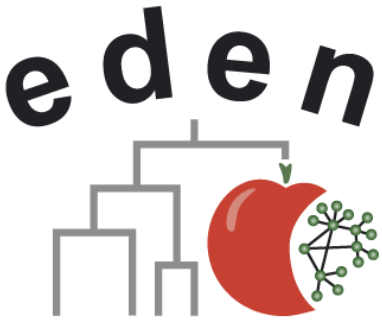


Project no. 043251

EDEN

ECOLOGICAL DIVERSITY AND EVOLUTIONARY NETWORKS

Instrument: Specific Targeted Project (STREP)

Thematic Priority: Integrating and strengthening the European Research Area
NEST Pathfinder initiative Tackling Complexity in Science

Deliverable 5.2.: Report on network phylogenies and their properties

Due date of deliverable: 31 December 2009
Actual submission date: 8 February 2009

Start date of project: 1 January 2007 Duration: 36 months

Organisation name of lead contractor for this deliverable: University of Leipzig
Other contributors:

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Introduction

This report comprises publications related to the reconstruction of phylogenetic networks and their analyses, produced within the EDEN project.

Rate Variations, Phylogenetics, and Partial Order, Prohaska et al, *Fifth International Workshop on Computational Systems Biology, WCSB 2008*, 133-136.

The systematic assessment of rate variations across large datasets requires a systematic approach for summarizing results from individual tests. Often, this is performed by coarse-graining the phylogeny to consider rate variations at the level of sub-claded. In a phylo-geographic setting, however, one is often more interested in other partitions of the data, and in an exploratory mode a pre-specified subdivision of the data is often undesirable. We propose here to arrange rate variation data as the partially ordered set defined by the significant test results.

***Hox* cluster duplication in the basal teleost *Hiodon alosoides* (Osteoglossomorpha),** Chambers et al, *Theory in Biosciences,* (2009) 128:109–120.

Large-scale—even genome-wide—duplicationsh ave repeatedly been invoked as an explanation for major radiations. Teleosts, the most species-rich vertebrate clade, underwent a ‘‘fish-specific genome duplication’’ (FSGD) that is shared by most ray-finned fish lineages. We investigate here the *Hox* complement of the goldeye (*Hiodon alosoides*), a representative of Osteoglossomorpha, the most basal teleostean clade. An extensive PCR survey reveals that goldeye has at least eight *Hox* clusters, indicating a duplicated genome compared to basal actinopterygians. The possession of duplicated *Hox* clusters is uncoupled to species richness. The *Hox* system of the goldeye is substantially different from that of other teleost lineages, having retained several duplicates of *Hox* genes for which crown teleosts have lost at least one copy. A detailed analysis of the PCR fragments as well as full length sequences of two *HoxA13* paralogs, and *HoxA10* and *HoxC4* genes places the duplication event close in time to the divergence of Osteoglossomorpha and crown teleosts. The data are consistent with—but do not conclusively prove—that Osteoglossomorpha shares the FSGD.

Evolution of Spliceosomal snRNA Genes in Metazoan Animals, Marz et al, *J Mol Evol* (2008) 67:594–607,

While studies of the evolutionary histories of protein families are commonplace, little is known on noncoding RNAs beyond microRNAs and some snoRNAs. Here we investigate in detail the evolutionary history of the nine spliceosomal snRNA families (U1, U2, U4, U5, U6, U11, U12, U4atac, and U6atac) across the completely or partially sequenced genomes of metazoan animals. Representatives of the five major spliceosomal snRNAs were found in all genomes. None of the minor splicesomal snRNAs were detected in nematodes or in the shotgun traces of Oikopleura dioica, while in all other animal genomes at most one of them is missing. Although snRNAs are present in multiple copies in most genomes, distinguishable paralogue groups are not stable over long evolutionary times, although they appear independently in several clades. In general, animal snRNA secondary structures are highly conserved, albeit, in particular, U11 and U12 in insects exhibit dramatic variations. An analysis of genomic context of snRNAs reveals that they behave like mobile elements, exhibiting very little syntenic conservation.

Complete HOX cluster characterization of the coelacanth provides further evidence for slow evolution of its genome, Amemiya et al, *Proc Natl Acad Sci U S A.* 2010 Feb 5.

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 to 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 noncoding regions, and relative rates of sequence evolution in both coding and noncoding tracts. Our results demonstrate conclusively that the coelacanth HOX clusters are evolving comparatively slowly and that this taxon should serve as a viable outgroup for interpretation of the genomes of tetrapod species.

RATE VARIATIONS, PHYLOGENETICS, AND PARTIAL ORDERS

Sonja J. Prohaska[1,2], *Guido Fritzsch*[3,4], *and Peter F. Stadler*[5,1,2,4,6]

[1]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe NM 87501, USA
[2]Department of Theoretical Chemistry, University of Vienna,
Währingerstraße 17, A-1090 Wien, Austria;
[3]Institute of Biology II: Zoologie, Molekulare Evolution und Systematik der Tiere,
University of Leipzig, Talstrasse 33, D-04103 Leipzig, Germany
[4]Interdisciplinary Center for Bioinformatics, and
[5]Bioinformatics Group, Department of Computer Science
University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
[6]RNomics Group, Fraunhofer Institut for Cell Therapy and Immunology (IZI),
Deutscher Platz 5e, D-04103 Leipzig, Germany
sonja@santafe.edu, gfritz@rz.uni-leipzig.de, studla@bioinf.uni-leipzig.de

ABSTRACT

The systematic assessment of rate variations across large datasets requires a systematic approach for summarizing results from individual tests. Often, this is performed by coarse-graining the phylogeny to consider rate variations at the level of sub-claded. In a phylo-geographic setting, however, one is often more interested in other partitions of the data, and in an exploratory mode a pre-specified subdivision of the data is often undesirable. We propose here to arrange rate variation data as the partially ordered set defined by the significant test results.

1. INTRODUCTION

Rate variations are an important source of information in evolutionary biology. Typically, one devises so-called relative-rate tests (RRTs) for statistically significant rate variations between two species [1, 2, 3, 4] or between subgroups of species [5, 6]. Group tests, however, require an initial hypothesis about which species to summarize. In particularly in an exploratory phase this is typically undesirable, since rate variations can be associated with many very different mechanisms, for clade-specific changes in mutation rates to differences in population structure.

In this contribution we therefore introduce an explorative approach to summarizing the results of many pairwise RRTs. The basic idea is to arrange the individual statistically significant pair-wise test results in a partially ordered set. Inspection of the Hasse diagram of this graph can then be used to identify systematic rate variations. In particular, this approach has the potential to highlight systematic rate variations even if they do not conform to a phylogenetic tree but correlate with other variables, such as migratory history.

2. RELATIVE RATE PO-SET

2.1. Po-Sets

Recall that a partially ordered set, *po-set* for short, is a set X together with a relation $\preceq$ satisfying
(P0) $x \preceq x$.
(P1) $x \preceq y$ and $y \preceq x$ implies $x = y$.
(P2) $x \preceq y$ and $y \preceq z$ implies $x \preceq z$.
A finite po-set $(X, \preceq)$ can be respresented as directed acyclic graph G (by drawing an arc $x \leftarrow y$ whenever $x \preceq y$ and $x \neq y$). The Hasse diagram of G is the subgraph H of G with the same vertex set X, and an arc $x \to y$ if $x \to y$ is an arc in G and there is no $z \neq x, y$ such that z lies on a directed path from x to y in G.

2.2. Substitution Rates

Let $\mathcal{X}$ be a set a taxa, which we represent here by their (aligned) nucleic acid or peptide sequences of length n. Furthermore, let $\mathfrak{T}$ be the underlying phylogenetic tree. Each interior vertex w of the tree can be specified as the *last common ancestor* $w = \mathrm{lca}(A, B)$ of two of the descents A and B of w so that the path connecting A and B runs through w.

The Hamming distance $d_{AB} = |\{i|A_i \neq B_i\}|$ counts the positions i in which the characters of the sequences differ. Now consider a triple (A, B, C) of sequences. The quantities

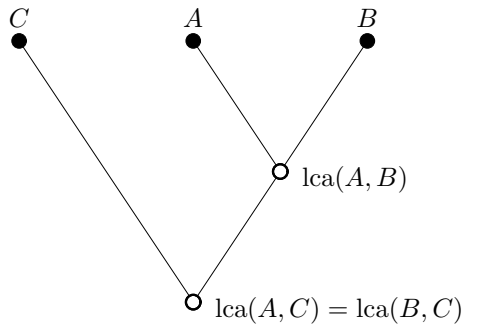
$$\begin{aligned} a_{ABC} &= |\{i|A_i = B_i = C_i\}|, \\ m_{AB|C} &= |\{i|A_i = B_i \neq C_i\}|, \\ m_{AC|B} &= |\{i|A_i = C_i \neq B_i\}|, \\ m_{BC|A} &= |\{i|B_i = C_i \neq A_i\}|, \\ w_{ABC} &= |\{i|A_i \neq B_i \neq C_i \neq A_i\}| \end{aligned} \tag{1}$$

distiguish five classes of alignment positions: (i) constant positions, (ii) positions in which all three sequence differ and (iii) three classes of positions in which two sequences are the same and the third one ins different.

The Hamming distance d_{AB} can be decomposed into three different components w.r.t. to a third sequence C. These correspond to the sequence position where C agrees with B (but not with A), the positions where C agrees with A (but not with B), and those where all three sequences differ:

$$d_{AB} = m_{BC|A} + m_{AC|B} + w_{ABC} \tag{2}$$

Now consider a subtree of $\mathfrak{T}$ consisting of three taxa A, B, C so that C is an outgroup to A and B:



(3)

Let us denote by a and b the lengths of branches between A, B and $\mathrm{lca}(A, B)$, respectively. We have

$$\begin{aligned} 2a = d_{AC} + d_{AB} - d_{BC} &= 2m_{BC|A} + w_{ABC} \\ 2b = d_{BC} + d_{AB} - d_{AC} &= 2m_{AC|B} + w_{ABC} \end{aligned} \tag{4}$$

and hence

$$a - b = m_{BC|A} - m_{AC|B}\,. \tag{5}$$

Note that $m_{BC|A}$ and $m_{AC|B}$ count independent sequence positions, while the Hamming distances are dependent via the common term w_{ABC}. Equ.(5) is the basis of Tajima's relative rate test [2], while the older Wu & Li test [3] uses the difference $d_{AC} - d_{BC}$. Alternatively, one might want to employ a suitable maximum likelihood test to assess the significance of branch length differences [1, 4].

We can estimate the relative rate of evolution along the branches a and b for those comparisons that are statistically signficant according to the relative rate test of choice. In the following, it will be more convenient to use the following logarithmic measure

$$\eta_{AB} = \begin{cases} \ln \frac{a}{b} & \text{if } a - b \text{ is statistically significant} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Next we show that for ideal data we do not have to fear contradictory results of relative rate tests involving different triples of taxa selected from the tree $\mathfrak{T}$. Recall that the distances d_{AB} of leafs A and B in a additive metric tree $\mathfrak{T}$ are defined as the sum of the lengths of the edges along the unique path that connects A and B in $\mathfrak{T}$.

More precise, we have the following

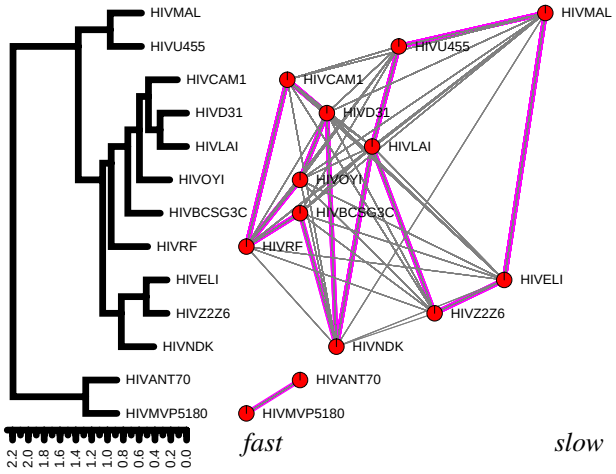
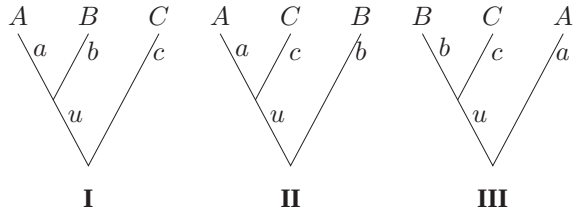


Figure 1. Example of a relative rate poset. Data are 5'UTRs of HIV-1. Thin lines in the r.h.s. panel indicate significant Tajima tests, the thick lines represent the associated Hasse diagram of the partially ordered set.

