether coelacanth HoxA14 is potentially functional, we constructed a synthetic *HoxA14* cDNA and fused it with *GFP* in order to assess activity in a transient transfection assay. Representative data from one such transfection experiment are given in Fig. S5. These results clearly indicate that the *Latimeria* HoxA14 fusion protein is localized to the nucleus of transfected cells as would be expected for a typical Hox transcription factor.

Discussion

We have cloned and sequenced the HOX clusters of *Latimeria menadoensis*. We identified 42 *Hox* genes in four clusters (Fig. 2), including all 33 genes that were previously identified by Koh *et al.* [29]. Genes not identified in the previous report are *HoxA3*, *HoxA5*, *HoxA14*, *HoxB8*, *HoxB9*, *HoxB10*, *HoxC3*, *HoxC6*, and *HoxC11*. We also identified two *Evx* genes, *Evx1* and *Evx2* located upstream of HOXA and HOXD, respectively. Within each cluster, *Hox* genes were oriented in the same transcriptional orientation and the intergenic spacing was found to be highly similar to that of the human HOX clusters (Fig.S1, *cf.* Fig. 2 and Fig.S2). As in other vertebrates, the *Evx* genes are in opposite transcriptional orientation to the *Hox* genes proper. The HOXD cluster was sequenced far upstream and downstream of its *Hox* genes and contained known coding and noncoding sequences that have been found in other HOXD clusters, including the *Lunapark* gene and the HOXD global control region at its 5' end, and the *Metaxin2* gene at its 3' end [41]. Identification of the complete *Hox* gene complement in *Latimeria* permits a more accurate reconstruction of the evolutionary history of HOX clusters among the jawed vertebrates (Fig. 1). However, in terms of overall gross organization, the coelacanth HOX clusters are unremarkable relative to those sequenced from other species with four clusters (Fig.1S), which speaks to the general conservation of the HOX system. The euteleost fishes, in which an independent round of whole genome duplication has occurred, appear to be an exception to this trend [26, 45, 22].

The vertebrate HOX clusters have been shown to be largely devoid of repetitive DNA [43, 36]. This has been interpreted to mean that the clusters are co-adapted gene complexes that are not readily disrupted by recombination [8, 46]. Although a repeat library does not yet exist for *Latimeria*, our analysis suggests that HOX clusters show typical strong depletion of repetitive sequences within the clusters. As observed in previous studies [43, 31], repeat densities close to genomic background are observed in those long intergenic regions where the coherence of the clusters weakens. This is shown in Fig. 4 for the *HoxB13-HoxB10* IGR, which is also enriched in repeats in other vertebrates, and the two regions of HOXD that deviate most from its human counterpart, namely the posterior end, which suffered the loss of *HoxD13*, and the *HoxD10-HoxD9* IGR, which is three-fold expanded in the coelacanth due to repeat insertion.

We had previously shown that paralog group- (PG-) 14 genes were present in both coelacanth (*HoxA14*) and horn shark (*HoxD14* and *HoxA14* pseudogene) [47], suggesting that PG-14 was, in fact, an ancestral condition for jawed vertebrates. The potential functionality of coelacanth *HoxA14* was assessed via a simple *in vitro* assay (Fig. S5) in which Hox14 was fused to GFP. The data confirm that the coelacanth HoxA14 protein can direct proper expression in the nuclei of transiently transfected human fibroblasts, as expected for a functional transcription factor. These data confirm that HOXA14 is potentially functional. PG-14 genes have also been found in two other cartilaginous fishes, the cloudy catshark, *Scyliorhinus torazame*, (*HoxD14*) [48] and the elephant shark (*HoxD14*, as well as *HoxA14* and *HoxC14* pseudogenes) [33]. Moreover, it was shown that the Japanese lamprey, a jawless vertebrate, also possesses a *Hox14* gene [48], suggesting that PG-14 existed before the divergence of lampreys and gnathostomes. Expression analysis of the lamprey and catshark *Hox14* genes by *in situ* hybridization indicated that the genes did not show a predicted posterior axial pattern of *Hox* expression; rather, the genes showed a noncanonical expression pattern in the gut that overlapped with that of *Hox13*, implying that the PG-14 genes may have arisen as a gene duplicate of *Hox13*, complete with gut-specific regulatory sequences [48]. The timing of this duplication and the relationship of vertebrate PG14 to amphioxus *Hox14* (and *Hox15*) are difficult to assess due to lack of phylogenetic signal [47].