Theorem 1. *The directed graph associated with $\boldsymbol{\eta}$ is acyclic provided d is an additive tree metric on $\mathcal{X}$.*

Proof. First, we observe that $\boldsymbol{\eta}$ is antisymmetric by construction, $\eta_{AB} = -\eta_{BA}$. Thus there are no cycles of length 2. Next assume $\eta_{AB} > 0$ and $\eta_{BC} > 0$. We have to consider the following three cases



Translating the assumption in inequalities of branch lengths in each of the three cases yields:

- **(I)** $a > b$ and $b + u > c$ implies $a + u > c$, i.e., $\eta_{AC} \geq 0$.
- **(II)** $a + u > b$ and $b > c + u$ implies $a > c$, i.e., $\eta_{AC} \geq 0$.
- **(III)** $a > b + u$ and $b > c$ implies $a > c + u$, i.e., $\eta_{AC} \geq 0$.

These three inequalities for η_{AC} assume that the underlying statistical test is "sane" in the sense that it never returns a significantly larger rate for the short branch. Thus $\eta_{AB} > 0$ and $\eta_{BC} > 0$ always implies $\eta_{AC} \geq 0$. Now consider a chain of taxa $\{A^j | 1 \leq j \leq m\}$ such that $\eta_{A^{j-1}A^j} > 0$ for $2 \leq j \leq m$. By repeated application of the this result we conclude $\eta_{A^k,A^l} \geq 0$ for any $l > k$, i.e., the $\{A^j\}$ cannot be part of a directed cycle. Since there is an edge from node i to node j iff $\eta_{i,j} > 0$, we conclude that the corresponding graph is a DAG, and hence the matrix $\boldsymbol{\eta}$ is acyclic. □

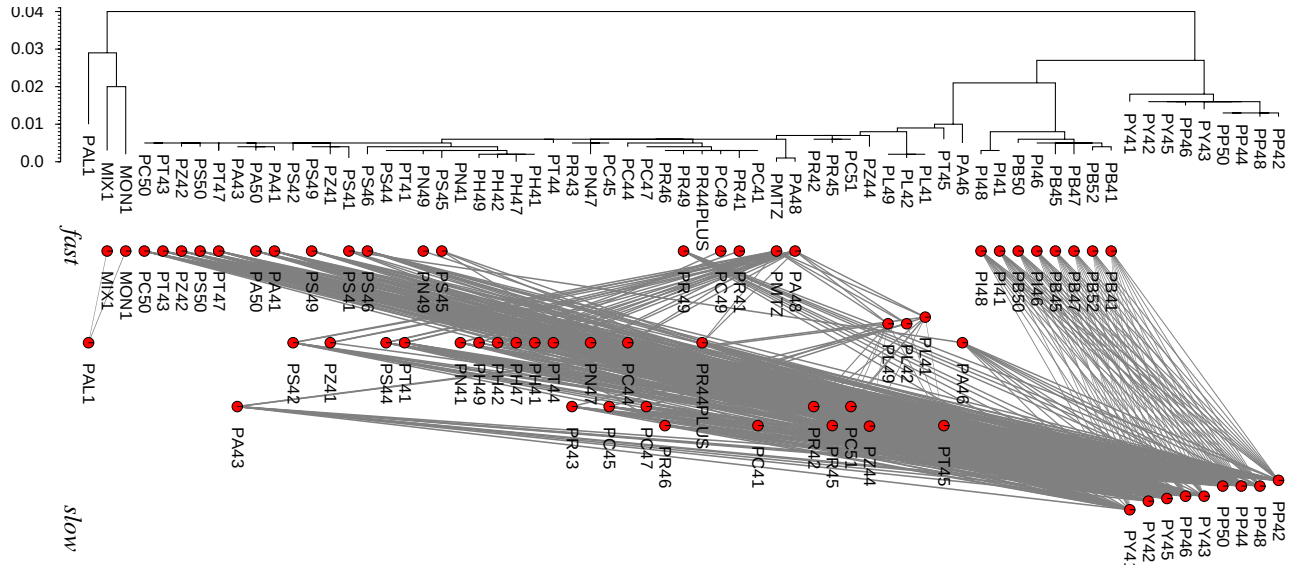


Figure 2. Phylogenetic tree (neighbor joining) and Hasse diagram of the relative-rate poset of mtND1 nucleotide sequence data of wolf spiders of the *Pardosa saltuaria* group [7]. Significance level for Tajima tests $p \leq 0.1$ ($\chi^2 = 2.706$), test results of all subtrees included. Labels refer to geographic locations: North/South Scandinavia PN, PS; Eastern/Western Riesengebirge PC, PR; Tatra Mountains PT; Alps PA, PL, PZ; Eastern/Western Pyrenees PP, PY; Balkans PB, PI; Bohemia PH; Lago di Garda area PMTZ. Outgroup: *P. palustris* PAL, *P. monticola* MON1, *P. mixta* MIX.

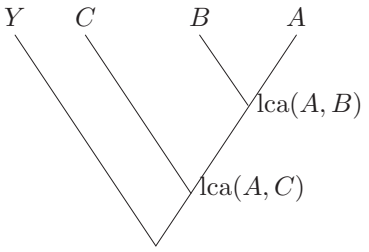
In order to work with real data, we have to relax the assumption that d is an additive tree metric. The estimates for a and b will then depend explicitly on the outgroup C. Note, however, that these variations are small as long as the data are at least approximately tree-like. We can therefore estimate η_{AB} as an *average* over all those triples (A, B, C) for which the Tajima test demonstrates a significant rate difference. The χ^2 value obtained from the Tajima test can be used as weight of the individual estimates. Numerically, we observe that $\boldsymbol{\eta}$ is indeed acyclic even when small χ^2 significance thresholds for the Tajima test are used.

The construction of the matrix $\boldsymbol{\eta}$ starting from a sequence alignment using Tajima's relative rate test has been implemented in a software prototype. It either uses a phylogenetic tree $\mathfrak{T}$ as additional input, or tests for all triples (A, B, C) with outgroup C if $d_{AC}, d_{BC} > d_{AB}$. In order to facilitate the interpretation of the data, it produces a graphical out that compares the phylogenetic tree with the Hasse diagram of the po-set derived from $\boldsymbol{\eta}$, Fig. 1. Points are positioned so that differences along the rate-axis are approximately proportional to differences in $\boldsymbol{\eta}$-values.

2.3. Loss of Phylogenetic Footprints

Relative rate tests can also be designed for more complex settings than substitution rates in homologous sequences. For example, the quantitative analysis of dynamical aspects of footprint loss and acquisition is complicated by the fact that individual regulatory DNA regions cannot be observed independently of sequence conservation. The reason is that phylogenetic footprinting [8, 9, 10, 11] always detects regulatory elements in (at least) pairs of sequences. As a consequence, even very simplistic models of footprint loss lead to rather sophisticated inference.

In the approach proposed in [12], *two* outgroups are required to first identify conserved sequence positions, before one tests for differential loss rates among two ingroup species. More precisely, consider a sub-tree of the following form:



(7)

Restricting the sequences to those positions for which $Y_i = C_i$ holds, we define

$$\begin{aligned} c_{CA} &= |\{i | Y_i = C_i = A_i\}|, \\ c_{CB} &= |\{i | Y_i = C_i = B_i\}|, \\ c_{CAB} &= |\{i | Y_i = C_i = A_i = B_i\}| . \end{aligned} \tag{8}$$

Note that $c_{CA} \geq c_{CAB}$ and $c_{CB} \geq c_{CAB}$ always holds. The number of conserved positions exclusively lost along the edge $A, \mathrm{lca}(A, B)$ is $m'_A = c_{CB} - c_{CAB}$ and similarly, for $B, \mathrm{lca}(A, B)$ we have $m'_B = c_{CA} - c_{CAB}$. One now tests whether m'_A and m'_B are significantly different. The corresponding matrix $\boldsymbol{\eta}$ has entries $\eta_{AB} = \ln(m'_A/m'_B)$ provided the difference is statistically signficant, and $\eta_{AB} = 0$, otherwise. For a fixed combination of outgroups Y, C, we immediately check that $m'_A - m'_{A'} > 0$ and $m'_{A'} - m'_{A''} > 0$ implies $m'_A - m'_{A''} > 0$. We therefore expect $\boldsymbol{\eta}$ to be acyclic. Since the choice of a different outgroup pair may lead to the selection of different conserved position, we cannot logically rule out contradictory

test results in this case, however. The implementation of this test is currently in progress.

3. EXAMPLE

The expansion of a species in a heterogeneous environment can be correlated with relative rates of evolution in geographically separated subpopulations. The rate variation may be due to adaptation to different environmental conditions and due to changes in population size or structure [13]. Slowly evolving populations are typically large and stable, while small unstable populations exhibit higher evolution rates. Multiple waves of migration thus may lead to rate variations that show little correlation with phylogenetic position.

As an example of a real-life data set we consider here a recent comprehensive European-wide phylogeographical study of the arctic-alpine distribution of wolf spiders of the *Pardosa saltuaria* group [7]. The data, mitochondrial ND1 gene sequences, show a complex picture of rate differences, with some clear regularities.

For instance, the substitution rates are increased in almost all lineages relative to the samples from the the Pyrenees. This suggests that the Pyrenees served as glacial refugia. The rate correlation between the sequences of the Pyrenees and the Balkan individuals indicates a secound glacial refugium in the Balkan mountains. However, the data indicate migration out of the Pyrenees refugia only. The data set also reflects one further cold period with refugia in the Alps, Sudeten Mountains, and the Upper Tatra.

4. DISCUSSION

We have introduced here an a convenient way to visualize and summarize information on significant rate differences across larger phylogenetic data sets. The poset-approach seems convenient for the exploratory phase of data analysis. As it stands our tool does not attempt to correct for multiple testing, although a strategy such as Bonferroni's correction could easily be incorporated. We also note that the $\mathcal{O}(N^3)$ RRTs that can be performed within a given tree are of course not independent from each other. It might therefore be desirable to restrict attention to a less redundant set of tests.

5. ACKNOWLEDGMENTS

This work was supported in part by the DFG Bioinformatics Initiative and the 6th Framwork Programme of the European Union as part of the EDEN project (contract no. 043251).

6. REFERENCES

[1] J. Felsenstein, "Phylogenies from molecular sequences: inference and reliability," *Annu. Rev. Genet.*, vol. 22, pp. 521–565, 1988.

[2] F. Tajima, "Simple methods for testing molecular clock hypothesis," *Genetics*, vol. 135, pp. 599–607, 1993.

[3] C.-I. Wu and W.-H. Li, "Evidence for higher rates of nucleotide substitution in rodents than in man," *Proc. Natl. Acad. Sci. USA*, vol. 82, pp. 1741–1745, 1985.

[4] Z. Yang, "Maximum-likelihood models for combined analyses of multiple sequence data," *J. Mol. Evol.*, vol. 42, pp. 587–596, 1996.

[5] P. Li and J. Bousqet, "Relative-rate test for nucleotide substitutions between two lineages," *Mol. Biol. Evol.*, vol. 9, pp. 1185–1189, 1992.

[6] M. Robinson, M. Gouy, C. Gautier, and D. Mouchiroud, "Sensitivity of relative-rate tests to taxonomic sampling," *Mol. Biol. Evol.*, vol. 15, pp. 1091–1098, 1998.

[7] C. Muster and T. U. Berendonk, "Divergence and diversity: lessons from an arctic-alpine distribution (*Pardosa saltuaria* group, lycosidae)," *Mol. Ecol.*, vol. 15, pp. 2921–2933, 2006.

[8] D. A. Tagle, B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess, and R. T. Jones, "Embryonic epsilon and gamma globin genes of a prosimian primate (galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints," *J. Mol. Biol.*, vol. 203, pp. 439–455, 1988.

[9] C. Mayor, M. Brudno, J. R. Schwartz, A. Poliakov, E. M. Rubin, K. A. Frazer, L. S. Pachter, and I. Dubchak, "VISTA: visualizing global DNA sequence alignments of arbitrary length," *Bioinformatics*, vol. 16, pp. 1046–1047, 2000.

[10] M. Blanchette and M. Tompa, "Discovery of regulatory elements by a computational method for phylogenetic footprinting," *Genome Research*, vol. 12, pp. 739–748, 2002.

[11] S. Prohaska, C. Fried, C. Flamm, G. Wagner, and P. F. Stadler, "Surveying phylogenetic footprints in large gene clusters: Applications to Hox cluster duplications," *Mol. Phyl. Evol.*, vol. 31, pp. 581–604, 2004.

[12] G. P. Wagner, C. Fried, S. J. Prohaska, and P. F. Stadler, "Divergence of conserved non-coding sequences: Rate estimates and relative rate tests," *Mol. Biol. Evol.*, vol. 21, pp. 2116–2121, 2004.

[13] C. Stringer and R. McKie, *African Exodus: The Origins of Modern Humanity*, J.Macrae/H.Holt, New York, 1996.

Preprint

Hox cluster duplication in the basal teleost *Hiodon alosoides* (Osteoglossomorpha)

Karen E. Chambers · Ryan McDaniell · Jeremy D. Raincrow · Maya Deshmukh · Peter F Stadler · Chi-hua Chiu

Manuscript date Thu Aug 14 02:02:53 CEST 2008

Abstract Large-scale – even genome-wide – duplications have repeatedly been invoked as an explanation for major radiations. Teleosts, the most species-rich vertebrate clade, underwent a "fish-specific genome duplication" (FSGD) that is shared by most ray-finned fish lineages. We investigate here the *Hox* complement of the goldeye (*Hiodon alosoides*), a representative of Osteoglossomorpha, the most basal teleostean clade. An extensive PCR survey reveals that goldeye has at least eight *Hox* clusters, indicating a duplicated genome compared to basal actinopterygians. The possession of duplicated *Hox* clusters is uncoupled to species richness. The *Hox* system of the goldeye is substantially different from that of other teleost lineages, having retained several duplicates of *Hox* genes for which crown teleosts have lost at least one copy. A detailed analysis of the PCR fragments as well as full length sequences of two *HoxA13* paralogs, and *HoxA10* and *HoxC4* genes places the duplication event close in time to the divergence of Osteoglossomorpha and crown teleosts. The data are consistent with — but do not conclusively prove — that Osteoglossomorpha shares the FSGD.

K.E.Chambers, R.McDaniell, J.D.Raincrow, M.Deshmukh, C.h.Chiu
Department of Genetics, Rutgers University, Piscataway, NJ, USA
E-mail: kchamber@wiley.com, rmmcdaniell@hotmail.com, raincrow@biology.rutgers.edu, maya.deshmukh@gmail.com, chiu@biology.rutgers.edu

P.F. Stadler
Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany; *and*
RNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie — IZI Perlickstrasse 1, D-04103 Leipzig, Germany; *and*
Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria; *and*
Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA
E-mail: studla@bioinf.uni-leipzig.de

Keywords *Hox* clusters, Fish-Specific Genome Duplication, goldeye *Hiodon alosoides*

1 Introduction

Genome duplication is a powerful evolutionary mechanism that has contributed to the diversity of the vertebrate lineage (Ohno, 1970). Present evidence supports that two rounds of genome duplication (called 1R and 2R) occurred in early chordate phylogeny and are common to the ancestor of jawed vertebrates (cartilaginous, lobe-finned, and ray-finned fishes) (Sidow, 1996). The clade of ray-finned fishes (Actinopterygii, Figure 1) underwent a third round of genome duplication dubbed the 3R or the FSGD (fish specific genome duplication, red arrow in Figure 1) (Taylor *et al.*, 2001; Christoffels *et al.*, 2004; Vandepoele *et al.*, 2004). The FSGD is proposed to be a whole genome event (Taylor *et al.*, 2003; Brunet *et al.*, 2006), a fact that is well supported by the observation that spotted green pufferfish (Teleostei; *Tetraodon nigroviridis*) has two syntenic regions (paralogons) corresponding to each single region in the human genome (Jaillon *et al.*, 2004). Comparative mapping, furthermore, shows that paralogons of pufferfish (*Tetraodon*), zebrafish (*Danio*) (Woods *et al.*, 2005) and medaka (*Oryzias*) (Kasahara *et al.*, 2007) are homologous. This supports the view that the FSGD occurred prior to the divergence of these teleosts.

The earliest inklings of the FSGD came from comparative analysis of *Hox* genes and clusters in different chordate lineages (Amores *et al.*, 1998, 2004; Chiu *et al.*, 2002, 2004). *Hox* genes, which encode transcription factors that play a central role in embryonic patterning of the body plan, are usually organized in clusters in the genome, although there are exceptions in some invertebrate lineages (Monteiro and Ferrier, 2006). Evidence to date suggests the basal state of *Hox* clusters in jawed vertebrates is four (**A**,**B**,**C**,**D**),

as is found in cartilaginous (shark (Chiu *et al.*, 2002; Kim *et al.*, 2000; Prohaska *et al.*, 2004; Venkatesh *et al.*, 2007)), lobe-finned (human (Krumlauf, 1994), latimeria (Koh *et al.*, 2003; Powers and Amemiya, 2004)), and basal ray-finned (bichir (Chiu *et al.*, 2004)) fishes.

In contrast, zebrafish has 7 Hox clusters that house expressed genes (**HoxAa**, **Ab**, **Ba**, **Bb**, **Ca**, **Cb**, **Da** (Amores *et al.*, 1998), where **Aa** and **Ab** duplicated clusters are each orthologous to the single **HoxA** cluster of outgroup taxa such as human (Amores *et al.*, 1998, 2004; Chiu *et al.*, 2002) Recently, the **Db** cluster (the 8th cluster) in zebrafish has been found to contain a single microRNA and no open reading frames (ORFs) (Woltering and Durston, 2006). Evidence of duplicated *Hox* clusters is reported for additional teleosts including pufferfishes (*Takifugu rubripes* and *Tetraodon nigroviridis* (Jaillon *et al.*, 2004; Amores *et al.*, 2004; Aparicio *et al.*, 2002), medaka (*Oryzias latipes* (Kasahara *et al.*, 2007; Kurosawa *et al.*, 2006; Naruse *et al.*, 2000), striped bass (*Morone saxatilis* (Snell *et al.*, 1999)), killifish (*Fundulus heteroclitus* (Misof and Wagner, 1996)), cichlids (*Oreochromis niloticus* (Santini and Bernardi, 2005), *Astatotilapia burtoni* (Hoegg *et al.*, 2007; Thomas-Chollier and Ledent, 2008)), salmon (*Salmo salar* (Moghadam *et al.*, 2005b; Mungpakdee *et al.*, 2008)), rainbow trout (*Oncorhynchus mykiss* (Moghadam *et al.*, 2005a)), goldfish (*Carassius auratus* (Luo *et al.*, 2007)), and Wuchang bream (*Megalobrama amblycephala* (Zou *et al.*, 2007)).

Comparative analysis of *Hox* clusters and genes in teleosts showed that the duplicated Hox **a** and **b** clusters have experienced divergent resolution producing variation in gene content (Lynch and Force, 2000; Prohaska and Stadler, 2004) and increased rates of substitution in both protein coding (Chiu *et al.*, 2000; Wagner *et al.*, 2005; Crow *et al.*, 2006) and noncoding (Chiu *et al.*, 2002, 2004; Tumpel *et al.*, 2006) sequences. Consistent with a shared duplication, the *Hox* paralogs form two distinct **a** and **b** clades (Amores *et al.*, 2004). All teleosts examined to-date represent only two species-rich actinopterygian clades, the Ostariophysi (e.g. zebrafish), and Euteleostei (Acanthopterygii: pufferfishes, killifish, medaka, bass, and cichlids; Salmoniformes: salmon, trout), comprising 6,000 and 16,000 species, respectively (Nelson, 1994) (Figure 1).

One may ask whether the FSGD is directly responsible for the biological diversification (i.e. speciosity) of ray-finned fishes (Vogel, 1998; Wittbrodt *et al.*, 1998; Meyer and Schartl, 1999; Venkatesh, 2003; Postlethwait *et al.*, 2004; Meyer and Van de Peer, 2005; Volff, 2005). Alternatively, species-richness and large-scale duplications have to be considered as independent phenomena. The examination of the actinopterygian fossil record (Donoghue and Purnell, 2005) shows that there are 11 extinct clades between teleosts and their closest living relatives. The authors conclude that the character acquisitions often attributed as synapomorphies of

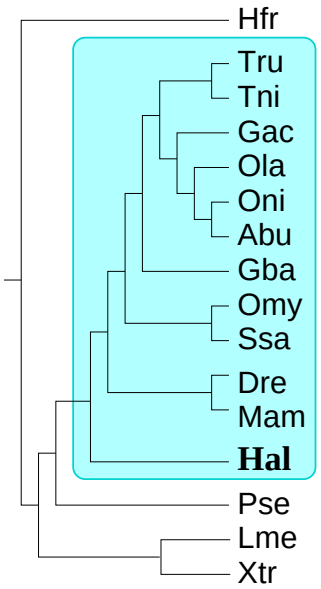


Fig. 1 Simplified phylogeny of jawed vertebrates, with focus on ray-finned fishes (actinopterygians). The jawed vertebrate clade consists of three branches, the cartilaginous (Chondrichthyes), the lobe-finned (Sarcopterygii), and ray-finned (Actinopterygii) fishes (Le *et al.*, 1993; Venkatesh *et al.*, 2001; Kikugawa *et al.*, 2004; Inoue *et al.*, 2003); the close relationship of cichlids is supported by both nuclear genes and phylogenomics data (Chen *et al.*, 2004; Steinke *et al.*, 2006).
Abbreviations: Hfr, *Hetrodontus francisci* (horn shark); Xtr, *Xenopus tropicalis* (frog); Lme, *Latimeria menadoensis* (coelacanth); Pse, *Polypterus senegalus* (bichir); Hal, *Hiodon alosoides* (goldeye); Dre, *Danio rerio* (zebrafish); Mam, *Megalobrama amblycephala*; Ssa, *Salmo salar* (salmon); Omy, *Onkorhynchus mykiss* (rainbow trout); Gba, *Gonostoma bathyphilum* (lightfish); Gac, *Gasterosteus aculateus* (three-spined stickleback); Ola, *Oryzias latipes* (medaka); Oni, *Oreochromis niloticus* (nile tilapia); Abu, *Astatotilapia burtoni*; Tni, *Tetraodon nigroviridis* (spotted green pufferfish); Tru, *Takifugu rubripes* (Japanese pufferfish)

derived teleost fishes arose gradually in ray-finned fish phylogeny with many innovations already predating the FSGD. Many of these extinct clades that have been shown to predate the FSGD were species rich themselves. Hence fossil evidence suggests that the FSGD is uncoupled to species richness. By showing that the species-poor Osteoglossomorpha exhibit duplicated *Hox* clusters, we add molecular evidence to this view.

Evidence from a handful of molecular evolution studies is consistent with this hypothesis. Phylogenetic analyses of four Hox genes (*HoxA11*, *HoxB5*, *HoxC11*, and *HoxD4*) (Crow *et al.*, 2006), duplicated ion and water transporter genes in eels (Cutler and Cramb, 2001), three nuclear genes (fzd8, sox11, tyrosinase (Hoegg *et al.*, 2004), the ParaHox cluster (Mulley *et al.*, 2006), and combined datasets (Hurley *et al.*, 2007) in basal, intermediate and derived actinopterygians together suggest that the FSGD is coincident with the origin of teleosts. More precisely, the data place the duplication event after the divergence of bowfin (Amia) and

Fig. 2 *Hox* cluster complement of chordates with focus on actinopterygians. The Hox cluster of Amphioxus is shown at the top. The *Hox* genes are depicted as colored rectangles for coelacanth (outgroup; red); zebrafish (blue), medaka (light green), tilapia (dark green), Tetraodon (pink) and Fugu (magenta) are shown in the top panel. Putative goldeye *Hox* genes, as inferred from the PCR fragments, are depicted as colored rectangles in the bottom panel. Black rectangles indicate homeoboxes that are assigned to a specific paralog group and cluster (e.g. **B**) but not to a teleostean **a** or **b** clade (see text). Fuscia rectangles indicate homeoboxes that are assigned to a specific paralog group, cluster and clade. Green rectangles depict homeobox fragments assigned to a specific paralog group but not cluster.

sturgeon but prior to the appearance ∼135 mya of the lineages leading to 23,637 (93%) of the 23,681 extant species of present-day teleosts (Benton, 2005).