Vertebrate HOX clusters are well known to exhibit a high level of conservation in their non-protein-coding regions [40, 24, 39, 42, 36, 33, 32]. VISTA plots, Fig.S3, readily show that the coelacanth is no exception, and reveal conspicuously conserved regions, among them several footprints whose function has been studied in previous work [3, 38]. A more sensitive quantitative method [39] reveals that nearly 10% of the HOX cluster IGR sequences are conserved between *Latimeria* and tetrapods or cartilaginous fishes, a percentage that exceeds genomic background levels by an order of magnitude. In the light of the large evolutionary distance with its vertebrate relatives, this degree of phylogenetic footprint conservation is substantial, and is interpreted as a consequence of the tight and complex cross-regulatory network that characterizes vertebrate *Hox* genes.

The highly conserved structure of coelecanth HOX cluster is consistent with the observation that its evolutionary rate is slower than that of both human and zebrafish [49, 50]. Relative rate tests performed for protein sequences showed a systematic retardation in evolutionary rate in all four clusters relative to both human and zebrafish (Fig. 5a,b). For the HOXA cluster, where sequence data for two suitable outgroups (shark and bichir) were available, it was also possible to test evolutionary rates of conserved non-coding regions. The tests remain significant under the assumption that both outgroups and the alternative in-group evolve at the same constant rate (Fig. 5c), supporting the interpretation that the evolution of *Latimeria* HOX is indeed retarded relative to the in-groups assayed.

In this paper we report the procurement and analysis of the complete sequences of the four HOX clusters in the Indonesian coelacanth, *Latimeria menadoensis*. We show that its HOX clusters exhibit a high level of conservation and slow evolutionary rate, observations that are in keeping with findings from our previous study on the protocadherin gene clusters in the coelacanth [49]. In addition, the *Latimeria* genome has been shown to be evolving slowly with regard to the turnover of interspersed repeats (SINE-type retroposons) [51, 52, 53]. Whereas most retroposon families undergo expansion and rapid turnover during evolution, at least two SINE families that predate the coelacanth-tetrapod divergence show a differential retention pattern in coelacanth. These SINEs are propagated and maintained in the coelacanth genome as typical SINE-like families, but have undergone substantial turnover in the tetrapod genomes, even adopting new functions in both coding and non-coding regions (exaptation) [51, 52, 53]. *In toto*, these characteristics of the coelacanth genome are highly favorable for using it as a viable outgroup in order to better inform the genome biology and evolution of tetrapod species including humans. Moreover, the coelacanth genome will also help to decipher, from the inside-out, the unique biology of this fascinating creature.

Materials and Methods

Library Construction and Screening. High molecular weight genomic DNA was isolated from frozen heart tissue of the Indonesian coelacanth *Latimeria menadoensis* (the kind gift of Mark Erdmann). Two BAC genomic DNA libraries were constructed, the first, a pooled library, and the second, an arrayed library (described in [2]). For the former, genomic DNA was cloned into the pBACe3.6 cloning vector and transformed into *E. coli* DH10B cells. Transformants were then collected into 188 pools averaging 700 clones each. Genomic clones were obtained in a series of three steps. First, a genomic PCR survey of *Hox* sequences was performed via PCR amplification and sequencing of a portion of the homeobox using the universal *Hox* degenerate primer set `ELEKEF` and `WFQNRR` (primers 334 and 335, Suppl.Tab.T1), capable of amplifying the homeoboxes in *Hox* paralog groups PG1 through PG10. Second, the homeobox primers plus additional paralog group-specific primers were used in the isolation and identification of BAC clones from the BAC clone pools. Third, the arrayed library was screened using hybridization of PCR generated probe DNAs from the clone sets obtained in the PCR screens of the pooled library. Sequences of primers and probes are provided in the Electronic Supplement[5]. Average insert size in the arrayed library is 170Kb facilitating the isolation of complete HOX clusters. A minimal set of clones spanning the HOX clusters was then sent to the Stanford Human Genome Center (Palo Alto, CA) for complete DNA sequencing [49].

Sequencing. Sequencing of BAC ends and PCR products was performed by the Benaroya Research Institute Sequencing Facility using the ABI Prism DNA Sequencing Kit and the ABI 3100 Genetic Analyzer.

Annotation. DNA sequences were first analyzed using the Informax Vector NTI software package. *Hox* coding sequences were identified in part using the `GenomeScan` [54] web site[6] with known vertebrate *Hox* sequences as training set. Initial annotations were then refined using `ProSplign` (for coding sequences) and `Splign` (for UTRs) [55]. Putative start codons were evaluated based on the position specific weight matrix reported by [56]. A few intron positions (in the 5' part of *lnp* and in *HoxB10*) were corrected manually to use common splice donor motifs.

MicroRNA precursors were identified by a `blast` comparison with `MirBase` (version 10) [57], and with `GotohScan` [58] based on the HOX cluster associated microRNAs described in [35]. Furthermore, tRNAs and tRNA pseudogenes were detected with `tRNAscan-SE` [59]. tRNA pseudogenes for which the ancestral tRNA remained undetermined by `tRNAscan-SE` were aligned with the complete set of human nuclear tRNAs [60] with `clustalw` [61]. A Neighbor-Joining tree was used to determine their relationship to functional tRNAs.