In order to assess the *Hox* complement in the earliest teleost lineages we identified *Hox* genes in the goldeye (*Hiodon alosoides*), a member of the species-poor Osteoglossomorpha (Nelson, 1994; Hurley *et al.*, 2007; Benton, 2005). Results of a PCR survey of Hox genes in the goldeye coupled with phylogenetic analyses of four individual Hox orthologs (*HoxA10*, *HoxA13-1*, *HoxA13-2*, *HoxC4*) provide conclusive evidence that the goldeye has duplicated *Hox* clusters. The organization of the goldeye *Hox* clusters, however, is significantly different from that of other teleosts, in that it has retained *Hox* genes in all eight clusters.

2 Materials and Methods

2.1 Gnathostome *Hox* Genes

Nucleotide and amino acid sequences of individual *Hox* genes analyzed in this study came from three sources: genome databases, published literature, and targeted PCR amplification using degenerate primers designed here (see below). Amphioxus (*Brachiostoma floridae*) homebox sequences are from (Garcia-Fernàndez and Holland, 1994; Ferrier *et al.*, 2000). The representative of the cartilaginous fishes is horn shark (*Heterodontus francisci*): **HoxA** cluster, ***AF479755***; **HoxD**, cluster ***AF224262***. The representatives of the lobe-finned fishes are coelacanth (*Latimeria menadoensis*) and frog (*Xenopus tropicalis*). Coelacanth homeobox fragments are listed in (Koh *et al.*, 2003); we (Chiu *et al.*, 2000) also sequenced the *HoxA11* ortholog (***AF287139***). Frog *Hox* clusters were taken from the Ensembl Web Browser Xenopus tropicalis genome JGI3: **HoxA**, scaffold29 1,777,789-2,133,531; **HoxB**, scaffold329 415,000-1,016,000; **HoxC**, scaffold280 199,492-581,365; **HoxD** scaffold353 474,676-800,000.

The representatives of the ray-finned fishes include bichir (*Polypterus senegalus*) and several teleost fishes. The bichir **HoxA** cluster was assembled from two BAC clones with accession numbers ***AC126321*** and ***AC132195*** as in (Chiu *et al.*, 2004). Zebrafish (*Danio rerio*) *Hox* clusters were assembled from PAC clones: **HoxAa**, ***AC107364***; **HoxAb**, ***AC107365*** (with an alteration of nucleotide 79,324 from T to C to avoid a premature stop codon); **HoxBa**, ***BX297395***, ***AL645782***; **HoxBb**, ***AL645798***; **HoxCa**, ***BX465864*** and ***BX005254***; the **HoxCb** cluster was taken from Ensembl Web Browser Danio rerio genome (Zv5); **HoxDa**, ***BX322661***. The zebrafish **HoxDb** cluster does not house *Hox* genes (Woltering and Durston, 2006) and was excluded in this study. Nile tilapia (*Oreochromis niloticus*) **HoxAa**, ***AF533976***; striped bass (*Morone saxatilis*) **HoxAa**, ***AF089743***. Medaka (*Oryzias latipes*) **Hox** clusters ***AB232918-AB232924***. Spotted-green pufferfish (*Tetraodon nigroviridis*) Hox clusters were

extracted from the Tetraodon Genome Browser[1]: **HoxAa**, chr21. 2,878,001-3,153,406; **HoxAb**, chr.8 6,506,471-6,727,504; **HoxBa** chr.Un 37,928,410-38,293,032; **HoxBb**, chr.2 1,321,876-1,537,033; **HoxC**, chr.9 4,083,941-4,353,227; **HoxDa**, chr.2 10,975,763-11,218,409 (a T was deleted at position 11,134,740 in order to shift back to correct frame); **HoxDb**, chr.17 9,471,3559,694,740. Japanese pufferfish (*Takifugu rubripes*) *Hox* clusters were acquired from the Ensembl genome browser (assembly FUGU 2.0). The **HoxAa** cluster is constructed from the entire scaffold 47, the **HoxAb** cluster is constructed from scaffold 330, see (Chiu *et al.*, 2002). Short homeobox fragments for QM analysis were in addition taken from (Prohaska and Stadler, 2004).

2.2 PCR amplification, cloning, and sequencing

Whole genomic DNA was extracted from ~ 80 milligrams of ethanol preserved tissue of goldeye (*Hiodon alosoides*) and lightfish (*Gonostoma bathyphilum*) using the DNeasy kit (Qiagen) and protocols.

PCR amplification of an 81 base pair (bp) fragment of the highly conserved homeobox of PG1-8 was performed using a degenerate homebox primer pair [*334*: 5-`GAR YTI GAR AAR GAR TTY`-3; *335*: 5-`ICK ICK RTT YTG RAA CAA`-3]. PCR amplification of an 114 bp fragment of the highly conserved homeobox of PG913 was performed using the degenerate primers [*HB913Forward*: 5-`AAA GGA TCC TGC AGA ARM GNT GYC CNT AYA SNA A`-3; *HB113Reverse*: 5-`ACA AGC TTG AAT TCA TNC KNC KRT TYT GRA ACC A`-3]. PCR amplifications were performed with AmpliTaq Gold DNA polymerase (Applied Biosystems) using the following cycling parameters: initial denaturation at 95°C for 5 min, 30 cycles of 95°C for 1 min, 50°C for 1 min, and 72°C for 1 min, and final extension at 72°C for 10 min. Final concentration of $MgCl_2$ was 3.5 millimolar. Amplified fragments were purified by agarose gel extraction (Qiagen) and cloned into a pGEM-T Easy vector (Promega) following the manufacturers protocol. Clones containing inserts of the correct size were identified using colony PCR and sequenced at the UMDNJ-RWJMS DNA Sequencing and Synthesis Core Facility[2]. For each clone, both strands were sequenced using T7 and SP6 sequencing primers.

2.3 Initial assignment of PCR fragments

The 81 bp and 114 bp long sequences of PG1-8 and PG9-13 homeoboxes, respectively, were compared with the corresponding sequence fragments from a range of chordates (see above). The membership of each PCR fragment to one of the paralog groups *Hox1-Hox13* was initially determined based on nucleotide and amino acid sequence similarity to published Hox sequences using `blast` (Altschul *et al.*, 1990, 1997). The second layer of analysis used neighbor-joining (Saitou and Nei, 1987) trees with deduced amino acid sequences (see Electronic Supplement) and assigned goldeye PCR fragments based on assigned the identity of the subtree in which they are located. With the exception of the "middle-group paralogs" Hox4-Hox7, we find that the paralog-groups are reconstructed as monophyletic clades (with the exception of the posterior sequences from Amphioxus (Garcia-Fernàndez and Holland, 1994; Ferrier *et al.*, 2000).

2.4 Assignment by Quartet Mapping

All subsequent analyses were performed using homeobox nucleotide sequences. Middle-group genes were identified using Quartet Mapping (QM), see (Nieselt-Struwe and von Haeseler, 2001) and an application of QM to homeobox PCR fragments from lower vertebrates (Stadler *et al.*, 2004) for additional details. To this end, we use the teleost homeobox sequences from (Amores *et al.*, 2004), the collection of homeobox fragments from (Prohaska and Stadler, 2004), sequences of human, shark, coelacanth and the bichir **HoxA** cluster (Chiu *et al.*, 2004) as well as sequences from our own unpublished PCR study of the bichir (Raincrow *et al.*, in preparation). We first determine QM support for paralog groups PG4, PG5, and the combination of PG6 and PG7. For those sequences that are not identified as PG4 homeoboxes, we rerun the analysis computing support for PG5, PG6, and PG7.

In a second experiment we then consider trees of the form $((\{x\},R),(U,(V,W)))$ or $((\{x\},(R,U)),(V,W))$, where $\{x\}$ denotes the query sequence from Hiodon and $\{R,U,V,W\} = \{PG4, PG5, PG6, PG7\}$ are the sets of known homeobox sequences from the four middle paralog groups. Together with the query sequence, we thus consider quintets, which can be represented in the form of six inequivalent quartets depending on which pair of paralog groups form a common subtree:
$((\{x\},R)|(U,(V,W)))$; $((\{x\},R)|(V,(U,W)))$; $((\{x\},R)|(W,(U,V)))$; $((\{x\},(R,U))|(V,W))$; $((\{x\},(R,V))|(U,W))$; $((\{x\},(R,W))|(U,V))$.
We analyze each of these six quartets using quartet mapping, i.e., we determine which assignment of the four paralog groups to R, U, V, W yields the maximal support for the tree. This yields a support value for each Hiodon query sequence x to be placed in a common subtree with either a single paralog group or with a pair of paralog groups. Ideally, x is placed together with the same paralog group R three times and placed together with the combination of R and one other paralog group in the remaining three quartets. Our implementation `quartm` of the Quartet Mapping method performs

[1] `http://www.genoscope.cns.fr/externe/tetranew/entry_ggb.html`

[2] `http://www2.umdnj.edu/dnalbweb`

this quartet analysis of quintets automatically. The program can be free downloaded from the authors' website[3].

2.5 Assignment by phylogenetic analysis

The quartet mapping analysis was complemented by the construction of neighbor joining (Saitou and Nei, 1987) and maximum parsimony (Swofford, 2003) trees from the same datasets. In the next step we used the same procedure separately for each paralog group to assign a sequence to one of the four gnathostome clusters **HoxA**, **HoxB**, **HoxC**, **HoxD**. In the final step we then attempted to resolve the assignment of the Hiodon PCR fragments from each class to one of the two teleost-specific paralog groups.

2.6 Sequencing of four *Hox* orthologs

All PCR amplifications were performed with AmpliTaq Gold DNA polymerase (Applied Biosystems). Cloning and sequencing were performed as described above.

Goldeye duplicated *HoxA13-1* and *HoxA13-2* sequences and the lightfish **HoxA13b**-like sequence (Figures 3a and 4) were PCR amplified using universal *HoxA13* primers sequences (Chiu *et al.*, 2004) using the following PCR conditions (initial denaturation at 95°C for 5 min, 30 cycles of 95°C for 1 min, 53°C for 1 min, and 72°C for 3 min, and final extension at 72°C for 10 min. Final concentration of $MgCl_2$ was 2.0 millimolar). The lightfish *Hoxa13b*-like sequence is deposited in Genbank (***bankit1122802***); the goldeye duplicated *HoxA13.1* and *HoxA13.2* sequences have accession numbers ***bankit1122788*** and ***bankit1122792***, respectively.

Two overlapping primer pairs were used to PCR amplify the goldeye *HoxA10*-like sequence (Figure 3c and Supplemental Figure 2). The first set of degenerate primers (*HoxA10Uforward*: 5-`CDG TNC CVG GYT ACT TCC G`-3; *HoxA10Ureverse*: 5-`CCC AAC AAC AKR ARA CTA CC`-3) amplify approximately the last third of exon 1, the intron, and most of exon 2 using the following cycling parameters (initial denaturation at 95°C for 5 min, 30 cycles of 95°C for 1 min, 55°C for 1 min, and 72°C for 1 min, and final extension at 72°C for 10 min. Final concentration of $MgCl_2$ was 3.5 millimolar). To amplify the N terminal portion of exon 1 we designed a forward primer (*PFCA75*: 5-`TTT GYW CRA GAA ATG TCA GC`-3) from an evolutionarily conserved noncoding sequence (*PFCAEF75*; Raincrow *et al.*, in preparation) immediately upstream of the *HoxA10* start codon. PCR using this forward primer and a reverse primer (*Halexon1R*: 5-`CCT TAG AAG TTG CAT AAG CC`-3)that is specific to the goldeye *HoxA10*-like exon 1 sequence (described above), was performed under the reaction conditions (initial denaturation at 95° C for 5 min, 30 cycles of 95°C for 1 min, 55°C for 1 min, and 72°C for 1 min, and final extension at 72°C for 10 min. Final concentration of $MgCl_2$ was 3.0 millimolar). The *HoxA10*-like sequence of goldeye built from a contig of these overlapping PCR fragments, spanning from the promoter to exon 2, is deposited in Genbank (***bankit1122799***).