The sequences of the four clusters and their annotation are deposited in `GenBank` with accession numbers **FJ497005**-**FJ497008**.

Repetitive Elements. Repetitive elements were annotated using `RepeatMasker`[7] in "vertebrate" mode. The density of interspersed repetitive elements was determined by counting the number of intergenic nucleotides that were annotated as interspersed elements (i.e., excluding simple and low complexity repeats). In order to visualize the repeat-content of the HOX cluster regions, we computed "dot-plots" comparing the nucleic acids sequence of a cluster against itself with `blastn`, as described in [36].

Analysis of Non-Coding Sequences. Long range sequence comparisons of HOX clusters from *Latimeria* and other vertebrates were performed using the `VistaPlot` web server [62], see Electronic Supplement. A systematic quantitative analysis of conserved non-coding sequence elements was performed in comparison with the following collection of species (HOX clusters): Hf – horn shark (*Heterodontus francisci*) **A**, **B**, **D**; Ps – bichir (*Polypterus senegalus*) **A**; Xt – frog (*Xenopus tropicalis*) **A**, **B**, **C**, **D**; Gg – chicken (*Gallus gallis*) **A**; Md – oppossum (*Monodelphis domestica*) **A**, **B**, **C**, **D**). Cf – dog (*Canis familiaris*) **A**, **B**, **C**, **D**; Hs – human (*Homo sapiens*) **A**, **B**, **C**, **D**; Mm – mouse (*Mus musculus*) **A**, **B**, **C**, **D**; Rn – rat (*Rattus norvegicus*) **A**, **B**, **C**, **D**. These sequences and their annotations can be found in the Electronic Supplement. For each of the four paralogous clusters we used `tracker` [39], a phylogenetic footprinting program based on `blast`, to determine an initial set of footprints. The complete lists of `tracker` footprints and the positions of the *Hox* genes were then used as weighted anchors for `dialign-2` [63]. This software produces global so-called segment-based alignments that emphasize local conservation. By construction, these alignments contained a maximal consistent set of `tracker` footprints together with additional local alignments detected by `dialign-2` only. As a consequence, this procedure increased the sensitivity relative to `tracker` alone. For these alignments, only short flanking regions outside the HOX cluster were used to reduce computational efforts.

The global `dialign-2` alignments were then further processed by a `perl` script (available from the Supplement website) that distinguishes conserved blocks from intervening variable regions in a multiple sequence alignment: Let p_α, $\alpha \in \{A, T, G, C\}$ be the frequency of nucleotide α in the entire alignment. For each alignment column, let f_α, $\alpha \in \{A, T, G, C, \_\}$ be the frequency of characters. In evaluating f_α we ignore all rows in which $\alpha =' \_'$ is part of a deletion longer than 9nt. We assign the score

$$S = \sum_{\alpha\in\{A,T,G,C\}} f_\alpha \log(f_\alpha/p_\alpha) + f_{\_} \log f_{\_} \qquad \textbf{[1]}$$

to each column. The first term measures the information content of the column, which is positive for well-conserved columns and approaches 0 when the column reflects the background nucleotide distribution. The second term is an entropy-like penalty for gaps, which is always non-positive. Alignment column k is considered as conserved if the running average of S over the interval $[k-L, k+L]$ reaches a threshold value S^*. Here we used the parameters $L = 4$, i.e., averages over windows of length 9 and a threshold value $S^* = 0.75$. A conserved block is defined as at least 6 consecutive conserved columns. Lists of all conserved blocks (excluding the sequence located between start and stop codon of the same protein) for the four HOX clusters can be found in the Electronic Supplement. These blocks were then used for statistical analysis.

Relative Rate Tests. **Protein Coding Sequences.** Tajima's relative rate test (RRT) [64] as implemented in the MEGA package [65] was applied to all exon-1 sequences of coelacanth, human, and zebrafish *Hox* proteins, using horn shark (HOXA, HOXB, HOXD) or elephant shark (HOXC) sequences as outgroup. Multiple RRTs can be combined to form a partial order encoding the relative evolutionary speeds of several species. Such data can be represented by the so-called Hasse diagram of the poset, in which faster-evolving genes are placed above the slower ones. A subset of significant tests are drawn as edges, so that all significant tests correspond to pairs of genes that are connected by a directed path [66]. **Noncoding Conserved Nucleotides.** Relative rates of evolution of conserved non-coding nucleotides (CNCNs) were evaluated following the procedure described in [44]. This test measures the differential loss of conservation in two ingroups of alignment positions that are conserved in two outgroups. Since two suitable outgroups, namely shark and bichir, were available for HOXA only, this analysis was confined to this cluster.