The *HoxC4* ortholog of bichir (*Polypterus senegalus*, Pse; (***bankit1123044***,***bankit1123047*** and the *HoxC4a*-like paralog of goldeye (Hal; Genbank ***bankit1122797*** were amplified with a degenerate primer pair (*HoxC4Forward*: 5-`CAT GAG CTC GTY TTT GAT GGA`3; *HoxC4Reverse*: 5-`AYT TCA TCC TKC GGT TCT GA`-3) using the following PCR conditions (initial denaturation at 95°C for 5 min, 30 cycles of 95°C for 1 min, 53°C for 1 min, and 72°C for 3 min, and final extension at 72°C for 10 min. Final concentration of $MgCl_2$ was 2.0 millimolar).

2.7 Phylogenetic analysis of exon 1 sequences

Alignments of Hox gene nucleotide sequences were done using the `clustalW` algorithm (Thompson *et al.*, 1994) in the software package MacVector, version 8.1.1, using default settings. Nucleotide sequences were trimmed so each sequence was of equal length. Alignments of *Hox* gene predicted amino acid sequences were done using the `clustalW` algorithm in the software package MacVector version 8.1.1 using default settings. Amino acid alignments were corrected by eye and trimmed so each sequence was of equal length. Alignments can be viewed in the Electronic Supplement.

Maximum Parsimony trees were created using PAUP* v4.0b10 (Swofford, 2003) under the parsimony optimality criterion. Heuristic searches were performed under default settings. Neighbor-Joining (Saitou and Nei, 1987) trees were also created using the PAUP* v4.0b10 package using the distance optimality criterion with default settings. Maximum Likelihood trees were obtained using GARLI v0.951 (Zwickl, 2006). Default settings were used unless otherwise stated below. Starting trees were obtained using heuristic search under the likelihood optimality criterion in PAUP* v4.0b10 (Swofford, 2003), default settings were used. The substitution model was set to the 2 rate model which corresponds to the HKY85 model. Under the Run Termination criteria "Bootstrap repetitions" was set to 2,000 and "Generations without improving topology" was set to 5,000 as suggested in the `GARLI` manual when using bootstrap repetitions. For all three methodes, node confidence was scored using the bootstrap resampling method and 50% cutoff.

Bayesian trees were obtained using `MrBayes` v3.1.2 (Ronquist and Huelsenbeck, 2003) and the parallel version of `MrBayes` v3.1.2 (Altekar *et al.*, 2004). `MrBayes` settings were as follows: 2 rate substitution model, relative rate distribution = gamma, number of generations = 1,000,000, sam-

[3] `http://www.bioinf.uni-leipzig.de/Software/quartm/`

ple freq = 1,000, number of chains = 4, and temperature = 0.2. ”Burn-in” was assessed using the ”sump” command. Normally, the first 1 or 2 trees were discarded as ”burn-in” before creating the final consensus tree. Node confidence was scored using the Bayesian posterior probability provided by the program.

Phylogenetic networks were computed using the neighbor-net algorithm (Bryant and Moulton, 2004) implemented in the `SplitsTree` package (Huson and Bryant, 2006) using the same distance matrices that also underlie the neighbor-joining trees.

3 Results

The first step of this study is to estimate the number of *Hox* clusters in the goldeye (*Hiodon alosoides*). Using degenerate primers that target homeoboxes (see Methods), we cloned and sequenced a total of 421 *Hox* fragments (81 and 114 bp long, depending on the primer set utilized) and 23 non-*Hox* fragments (not further analyzed). Using a combination of `blast` (Altschul *et al.*, 1990, 1997), similarity, Quartet Mapping (QM; (Nieselt-Struwe and von Haeseler, 2001), and phylogenetic analyses (Electronic Supplement[4], the 421 Hox sequences group into 41 unique sequences (Figure 2). For each sequence, allelic exclusion tests were performed as described in (Misof and Wagner, 1996). The 41 homeobox sequences of goldeye found in this study have been deposited in GenBank ***FJ015270-FJ015310***. A full list is provided in the Electronic Supplement.

As shown in Figure 2 (bottom panel), the goldeye has duplicated paralogs on each of the four Hox clusters. For **HoxA**-like clusters, there is evidence for duplicated group 10, 11, and 13 paralogs; **HoxB**-like clusters, group 4; **HoxC**-like clusters, groups 5, 6, 9, 12, 13; and **HoxD**-like clusters, groups 3 and 10. Strikingly, the goldeye is the only teleost fish examined to date that has evidence for retained *Hox* genes on each of the eight Hox clusters (**Aa**, **Ab**, **Ba**, **Bb**, **Ca**, **Cb**, **Da**, **Db**.

Phylogenetic analysis and QM mapping, however, assigned only thirteen sequences to **a** or **b** paralog clades observed in advanced teleost fishes (Figure 2). About the same number of sequences is preferentially classified with the unduplicated genes in bichir, shark, or sarcopterygians. The PCR fragments therefore do not provide enough information to decide whether the goldeye shares the *Hox* duplication with the crown teleosts, i.e., whether its eight *Hox* clusters are orthologous to the eight teleost *Hox* loci, or whether an independent duplication event occured in Osteoglossomorpha.

Because the homeobox sequence amplified in a genomic PCR survey is so short, we chose to further investigate this problem by examining exon sequences of four *Hox* orthologs, *HoxA13* (two paralogs), *HoxA10* and *HoxC4*. For the *HoxA13* locus, we cloned and sequenced the gene proper region of two *HoxA13*-like paralogs (*Hal13.1* and *Hal13.2*) including the beginning of exon 1 (12aa from the start codon), intron, and most of exon 2 including the homeobox. Notably, the homeodomain sequences of *Hal13.1* and *Hal13.2* are identical to homeobox fragments 13.1 and 13.2, respectively, isolated in our independent PCR survey of *H. alosoides* whole genomic DNA.

Interestingly, while homebox fragments *13.1* and *13.2* are tentatively assigned as *HoxA13a* and *HoxA13b* (Figure 2), gene tree reconstructions using *Hal13.1* and *Hal13.2* exon 1 amino acid sequences (Figure 3a) show that both HoxA13-like paralogs of goldeye do not group in either the *HoxA13a* or *HoxA13b* clades of teleost fishes. Instead, both *HoxA13* paralogs of goldeye branch at the base of teleosts, prior to the duplication but after divergence of bichir (*P. senegalus*), the most basal living lineage (Chiu *et al.*, 2004; Mulley *et al.*, 2006). Gene trees reconstructed using exon 1 nucleotide sequences do not resolve the phylogenetic position of the two *HoxA13*-like paralogs (see also Supplemental Figure 1a).

We examined the exon 1 nucleotide sequences of each HoxA13-like paralog in goldeye and did not detect evidence for gene conversion (data not shown). Interestingly though, when we examined the predicted primary amino acid sequence of *Hal13.1* and *Hal13.2* paralogs, we found that they share many amino acids at positions that have diverged in the duplicated paralogs of all crow teleosts (zebrafish (Chiu *et al.*, 2002), medaka (Kasahara *et al.*, 2007; Naruse *et al.*, 2000; Kurosawa *et al.*, 2006), tilapia (Santini and Bernardi, 2005), lightfish (this study) and pufferfishes (Jaillon *et al.*, 2004; Aparicio *et al.*, 2002)), see Fig. 4. The amino acid positions shared by the duplicated *HoxA13*-like paralogs in goldeye are the ancestral sites, as determined by their shared presence in the bichir (*Polypterus senegalus*), which has a single **HoxA** cluster (Chiu *et al.*, 2004). We examined whether there is selection acting on synonymous substitutions (Ks) at these two loci in the goldeye (Yang, 1997), but we did not find any statistical support (data not shown). Our findings for the goldeye *HoxA13*-like paralogs are striking because they do not exhibit a pattern of sequence evolution consistent with intensive diversifying selection (van de Peer *et al.*, 2001; Crow *et al.*, 2006) following duplication. The goldeye thus may be a good model to test the predictions of the DDC model (Force *et al.*, 1999), whereby amino acid sequence divergence of duplicated paralogs may be small but divergence in regulatory sequences is large.

Using overlapping primer sets (see below), we cloned and sequenced the gene proper region of a *HoxA10*-like sequence (Figure 3b) including a promoter sequence (not shown). The homeodomain sequence of the *HoxA10*-like ortholog is an exact match to fragment 10-1 (Figure 2), assigned as a

[4] `http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/Hiodon/`

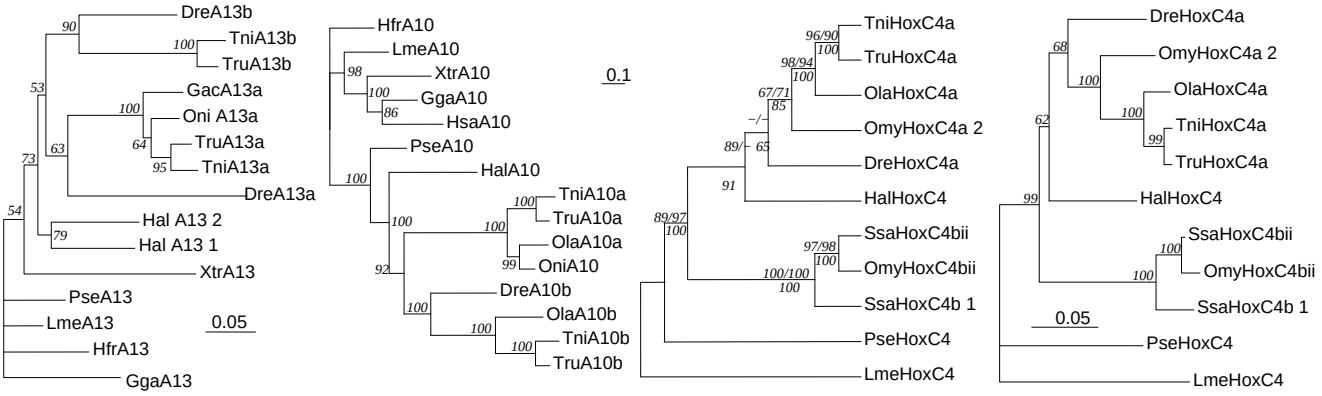


Fig. 3 Examples of phylogenetic analysis of *Hox* exon 1 sequences. Species abbreviations as in Fig. 1. (A) *HoxA13* tree reconstructed using neighbour-joining (Saitou and Nei, 1987) analysis of *HoxA13* amino acid sequences. Bootstrap support (2000 replications) are shown at the nodes. (B) *HoxA10* tree reconstructed using Bayesian (Ronquist and Huelsenbeck, 2003; Altekar *et al.*, 2004) analysis of amino acid sequences. Node confidence values of 1,000,000 generations are shown. (C) Consensus *HoxC4* tree reconstructed using Neighbor joining (Saitou and Nei, 1987), heuristic maximum parsimony (Swofford, 2003), and maximum likelihood *Swofford:03,Zwickl:06* analyses of amino acid sequences. Node confidence values are listed as NJ/HMP/B. (D) Consensus *HoxC4* tree reconstructed using Neighbor joining analysis of nucleotide sequences. Node confidence values are listed as NJ/MP/B/ML. See text for details of phylogenetic analysis.

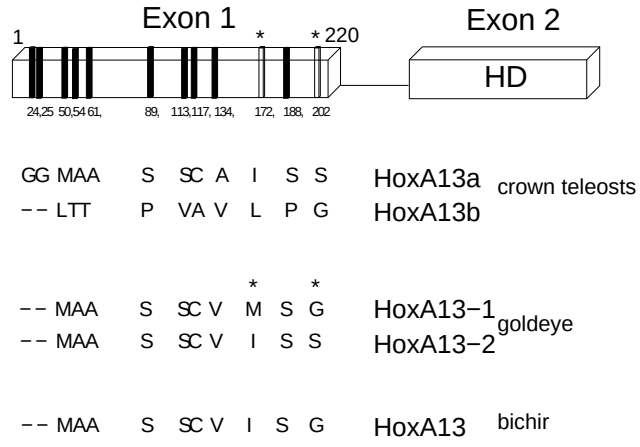


Fig. 4 Goldeye duplicated *HoxA13*-like paralogs do not diverge at the amino acid level. Cartoon depiction of *HoxA13* exon 1 and exon 2 domains. Amino acid numbers according to *HoxA13a* of pufferfish (*Takifugu*), see text. Amino acid positions (black bars) that diverge in the duplicated *HoxA13a* and *Hoxa13b* paralogs of species-rich teleosts are shown and contrasted with the duplicated *HoxA13*-like paralogs of goldeye. Only two of amino acid positions diverge in goldeye (asterisks). See text for further description.