In extension of [44], we also implemented a bootstrapping procedure for this test to evaluate the stability of the data. As observed in [44] CNCNs typically contain short blocks of consecutive nucleotides that are conserved between the two outgroups. The average length of these blocks roughly matches the expected size of individiual footprints ($b \approx 6$) Conservatively, one assumes that these blocks evolve in a correlated fashion due to selective constraints. This is reflected in the testing procedure as an effective reduction of the variance. A bootstrapping approach has to incorporate this fact. The resampling of the alignment therefore proceeds by randomly picking $N/(2b)$ blocks of length $2b$ to obtain a new alignment of length N.

Cellular Localization of *HoxA14*. A synthetic *HoxA14* cDNA was generated using primers 791-796 (Supplemental Material) and overlap PCR. This cDNA was directionally cloned upstream and in-frame into the GFP gene of pEGFP-C3 [67]. Purified DNA was transfected into adherent GM0637 cells (human fibroblasts) using FuGene 6 cationic lipid transfection reagent (Roche) following the manufacturer's recommendations. Control transfections included a construct containing mouse *HoxA11* (positive control), as well as a mouse *HoxA11* construct that lacked the nuclear localization site [67] and empty vector (negative controls). Images were taken with a confocal microscope (Bio-Rad MRC-1024).

Acknowledgments

We thank Mandy Ranisch for help with checking the final annotation of the HOX cluster sequences, Karen Cerosaletti for help with transfection experiments and confocal microscopy, and Chi-hua Chiu for help with the first-generation coelacanth BAC library. This work was funded, in part, from grants from the National Science Foundation (IOS-0321461 and MCB-0719558 to CTA, and IOS-0321470 to FHR), the United States Department of Energy (DE-

[5] `http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-002/`

[6] `http://genes.mit.edu/genomescan.html`

[7] `http://www.repeatmasker.org/`

FG03-01ER63273 to CTA, DE-FG03-01ER63272 to RM, and DE-F602-01ER63274 to FHR), the Deutsche Forschungsgemeinschaft (Bioinformatics Initiative BIZ-6/1-2 to PFS), and the 6th Framework Programme of the European Union (projects 043312 "SYNLET" and 043251 "EDEN").

1. Balon EK, Bruton MN, Fricke H (1988) A fiftieth anniversary reflection on the living coelacanth, *Latimeria chalumnae*: some new interpretations of its natural history and conservation status. *Environ Biol Fishes* 23:241–280.
2. Danke J, et al. (2004) Genome resource for the indonesian coelacanth, *Latimeria menadoensis*. *J Exp Zool A: Comp Exp Biol* 301:228–234.
3. Shashikant C, Bolanowski SA, Danke J, Amemiya CT (2004) Hoxc8 early enhancer of the indonesian coelacanth, *Latimeria menadoensis*. *J Exp Zoolog B: Mol Dev Evol* 302:557–563.
4. Carroll RL (1988) *Vertebrate paleontology and evolution*. H. Freeman and Co., New York.
5. Takezaki N, Figueroa F, Zaleska-Rutczynska Z, Takahata N, Klein J (2004) The phylogenetic relationship of tetrapod, coelacanth, and lungfish revealed by the shoxa14.epsequences of forty-four nuclear genes. *Mol Biol Evol* 21:1512–1524.
6. Zardoya R, Cao Y, Hasegawa M, Meyer A (1998) Searching for the closest living relative(s) of tetrapods through evolutionary analyses of mitochondrial and nuclear data. *Mol Biol Evol* 15:506–517.
7. Rock J, Eldridge M, Champion A, Johnston P, Joss J (1996) Karyotype and nuclear DNA content of the australian lungfish, *Neoceratodus forsteri* (Ceratodidae: Dipnoi). *Cytogenet Cell Genet* 73:187–189.
8. Lewis EB (1978) A gene complex controlling segmentation in *Drosophila*. *Nature* 276:565–575.
9. Gehring WJ (1998) *Master Control genes in development and evolution: the Homeobox story (Terry Lecture Series)*. Yale University Press, New Haven, CT.
10. McGinnis W Krumlauf R (1992) Homeobox genes and axial patterning. *Cell* 68:283–302.
11. Schubert FR, Nieselt-Struwe K, Gruss P (1993) The antennapedia-type homeobox genes have evolved from three precursors separated early in metazoan evolution. *Proc Natl Acad Sci USA* 90:143–147.
12. Hunt P Krumlauf R (1991) Deciphering the Hox code: clues to patterning branchial regions of the head. *Cell* 66:1075–1078.
13. Holland PWH, Garcia-Fernández J, Williams NA, Sidow A (1994) Gene duplication and the origins of vertebrate development. *Development* (Suppl.):125–133.
14. Holland PW Garcia-Fernandez J (1996) Hox genes and chordate evolution. *Dev Biol* 173:382–395.
15. Amores A, et al. (1998) Zebrafish *Hox* clusters and vertebrate genome evolution. *Science* 282:1711–1714.
16. Taylor J, Braasch I, Frickey T, Meyer A, Van De Peer Y (2003) Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res* 13:382–390.
17. Prohaska SJ, et al. (2004) The shark HoxN cluster is homologous to the human HoxD cluster. *J Mol Evol* p. 58. 212-217.
18. Meyer A Málaga-Trillo E (1999) Vertebrate genomics: More fishy tales about Hox genes. *Curr Biol* 9:R210–R213.
19. Prince VE (2002) The Hox paradox: More complex(es) than imagined. *Developmental Biology* 249:1–15.
20. Amores A, et al. (2004) Developmental roles of pufferfish hox clusters and genome evolution in ray-fin fish. *Genome Res* 14:1–10.
21. Woltering JM Durston AJ (2006) The zebrafish *hoxDb* cluster has been reduced to a single microRNA. *Nat Genet* 38:601–602.
22. Chambers KE, et al. (2009) Hox cluster duplication in a basal teleost fish, the goldeye (*Hiodon alosoides*). *Th Biosci* 128:109–120.
23. Moghadam HK, Ferguson MM, Danzmann RG (2005) Evolution of *Hox* clusters in salmonidae: A comparative analysis between atlantic salmon (*Salmo salar*) and rainbow trout (*Oncorhynchus mykiss*). *J Mol Evol* 61:636–649.
24. Chiu CH, et al. (2004) Bichir *HoxA* cluster sequence reveals surprising trends in ray-finned fish genomic evolution. *Genome Res* 14:11–17.
25. Hoegg S, Brinkmann H, Taylor J, Meyer A (2004) Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J Mol Evol* 59:190–203.
26. Crow KD, Stadler PF, Lynch VJ, Amemiya C, Wagner GP (2006) The fish-specific hox cluster duplication is coincident with the origin of teleosts. *Mol Biol Evol* 23:121–136.
27. Taylor JS, Van de Peer Y, Meyer A (2001) Genome duplication, divergent resolution and speciation. *Trends Genet* 17:299–301.
28. Wagner GP, Amemiya C, Ruddle F (2003) Hox cluster duplications and the opportunity for evolutionary novelties. *Proc Natl Acad Sci USA* 100:14603–14606.
29. Koh EGL, et al. (2003) *Hox* gene clusters in the indonesian coelacanth, *Latimeria menadoensis*. *Proc Natl Acad Sci USA* 100:1084–1088.
30. Hoegg S Meyer A (2005) Hox clusters as models for vertebrate genome evolution. *Trends Genet* 21:421–424.
31. Prohaska SJ, Stadler PF, Wagner GP (2006) Evolutionary genomics of *Hox* gene clusters. In S Papageorgiou, ed., *HOX Gene Expression*, pp. 68–90. Landes Bioscience & Springer, New York.
32. Di-Poï N, Montoya-Burgos JI, Duboule D (2009) Atypical relaxation of structural constraints in Hox gene clusters of the green anole lizard. *Genome Res* 19:602–610.
33. Ravi V, et al. (2009) Elephant shark (*Callorhinchus milii*) provides insights into the evolution of Hox gene clusters in gnathostomes. *Proc Natl Acad Sci USA* 106:16327–16332.