HoxA10 homeobox. As illustrated in phylogenetic analysis of exon 1 amino acid sequences, the *HoxA10*-like sequence of goldeye branches outside the duplicated *HoxA10a* and *HoxA10b* clades (Figure 3b), similarly to the *HoxA13*-like paralogs (Figure 3a). The topology of this gene tree is similar to that reported in (Hurley *et al.*, 2007) for other nuclear genes. Interestingly, the promoter of the goldeye *HoxA10*-like ortholog also has not acquired diagnostic teleostean paralog **a** and **b** specific nucleotides (not shown). There are at least two possibilities that could account for these results. First, following *Hox* cluster duplication, goldeye retains only a single *HoxA10* locus that did not accumulate substitutions at an increased rate observed when both duplicated paralogs are retained following duplication in teleost crown groups (Chiu *et al.*, 2000; Wagner *et al.*, 2005; van de Peer *et al.*, 2001). In fact, phylogenetic analysis of exon 1 of the single HoxA10b locus in zebrafish provides strong support for branching within the teleostean **b** clade only at the amino acid (Figure 3b), but not nucleotide sequence (Supplemental Figure 1b) level. Hence, following a duplication, if one of the paralogs is immediately lost, the rate of nucleotide substitution of the remaining singlet may be conservative. A second possibility raised by our findings is that goldeye experienced a duplication that is independent from that in the crown group of ostariphysians and acanthomorphs. A third scenario, although not tenable with available data, is that goldeye experienced massive gene loss shortly after the FSGD and subsequently experienced lineage specific duplications of all or parts of its genome, including the *Hox* clusters, minimally the HoxA-like cluster.

Intriguingly, phylogenetic analysis of the majority of exon 1 of a *HoxC4*-like sequence found in this study provides strong support that this locus is *HoxC4a*-like at the level of amino acid (Figure 3c) and nucleotide (Figure 3d) sequences. Hence, this result supports that goldeye shares the FSGD. Importantly, the homeodomain sequence of this *HoxC4a*-like locus is an identical match to our PCR homeobox survey fragment *4-5* (Figure 2) that we independently assigned as *HoxC4a* using phylogenetic methods and QM (Table 1 in the Electronic Supplement). This result, i.e., that goldeye experienced the FSGD, is consistent with the phylogenetic branching arrangement of three Hox genes *HoxA11*α, *Hoxa11*β, and HoxB5β in goldeye into HoxA11a, HoxA11b, and HoxB5b teleostean clades, respectively (Crow *et al.*,

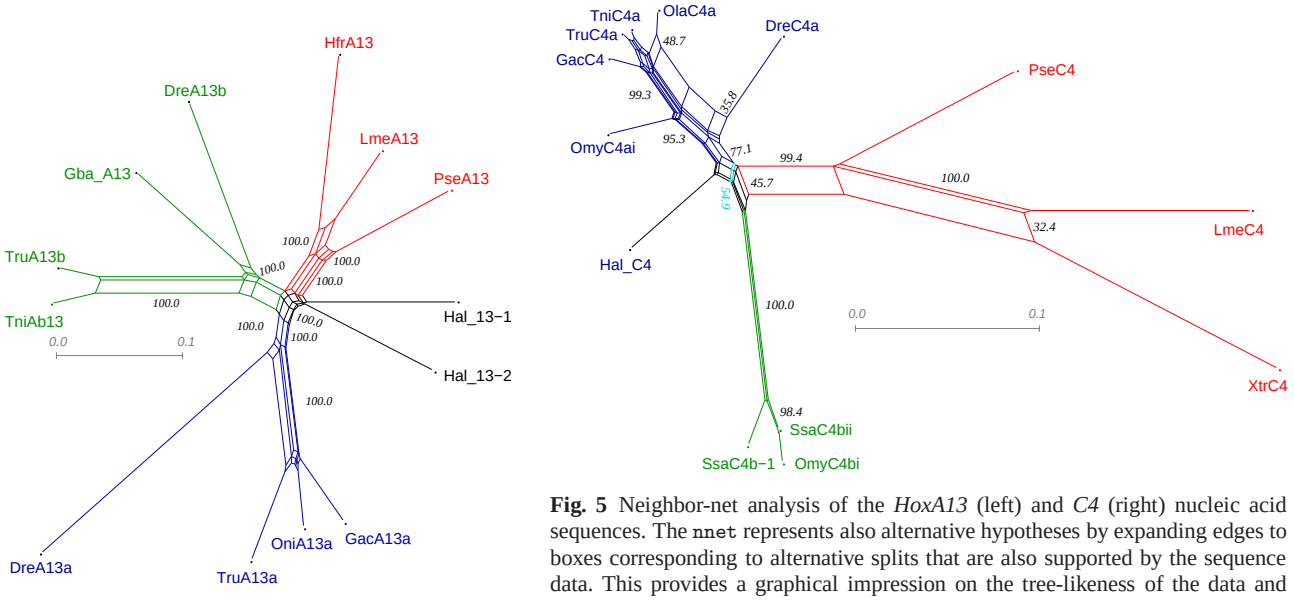


Fig. 5 Neighbor-net analysis of the *HoxA13* (left) and *C4* (right) nucleic acid sequences. The `nnet` represents also alternative hypotheses by expanding edges to boxes corresponding to alternative splits that are also supported by the sequence data. This provides a graphical impression on the tree-likeness of the data and vizalizes the signal to noise ratio of the data set.

2006). Interestingly, our PCR survey above detected two unique HoxA11-like homeobox fragments (*11-1*, *11-2*, Figure 2 that both are assigned, with weak support, to be HoxA11b-like. Our PCR screen did not yield HoxB5-like homeobox sequences.

4 Discussion

Our findings contribute to the understanding of the Hox complement in a basal teleost lineage (Figure 2) and permit inferences on when duplicate Hox paralogs have been lost in actinopterygian phylogeny.

While acantomorpha have completely lost one of the **HoxC** duplicates, and ostariophysi as well as Salmoniformes have lost all protein coding genes from one of the **HoxC** duplicates, goldeye has retained *Hox* genes of all eight clusters. As illustrated in Figure 2, goldeye in particular possesses duplicate paralogs of *HoxB4*, *HoxC5*, *HoxC6*, *HoxD3*, and *HoxD10*.In contrast zebrafish, with the exception of HoxC6 (Amores *et al.*, 1998), medaka (Kasahara *et al.*, 2007; Naruse *et al.*, 2000; Kurosawa *et al.*, 2006) cichlids (Santini and Bernardi, 2005; Hoegg *et al.*, 2007; Thomas-Chollier and Ledent, 2008), and pufferfishes (Aparicio *et al.*, 2002; Jaillon *et al.*, 2004), each possess at most a single copy of these loci (Figure 2). Based on fossil evidence, we infer that these genes were lost in the time interval spanning from 250 million years ago (Amia) to 135 million years ago (appearance of ostariophysans) (Benton, 2005).

The functional consequences of this seeming bias in gene losses remain to be explored. One prediction is that the remaining single ortholog of each locus may exhibit a pattern of sequence evolution diagnostic of negative or stabilizing selection, which is in contrast to the pattern of strong positive selection (i.e. molecular adaptation with $K_a/K_s > 1$) that has been reported when duplicated paralogs are retained, such as the zebrafish *HoxC6a* and *HoxC6b* paralogs (van de Peer *et al.*, 2001), **HoxA** cluster duplicated paralogs of ostariophysan and acanthomorph lineages (Chiu *et al.*, 2000; Wagner *et al.*, 2005) and other nuclear loci (Brunet *et al.*, 2006).

The duplication of the *Hox* gene system in goldeye together with previously reported duplications (relative to the gnathostome ancestor) of several other nuclear genes in other bony tongues (Hoegg *et al.*, 2004) suggests that we are dealing with a whole-genome duplication. A genome duplication, or the possession of a duplicated *Hox* system in particular, is therefore uncoupled from species-richness. Our results emphasize the genome plasticity of actinopterygians in general and suggest that different mechanisms may be at work in the earliest (species poor) versus later (species rich) teleost fishes.

Strictly speaking, our data fail to conclusively resolve the question whether or not the duplicated *Hox* clusters in goldeye are true orthologs of the eight teleostean clusters. As illustrated in Figure 3a, the branch length of each *HoxA13*-like sequence in goldeye is long, suggesting they derive from an ancient duplication and not a lineage specific duplication as observed in paddlefish for *HoxB5* duplicated paralogs (Crow *et al.*, 2006). The ambiguity of the phylogenetic analysis, furthermore, in itself implies that the duplication observed in osteoglossomorpha must have been very *close* in time to the divergence of this lineage from crown teleosts, a conclusion also drawn in (Crow *et al.*, 2006). This is illustrated nicely by the phylogenetic networks in Figure 5,

which show that the phylogenetic signal (branch lengths) separating the FSGD from the divergence of Osteoglossomorpha and crown teleosts is comparable to the noise inherent in the available data.

In conclusion, our analysis is consistent both with independent duplications in both lineages shortly after the osteoglossomorpha-crown teleost split, and with the — more parsimonious — interpretation of a single FSGD pre-dating this divergence (Crow *et al.*, 2006). We suspect that a definitive resolution of this question will require genome-wide data as well as a denser taxon sampling at key points in actinopterygian phylogeny.

Acknowledgements Goldeye and lightfish genomic DNAs are a gift from Dr. Guillermo Orti. We are grateful to Ms. Inna Zamanskaya and Mr. Suley Kuyumcu for technical assistance, and Bärbel M. R. Stadler for carefully proofreading this manuscript. This work was supported by the National Science Foundation, MCB 0447478 (C.H.C.) and the Bioinformatics Initiative of the Deutsche Forschungs Gemeinschaft, BIZ-6/1-2 (P.F.S).

Supplemental Material

An extensive **Electronic Supplement** provides further details on the phylogenetic analysis of the exon-1 sequences (Supplementary Figure 1), Tables detailing the phylogenetic and quartet mapping analysis of the PCR fragments, as well as machine readable files containing sequences and alignments. The electronic Supplement can be found at `http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/Hiodon/`.

References

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F, 2004. Parallel Metropolis-coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20:407–415.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, 1997. Gapped `BLAST` and `PSI-BLAST`: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH, 1998. Zebrafish hox clusters and vertebrate genome evolution. Science 282:1711–1714.

Amores A, Suzuki T, Yan YL, Pomeroy J, Singer A, Amemiya C, Postlethwait J, 2004. Developmental roles of pufferfish *Hox* clusters and genome evolution in ray-fin fish. Genome Res 14:1–10.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia Jm, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MDS, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJK, Dogget N, Zharkikh A, Tavtigian SV, Pruss D, Barstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, H. TY, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S, 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301–1310.

Benton MJ, 2005. Vertebrate Paleontology. Malden: Blackwell, 3rd edn.

Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M, 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol Biol Evol 23:1808–1816.

Bryant D, Moulton V, 2004. Neighbor-net: An agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21:255–265.

Chen WJ, Orti G, Meyer A, 2004. Novel evolutionary relationship among four fish model systems. Trends Genet 20:424–431.

Chiu CH, Amemiya C, Dewar K, Kim CB, Ruddle FH, Wagner GP, 2002. Molecular evolution of the HoxA cluster in the three major gnathostome lineages. Proc Natl Acad Sci USA 99:5492–5497.

Chiu CH, Dewar K, Wagner GP, Takahashi K, Ruddle F, Ledje C, Bartsch P, Scemama JL, Stellwag E, Fried C, Prohaska SJ, Stadler PF, Amemiya CT, 2004. Bichir *HoxA* cluster sequence reveals surprising trends in ray-finned fish genomic evolution. Genome Res (p. 14). 11-17.