34. Raincrow JD, et al. (2009) *Hox* clusters of the bichir (*Polypterus senegalus*). Tech. Rep. BIOINF 09-040, U. Leipzig.
35. Tanzer A, Amemiya CT, Kim CB, Stadler PF (2005) Evolution of microRNAs located within *Hox* gene clusters. *J Exp Zool: Mol Dev Evol* 304B:75–85.
36. Amemiya CT, et al. (2008) The amphioxus *Hox* cluster: characterization, comparative genomics, and evolution. *J Exp Zool B: Mol Dev Evol* 310B:465–477.
37. Stark A, et al. (2007) Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Res* 17:1865–1879.
38. Suster ML, et al. (2009) A novel conserved *evx1* enhancer links spinal interneuron morphology and cis-regulation from fish to mammals. *Dev Biol* 325:422–433.
39. Prohaska S, Fried C, Flamm C, Wagner G, Stadler PF (2004) Surveying phylogenetic footprints in large gene clusters: Applications to Hox cluster duplications. *Mol Phyl Evol* 31:581–604.
40. Chiu Ch, et al. (2002) Molecular evolution of the HoxA cluster in the three major gnathostome lineages. *Proc Natl Acad Sci USA* 99:5492–5497.
41. Lee AP, Koh EGL, Tay A, Sydney B, Venkatesh B (2006) Highly conserved syntenic blocks at the vertebrate Hox loci and conserved regulatory elements within and outside Hox gene clusters. *Proc Natl Acad Sci USA* 103:6994–6999.
42. Hoegg S, Boore JL, Kuehl JV, Meyer A (2007) Comparative phylogenomic analyses of teleost fish *Hox* gene clusters: lessons from the cichlid fish *Astatotilapia burtoni*. *BMC Genomics* 8:317.
43. Fried C, Prohaska SJ, Stadler PF (2004) Exclusion of repetitive dna elements from gnathostome Hox clusters. *J Exp Zool, Mol Dev Evol* 302B:165–173.
44. Wagner GP, Fried C, Prohaska SJ, Stadler PF (2004) Divergence of conserved noncoding sequences: Rate estimates and relative rate tests. *Mol Biol Evol* 21:2116–2121.
45. Kuraku S Meyer A (2009) The evolution and maintenance of *Hox* gene clusters in vertebrates and the teleost-specific genome duplication. *Int J Dev Biol* 53:765–773.
46. Duboule D (2007) The rise and fall of Hox gene clusters. *Development* 134:2549–2560.
47. Powers TP Amemiya CT (2004) Evidence for a *Hox14*, paralog group in vertebrates. *Curr Biol* 14:R183–R184.
48. Kuraku S, et al. (2008) Noncanonical role of *Hox14* revealed by its expression patterns in lamprey and shark. *Proc Natl Acad Sci USA* 105:6679–6683.
49. Noonan JP, et al. (2004) Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. *Genome Res* 14:2397–2405.
50. Brinkmann H, Venkatesh B, Brenner S, Meyer A (2004) Nuclear protein-coding genes support lungfish and not the coelacanth as the closest living relatives of land vertebrates. *Proc Natl Acad Sci USA* 101:4900–4905.
51. Bejerano G, et al. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441:87–90.
52. Nishihara H Smit N A Fand Okada (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res* 16:864–874.
53. Xie X, Kamal M, Lander ES (2006) A family of conserved noncoding elements derived from an ancient transposable element. *Proc Natl Acad Sci USA* 103:11659–11664.
54. Yeh RF, Lim LP, Burge CB (2001) Computational inference of homologous gene structures in the human genome. *Genome Res* 11:803–816.
55. Kapustin Y, Souvorov A, Tatusova T, Lipman D (2008) `Splign`: algorithms for computing spliced alignments with identification of paralogs. *Biology Direct* 3:20.
56. Peri S Pandey A (2001) A reassessment of the translation initiation codon in vertebrates. *Trends Genet* 17:685–687.
57. tools for microRNA genomics m (2008) Griffiths-jones, s and saini, h k and van dongen, s and enright, a j. *Nucleic Acids Res* 36:D154–D158.
58. Hertel J, et al. (2009) Non-coding RNA annotation of the genome of *Trichoplax adhaerens*. *Nucleic Acids Res* 37:1602–1615.
59. Lowe TM Eddy S (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl Acids Res* 25:955–964.
60. Jühling F, et al. (2009) `tRNAdb` 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res* 37:D159–D162.
61. Thompson JD, Higgs DG, Gibson TJ (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucl Acids Res* 22:4673–4680.
62. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 32:W273–279.
63. Morgenstern B, et al. (2004) Multiple sequence alignment with user-defined constraints gobics. *Bioinformatics* 7:1271–1273.
64. Tajima F (1993) Simple methods for testing molecular clock hypothesis. *Genetics* 135:599–607.
65. Kumar S, Dudley J, Nei M, Tamura K (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings Bioinformatics* 9:299–306.
66. Prohaska SJ, Fritzsch G, Stadler PF (2008) Rate variations, phylogenetics, and partial orders. In M Ahdesmäki, K Strimmer, N Radde, J Rahnenführer, K Klemm, H Lähdesmäki, O Yli-Harja, eds., *Fifth International Workshop on Computational Systems Biology, WCSB 2008*, pp. 133–136. TU Tampere, Tampere, FI.
67. Roth JJ, Breitenbach M, Wagner GP (2005) Repressor domain and nuclear localization signal of the murine *Hoxa-11* protein are located in the homeodomain: no evidence for role of poly-alanine stretches in transcriptional repression. *J Exp Zoolog B: Mol Dev Evol* 304:468–475.

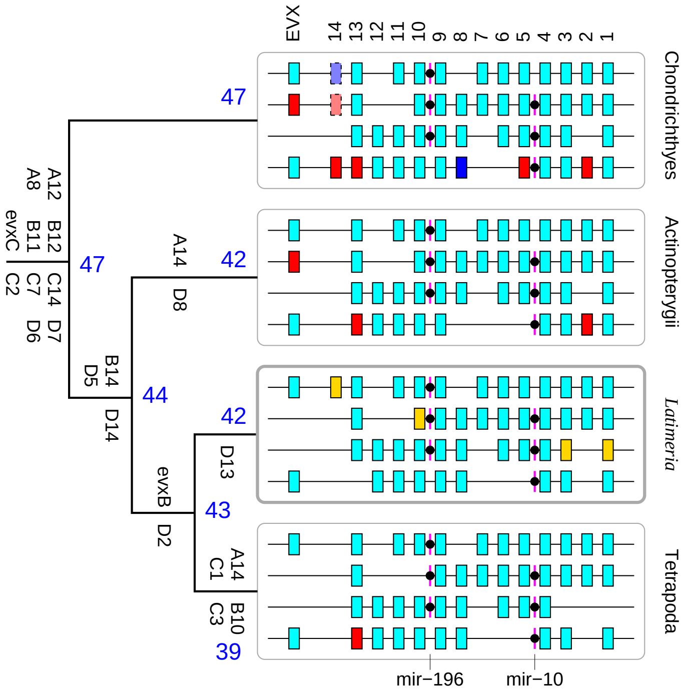

Fig. 1. Evolution of the HOX clusters in chordates. For each taxon, HOX clusters are illustrated from top to bottom, HOXA, HOXB, HOXC and HOXD. Genes shown in cyan inferred to constitute the ancestral states of the major chordate lineages. Dark blue boxes are losses in the actinopterygian stem linages; red boxes are genes that are absent from *Latimeria*, yellow boxes indicate *Latimeria* genes that are lost in the tetrapod stem-lineage. The number of retained *Hox* genes is indicated by blue numbers; the gene designations among the branches are those *Hox* genes which are inferred to have been lost. Ancestral gene complements are a composite of [22, 23, 30, 31, 32, 33, 34, 45]. Gene counts include *Hox* pseudogenes but exclude *Exv* paralogs. Most data from actinopterygian fishes come from teleosts, which have undergone an additional round of genome duplication. A gene is counted as present if it survived in at least one of the two teleostean copies. Duplicated paralogs are not added to the total.

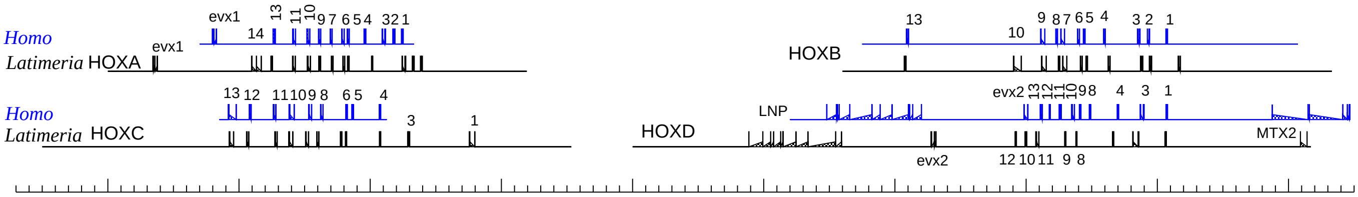

Fig. 2. Scale map of the *Latimeria menadoensis* HOX clusters compared to their human counterparts. Major tic marks are 100kb. Comparison of relative HOX cluster sizes and intergenic spacing among various vertebrates is given in Fig.S2.

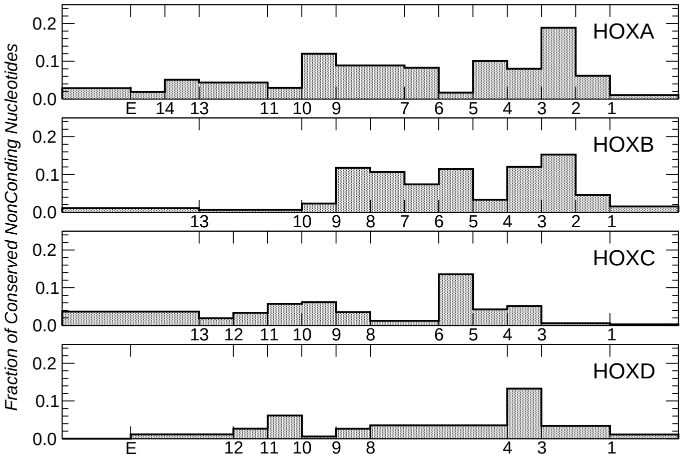

Fig. 3. Distribution of conserved non-coding DNA in intergenic regions between *Hox* genes. The figure summarizes the compilation of the conserved phylogenetic footprints as determined the `tracker` algorithm. A listing of all conserved footprints is given in the online supplement. For each intergenic region as well as the regions flanking the four *Latimeria* HOX clusters, the fraction of nucleotides contained in conserved noncoding elements is plotted. The highest totals are seen between *HoxA2* and *HoxA3*, *HoxB2* and *HoxB3*, *HoxC5* and *HoxC6*, and *HoxD3* and *HoxD4*. Functional aspects of these conserved footprints are largely unknown, though many are likely to represent *cis*-regulatory elements.