Chiu CH, Nonaka D, Xue L, Amemiya CT, Wagner GP, 2000. Evolution of *HoxA11* in lineages phylogenetically positioned along the fin-limb transistion. Mol Phylogenet Evol 17:305–316.

Christoffels A, Koh EGL, Chia JM, Brenner S, Aparicio S, Venkatesh B, 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol Biol Evol 21:1146–1151.

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.

Cutler CP, Cramb G, 2001. Molecular physiology of osmoregulation in eels and other teleosts: the role of transporter isoforms and gene duplication. Comp Biochem Physiology A 130:551–564.

Donoghue PCJ, Purnell MA, 2005. Genome duplication, extinction, and vertebrate evolution. Trends Ecol Evol 20:312–319.

Ferrier DEK, Minguillón C, Holland PWH, Garcia-Fernàndez J, 2000. The amphioxus Hox cluster: deuterostome posterior flexibility and *Hox14*. Evol Dev 2:284–293.

Force A, Lynch M, Pickett FB, Amores A, Yan Yl, Postlethwait J, 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.

Garcia-Fernàndez J, Holland PW, 1994. Archetypal organization of the amphioxus hox gene cluster. Nature 370:563–566.

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.

Hoegg S, Brinkmann H, Taylor JS, Meyer A, 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. J Mol Evol 59:190–203.

Hurley IA, Mueller RL, Dunn KA, Schmidt EJ, Friedman M, Ho RK, Prince VE, Yang Z, Thomas MG, , Coates MI, 2007. A new time-scale for ray-finned fish evolution. Proc Biol Sci 274:489–498.

Huson DH, Bryant D, 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267.

Inoue JG, Miya M, Tsukamoto K, Nishida M, 2003. Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the " ancient fish". Mol Phylog Evol 26:110–120.

Jaillon O, Aury J, Brunet F, Petit J, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biémont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau J, Gouzy J, Parra G, Lardier G, Chapple C, McKernan K, McEwan P, Bosak S, Kellis M, Volff J, Guigó R, Zody M, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quétier F, Saurin W, Scarpelli C, Wincker P, Lander E, Weissenbach J, Roest Crollius H, 2004. Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431:946–957.

Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y, Jindo T, Kobayashi D, Shimada A, Toyoda A, Kuroki Y, Fujiyama A, Sasaki T, Shimizu A, Asakawa S, Shimizu N, Hashimoto S, Yang J, Lee Y, Matsushima K, Sugano S, Sakaizumi M, Narita T, Ohishi K, Haga S, Ohta F, Nomoto H, Nogata K, Morishita T, Endo T, Shin-I T, Takeda H, Morishita S, Kohara Y, 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature 447:714–719.

Kikugawa K, Katoh K, Kuraku S, Sakurai H, Ishida O, Iwabe N, Miyata T, 2004. Basal jawed vertebrate phylogeny inferred from multiple nuclear DNA-coded genes. BMC Biol 2:3.

Kim CB, Amemiya C, Bailey W, Kawasaki K, Mezey J, Miller W, Minosima S, Shimizu N, P. WG, Ruddle F, 2000. Hox cluster genomics in the horn shark, *heterodontus francisci*. Proc Natl Acad Sci USA 97:1655–1660.

Koh EGL, Lam K, Christoffels A, Erdmann MV, Brenner S, Venkatesh B, 2003. *Hox* gene clusters in the indonesian coelacanth, *Latimeria menadoensis*. Proc Natl Acad Sci USA 100:1084–1088.

Krumlauf R, 1994. Hox genes in vertebrate development. Cell 78:191–201.

Kurosawa G, Takamatsu N, Takahashi M, Sumitomo M, Sanaka E, Yamada K, Nishii K, Matsuda M, Asakawa S, Ishiguro H, Miura K, Kurosawa Y, Shimizu N, Kohara Y, Hori H, 2006. Organization and structure of *hox* gene loci in medaka genome and comparison with those of pufferfish and zebrafish genomes. Gene 370:75–82.

Le HL, Lecointre G, Perasso R, 1993. A 28S rRNA-based phylogeny of the gnathostomes: First steps in the analysis of conflict and congruence with morphologically based cladograms. Mol Phylogenet Evol 2:31–51.

Luo J, Stadler PF, He S, Meyer A, 2007. PCR survey of Hox genes in the Goldfish *Carassius auratus auratus*. J Exp Zool B Mol Devel Evol 308B:250–258.

Lynch M, Force A, 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473.

Meyer A, Schartl M, 1999. Gene and genome duplications in vertebrates: the one-tofour (-to-eight in fish) rule and the evolution of novel gene functions. Curr Opin Cell Biol 11:699–704.

Meyer A, Van de Peer Y, 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). BioEssays 27:937–945.

Misof BY, Wagner GP, 1996. Evidence for four Hox clusters in the killifish *Fundulus Heteroclitus* (teleostei). Mol Phylog Evol 5:309–322.

Moghadam HK, Ferguson MM, Danzmann RG, 2005a. Evidence for *Hox* gene duplication in rainbow trout (*Oncorhynchus mykiss*): A tetraploid model species. J Mol Evol 61:804–818.

Moghadam HK, Ferguson MM, Danzmann RG, 2005b. 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.

Monteiro AS, Ferrier DEK, 2006. Hox genes are not always colinear. Int J Biol Sci 2:95–103.

Mulley JF, Chiu CH, Holland PW, 2006. Breakup of a homeobox cluster after genome duplication in teleosts. Proc Natl Acad Sci USA 103:10369–10372.

Mungpakdee S, Seo HC, Angotzi AR, Dong X, Akalin A, Chourrout D, 2008. Differential evolution of the 13 Atlantic salmon *Hox* clusters. Mol Biol Evol 25:1333–1343.

Naruse K, Fukamachi S, Mitani H, Kondo M, Matsuoka T, Kondo S, Hanamura N, Morita Y, Hasegawa K, Nishigaki R, Shimada A, Wada H, Kusakabe T, Suzuki N, Kinoshita M, Kanamori A, Terado T, Kimura H, Nonaka M, Shima A, 2000. A detailed linkage map of medaka, *Oryzias latipes*: Comparative genomics and genome evolution. Genetics 154:1773–1784.

Nelson JS, 1994. Fishes of the world. New York: John Wiley & Sons Inc., 3rd edn.

Nieselt-Struwe K, von Haeseler A, 2001. Quartet-mapping, a generalization of the likelihood mapping procedure. Mol Biol Evol 18:1204–1219.

Ohno S, 1970. Evolution by gene duplication. New York: Springer-Verlag.

Postlethwait J, Amores A, Cresko W, Singer A, Yan YL, 2004. Subfunction partitioning, the teleost radiation and the annotation of the human genome. Trends Genet 20:481–490.

Powers TP, Amemiya CT, 2004. Evidence for a hox14 paralog group in vertebrates. Current Biol 14:R183–R184.

Prohaska SJ, Fried C, Amemiya CT, Ruddle FH, Wagner GP, Stadler PF, 2004. The shark HoxN cluster is homologous to the human HoxD cluster. J Mol Evol (p. 58). 212-217.

Prohaska SJ, Stadler PF, 2004. The duplication of the hox gene clusters in teleost fishes. Th Biosci 123:89–110.

Ronquist F, Huelsenbeck JP, 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Saitou N, Nei M, 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425.

Santini S, Bernardi G, 2005. Organization and base composition of tilapia *Hox* genes: implications for the evolution of *Hox* clusters in fish. Gene 346:51–61.

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

Snell EA, Scemama JL, Stellwag EJ, 1999. Genomic organization of the hoxa4-hoxa10 region from *Morone saxatilis*: implications for hox gene evolution among vertebrates. J Exp Zool Mol Dev Evol 285:41–49.

Stadler PF, Fried C, Prohaska SJ, Bailey WJ, Misof BY, Ruddle FH, Wagner GP, 2004. Evidence for independent *Hox* gene duplications in the hagfish lineage: A PCR-based gene inventory of *Eptatretus stoutii*. Mol Phylog Evol 32:686–692.

Steinke D, Salzburger W, Meyer A, 2006. Novel relationships among ten fish model species revealed based on a phylogenomic analysis using ESTs. J Mol Evol 62:772–784.

Swofford DL, 2003. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4. Sunderland, MA: Sinauer Associates. Handbook and Software.

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.

Taylor JS, Van de Peer Y, Meyer A, 2001. Revisiting recent challenges to the ancient fish-specific genome duplication hypothesis. Curr Biol 11:R1005–R1008.

Thomas-Chollier M, Ledent V, 2008. Comparative phylogenomic analyses of teleost fish *Hox* gene clusters: lessons from the cichlid fish *Astatotilapia burtoni*: comment. BMC Genomics 9:35.

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.

Tumpel S, Cambronero F, Wiedemann LM, Krumlauf R, 2006. Evolution of *cis* elements in the differential expression of two Hoxa2 coparalogous genes in pufferfish (*Takifugu rubripes*). Proc Natl Acad Sci USA 103:5419–5424.

van de Peer Y, Taylor JS, I. B, Meyer A, 2001. The ghost of selection past: rates of evolution and functinal divergence of anciently duplicated genes. J Mol Evol 53:436–446.

Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y, 2004. Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci USA 101:1638–1643.

Venkatesh B, 2003. Evolution and diversity of fish genomes. Curr Opin Genet Dev 13:588–592.

Venkatesh B, Erdmann MV, Brenner S, 2001. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. Proc Natl Acad Sci USA 98:11382–11387.

Venkatesh B, Kirkness EF, Loh YH, Halpern AL, Lee AP, Johnson J, Dandona N, Viswanathan LD, Tay A, Venter JC, Strausberg RL, Brenner S, 2007. Survey sequencing and comparative analysis of the elephant shark (*Callorhinchus milii*) genome. PLoS Biol 5:e101.

Vogel G, 1998. Doubled genes may explain fish diversity. Science 281:1119–1121.

Volff JN, 2005. Genome evolution and biodiversity in teleost fish. Heredity 94:280–294.

Wagner GP, Takahashi K, Lynch V, Prohaska SJ, Fried C, Stadler PF, Amemiya CT, 2005. Molecular evo-

lution of duplicated ray finned fisch hoxa clusters: Increased synonymous substitution rate and asymmetrical co-divergence of coding and non-coding sequences. J Mol Evol (pp. 665–676).

Wittbrodt J, Meyer A, Schartl M, 1998. More genes in fish? Bioessays 20:511512.

Woltering JM, Durston AJ, 2006. The zebrafish *hoxDb* cluster has been reduced to a single microRNA. Nat Genet 38:601–602.

Woods IG, Wilson C, Friedlander B, Chang P, Reyes DK, Nix R, Kelly PD, Chu F, Postlethwait JH, Talbot WS, 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. Genome Res 15:1307–1314.

Yang Z, 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556.

Zou SM, Jiang XY, He ZZ, Yuan J, Yuan XN, Li SF, 2007. *Hox* gene clusters in blunt snout bream, *Megalobrama amblycephala* and comparison with those of zebrafish, fugu and medaka genomes. Gene 400:60–70.

Zwickl DJ, 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. thesis, The University of Texas at Austin.

Preprint manuscript No.
(will be inserted by the editor)

Manuela Marz, Toralf Kirsten, Peter F. Stadler

Evolution of Spliceosomal snRNA Genes in Metazoan Animals

March 20, 2008

Abstract While studies of the evolutionary histories of protein families are common place, little is known on noncoding RNAs beyond microRNAs and some snoRNAs. Here we investigate in detail the evolutionary history of the 9 spliceosomal snRNA families (U1, U2, U4, U5, U6, U11, U12, U4atac, and U6atac) across the completely or partially sequenced genomes of metazoan animals.
Representatives of the five major spliceosomal snRNAs were found in all genomes. None of the minor splicesomal snRNAs was detected in Nematodes and in the shotgun traces of *Oikopleura dioica*, while in all other animal genomes at most one of them is missing. Although snRNAs are present in multiple copies in most genomes, distinguishable paralog groups are not stable over long evolutionary times, although they appear independently in several clades. In general, animal snRNA secondary structures are highly conserved, albeit in particular U11 and U12 in insects exhibit dramatic variations. An analysis of genomic context of snRNAs reveals that they behave like mobile elements, exhibiting very little syntenic conservation.