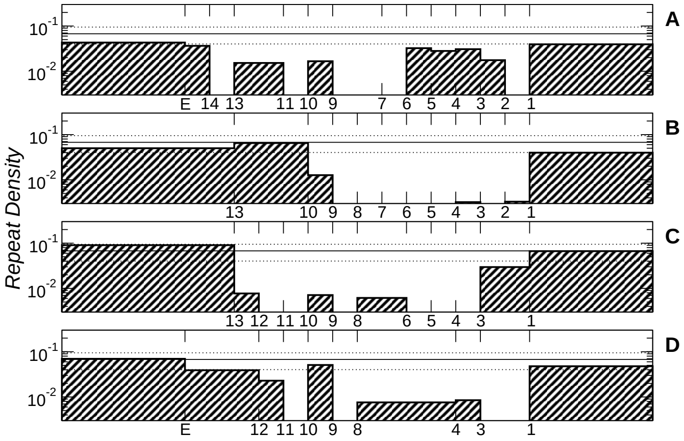

Fig. 4. Density of repetitive elements measured as the fraction of nucleotides annotated as interspersed repeats by `repeatmasker`. Numbers refer to *Hox* genens, E=*Evx*. The fraction of nucleotides in repetitive elements is shown on a log-scale for each IGR and the regions adjacent to the HOX clusters. The three horizontal lines indicate the distribution of the repeat density of the *Latimeria* genome determined from the 15 longest GenBank entries from *Latimeria menadoensis*. The middle line is the average density. In addition plus/minus one standard deviation is indicated. Repetitive elements are depleted only within the HOX clusters, while in the flanking regions the repeat density is consistent with the genomic distribution.

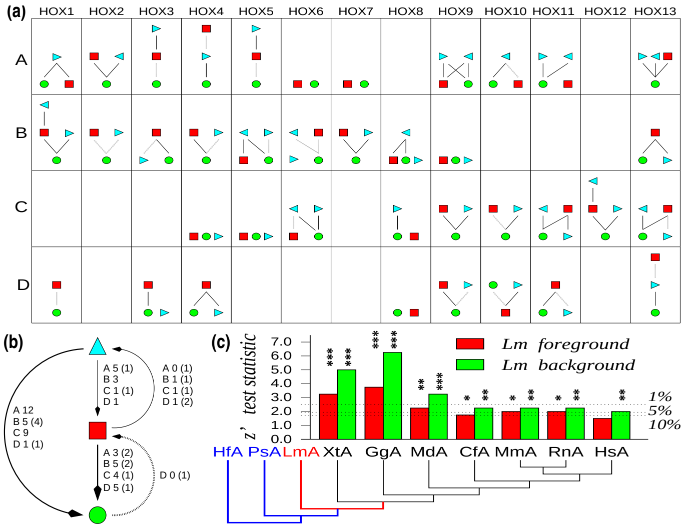

Fig. 5. Relative Rate Tests. **(a)** Summary of Tajima tests performed on Hox protein sequences using horn shark (HOXA, HOXB, HOXD) or elephant shark (HOXC) as outgroup. For each gene, a Hasse diagram shows highly significant ($p \leq 0.01$, full line) and significant ($0.01 < p \leq 0.05$, dotted line) comparisons, with the faster-evolving gene shown above the slower-evolving one. Lm ●, Hs ■, Dr-a ▶, Dr-b ◀. **(b)** Summary of significant relative rate tests at species level. Each arrow indicates that RRTs were significant for one or more genes between two species, with the arrow pointing towards the slower-evolving species. Full arrows imply that there are highly significant test results, dotted arrows refer tests that are only significant. The number of highly significant (significant) tests is indicated for each of the four HOX clusters. Except for the HOXD cluster, mostly zebrafish (▲) genes evolve faster than human (■) genes. For HOXD this situation is reverse. With a single marginally significant exception (*HoxD10*), *Latimeria* (●) never appears as the faster-evolving species. **(c)** Relative rate tests for conserved non-coding regions. Two outgroups are necessary to determine the conserved nucleotide positions. The test contrasts the evolutionary rate of one of two in-groups (foreground) against a constant rate among the two outgroups and the other in-group (background). *Latimeria* always appears slow evolving: as “foreground” it appears significantly retarded. When used as background in-group, each tetrapod in-group is significantly accelerated. Significance levels are * $p < 0.1$, ** $p < 0.05$, and *** for $p < 0.01$. Abbreviations: Dr – *Danio rerio* (zebrafish), Hf – *Heterodontus francisci* (horn shark), Ps – *Polypterus senegalus* (bichir), Lm – *Latimeria menadoensis* (coelacanth), Xt – *Xenopus tropicalis* (clawed frog), Gg – *Gallus gallus* (chicken), Md – *Monodelphis domestica* (opossum), Cf – *Canis familiaris* (dog), Mm – *Mus musculus* (mouse), Rn – *Rattus norvegicus* (rat), Hs – *Homo sapiens* (humans).