M. Marz
Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany
E-mail: manja@bioinf.uni-leipzig.de

T. Kirsten
Interdisziplinäres Zentrum für Bioinformatik, Härtelstrasse 16-18, D-04107 Leipzig, Germany
E-mail: kirsten@izbi.uni-leipzig.de,

P.F. Stadler
Bioinformatics Group, Department of Computer Science, and Interdisziplinäres Zentrum für Bioinformatik, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany;
RNomics Group, Fraunhofer Institute for Immunology and Cell Therapy, Leipzig;
Institute for Theoretical Chemistry and Structural Biology, University of Vienna, Währingerstrasse 17, A-1090 Wien, Austria; and
The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe NM 87501
E-mail: studla@bioinf.uni-leipzig.de

1 Introduction

In most eukaryote lineages, introns are spliced out of protein-coding mRNAs by the spliceosome, a huge RNP complex consisting of about 200 proteins and five small non-coding RNAs [58]. These snRNAs exert crucial catalytic functions in the process [86,88,87] in three distinct splicing machineries. The *major spliceosome*, containing the snRNAs U1, U2, U4, U5 and U6, is the dominant form in metazoans, plants, and fungi, and removes introns with GT-AG (as well as rarely AT-AC and GC-AG) boundaries. Another class of "non-canonical" introns with AT-AC (and rarely GT-AG [71]) boundaries is excised by the *minor spliceosome* [61], which contains the snRNAs U11, U12, U4atac, U5, and U6atac. Just as the major spliceosome, the minor spliceosome is present across most eukaryotic lineages and traces back to an origin very early in the eukaryote evolution [9,44,65]. Recently it was found that the minor spliceosome can also act outside the nucleus and controls cell proliferation [35]. Functional and structural differences of two spliceosomes are reviewed in [89]. The third type of splicing the *SL-trans-splicing*, in which a "miniexon" derived from the non-coding spliced-leader RNA (SL) is attached to each protein-coding exon. The corresponding spliceosomal complex requires the snRNAs U2, U4, U5, and U6, as well as an SL RNA [24]. Due to the high sequence variation of the short SL RNAs, and the patchy phylogenetic distribution of SL-trans-splicing, the evolutionary origin(s) of this mechanism, which is active at least in chordates, nematodes, cnidarians, euglenozoa, and kinetoplastids, is still unclear.

Previous studies on the evolutionary origin of the spliceosomes have been performed predominantly based on homology of the most important spliceosomal proteins. Thus relatively little detail is known on the evolution of the snRNA sequences themselves beyond the homology of nine famlilies of snRNAs across all eukaryotes studies so far [73,69,10,44, 9,65]. This may come as a surprise since it has been known for more than a decade that at least all of the snRNAs of the major spliceosome appear in multiple copies and that these paralogs are differentially regulated in at least some

species, see e.g. [43,80,79,5,52]. Very recently, however, some of these variants have been studied in more details, see e.g. [64,8,39,77,29,78] and the references therein. The only systematic study that we are aware of is the recent comprehensive analysis of 11 insect genomes [53] which reported that phylogenetic gene trees of insect snRNAs do not provide clear support for discernible paralog groups of U1 and/or U5 snRNAs that would correspond to the variants with tissue-specific expression patterns. Instead, the analysis supports a concerted mode of evolution and/or extreme purifying selection, a scenario previously described for snRNA evolution [42,40,57].

In this contribution we extend the detailed analysis of the nine spliceosomal snRNAs to metazoan animals. In particular in mammals, the analysis is complicated by high copy number of snRNAs of the major spliceosome and an associated large number of pseudogenes [13]. We focus here on four questions: (1) Is there evidence for discernible paralog groups of snRNAs in some clades? A dominating mode of concerted evolution does not necessarily prevent this, as demonstrated by the existence of two highly diverged copies of both LSU and SSU rRNA in Chaetognatha [83,60], which is probably associated with a duplication of the entire rDNA cluster. (2) Are there clades with deviant snRNA structures? The prime example for a highly divergent snRNA is the U11 in a subset of the insects [69]. (3) Are there interpretable trends in the copy number of snRNAs across metazoa? (4) How mobile are snRNA genes relative to the "background" of protein coding genes? In other words, to what extent are some or all of the snRNA genes off-springs of a locus that remains stably linked to its context over large time-scales.

2 Materials and Methods

2.1 Sequence Data

Known snRNA sequences were retrieved from `Genbank` [4], `Rfam` [23], and in some cases extracted directly from the literature. Genomic DNA sequences were downloaded from the websites of `ensembl`, the Joint Genome Institute, the Sanger Institute, WormBase, the Genome Sequencing Center, UCSC, CAF1, Broad Institute, BGI, and the NCBI trace archive. For some species, we also performed non-exhaustive searches in the `NCBI Trace Archive` using `megablast`. Details on the dataset can be found in the Electronic Supplement.[1]

Over all, the published experimental evidence on metazoan snRNAs is very unevenly distributed. For example, a large and phylogentically diverse set of U2 snRNA sequences is reported in [20], while most other snRNAs have mostly been reported for a few model organisms only. A recent experimental screen for snRNAs in *Takifugu rubripes* [55] resulted in copies of eight snRNAs families. U4atac was missing, but a plausible candidate can easily be found by `blast`. Only a few sequences of minor spliceosomal snRNAs have been reported so far, mostly in a few model mammals [82] and in Drosophilids [69,53].

[1] `http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/08-001/`

2.2 Homology Search

In a first automatic step we used a local installation of `NCBI blast` (v.2.2.10) with default parameters and $E < 10^{-6}$ to find candidate sequences in closely related genomes. If successful, the results of this search were aligned to the query sequence using `clustalw` (v.1.83). After a manual inspection using `clustalx`, the consensus sequence of the alignment was again used as a blast query with the same E-value cutoff.

If this automatic search was not successful, the best `blast` hit(s) were retrieved and aligned to a set of known snRNAs from related species. Candidate sequences were retained only when a visual inspection left no doubt that they were true homologs. This manual analysis step included a check whether the phylogenetic position of the candidate sequence in a neighborjoining tree was plausible, taking into account that the sequences are short and some parts of the alignments are of low quality.

In cases where no snRNA homologs were found as described above, we searched the genome again with a much less stringent cutoff of $E < 0.1$ (or even larger in a few cases) and extracted all short hits together with 200nt flanking sequence. We used Sean Eddy's `rnabob` with a manually constructed structure model to extract a structure-based match within the selected regions and attempted to align the candidate sequences manually to a structure-annotated alignment of snRNAs in the `emacs` editor using the `ralee` mode [22].

Finally, the resulting alignments of snRNAs where used to derive search patterns for `RNAmotif` [45] and `erpin` [19]. To this end, the consensus structure of the alignment was computed using `RNAalifold` [30] and converted into a form suitable as input for the two search programs.

2.3 Structure Models

Structure annotated sequence alignments were manually modified in the `emacs` text editor using the `ralee` mode [22] to improve local sequence-structure features based on secondary structure predictions for the individual sequences obtained from `RNAfold` [31]. Consensus structures were then computed using `RNAalifold` [30]. The structure models are compiled in the Electronic Supplement.

2.4 Upstream Region Analysis

With `MEME` (v.3.5.0) we discovered motifs upstream of the sequences for analysis of regulators and other possible depencies. They were manually compared with previously published sequence elements. We visually compared the `MEME` -patterns with the upstream elements in related species from

the following literature sources: [26] (general motifs), [14, 82, 2, 38] (human), [36, 5] (chicken), [53] (insects), [77] (*Bombyx mori*), [81] (*Strongylocentrotus purpuratus*), [84] (*Caenorhabditis elegans*).

2.5 Phylogenetic Analysis

Since the snRNA sequences are short and in addition there are several highly variable regions, we use split decomposition [1] and the neighbor net [7] algorithm (as implemented as part of the `SplitsTree4` package [33]) to construct phylogenetic networks rather than phylogenetic trees. The advantage of these method is that they are very conservative and that the reconstructed networks provide and easy-to-grasp representation of the considerable noise in the sequence data.

2.6 Synteny Information

In order to assess whether snRNA genes are mobile in the genome, we determined their flanking protein-coding genes. We used the `ensembl compara` annotation [17] to retrieve homolgous proteins in other genomes and compared whether these homologs also have adjacent snRNAs. For consistency, this analysis is performed based on `ensembl` (release 46) [32] using the data integration platform `BioFuice` [34]. More precisely, for each human snRNA G we examined that the relation of the left homologous $L_H(G)$ and right homologous $R_H(G)$ of flanking protein coding genes $L(G)$ and $R(G)$ on both sides of G. We only considered annotations in $L_H(G)$ and $R_H(G)$, resp., if the sequence distance between G_H and $L_H(G)$ and $R_H(G)$ was not more than twice (five times for mammals) the distance between G and $L(G)$ and $R(G)$.

3 Results

3.1 Homology Search

Tab. 1 summarizes the results of the sequence homology search. We find that, with few exceptions, `blast`-based homology search strategies are in general sufficient to find homologs of all nine spliceosomal snRNAs in most metazoa genomes. The procedure is hard to automatize, however, since in many cases the initial `blast` hits have poor E-values, while a multiple sequence alignment then leaves little doubt that a true homolog has been found. This is in particular true for searches bridging large evolutinary distances, in particular when the search extends beyond bilateria.

With very few exceptions we find multiple copies of all five major spliceosomal RNAs that exhibit the typical snRNA-like promoter elements and are hence mostly likely functional copies of the genes. The snRNA copy numbers vary substantially between different clades. The genus *Caenorhabditis*, for example, is set apart from other nematodes by a two to threefold increase in the number of major spliceosomal snRNAs. In contrast, the snRNAs of the minor spliceosome are in most cases single-copy genes.

Many genomes, most notably mammalian genomes, contain a sizeable number of major snRNA pseudogenes. Table 1 therefore lists only candidates that have plausible snRNA-like promoter structure, that fit the secondary structures of snRNAs in related species, and that exhibit strong sequence similarity in the unpaired regions of the molecule. These are rather restrictive criteria. In the Electronic Supplement, we therefore provide a corresponding table that is based only on sequence homology.

It is surprisingly difficult to compare the present snRNA survey with previous reports on vertebrate snRNAs. The main reason for discrepancies in the count of snRNAs is that distinguishing functional snRNAs from pseudogenes is still an unsolved problem. In this contribution, we use a very stringent criterion by insisting on a recognizable promoter structure. In some cases, however, it is known that snRNAs have internal promoters only [85]. These cases constitute false negatives in Tab. 1. On the other hand, much of the published literature considers sequence similarity to the known functional genes as the only criterion, thus most likely leading to the inclusion of a substantial fraction of pseudogenes. For instance, ref. [67] counts 16 U1, 6 U2 and 44 U6 snRNAs in the human genome (compared to our 8, 3, and 7, resp.), while [14] report 5-9 U6 snRNA genes, consistent with our list. Similarly, only a fraction of the major spliceosomal snRNAs reported for the chicken genome in [27] pass our promoter analysis.

For Drosophilids, on the other hand, our analysis is almost identical to the results of [53, Tab.1] and the data reported in [77]. Furthermore, we come close the results of a comparative genomics screen for non-coding RNAs in *C. elegans* [49], which reported 12 U1, 19 U2, 5 U4, 13 U5, and 23 U6, i.e., only a few more candidates than our present purely homology-based approach. A comparative screen of the two *Ciona* species for evolutionary conserved structured RNAs [48] missed a small number of snRNA genes that we indentified as most likely functional ones.

In a few species we failed to identify individual major spliceosomal snRNAs. Minor spliceosomal snRNAs are more often missing. In those cases where only some of the major or minor snRNAs remain undetected, the missing family member most likely escaped our detection procedure for one of several reasons:
(1) in the case of unassembled incomplete genomes for which only shotgun reads were searched, the snRNA may be located in the not yet sequenced fraction of the genome or it might not be completely contained within at least one single shotgun read.
(2) The snRNA in question may be highly derived in sequence. (For instance, the U11 snRNA in Drosophilids [69] cannot be found by be a simple `blast` search starting from non-insect sequences. It can be found however, by the combination of very un-specific blast and subsequent structure search as described in section 2.2.)

Table 1 Approximate copy number of snRNA genes.
We list here only those sequences that (1) are consistent with the secondary structures of related snRNAs, (2) show substantial sequence conservation in the unpaired regions of these structures, and (3) have recognizable promoter motifs. In some cases none of the candidates satisfies all these criteria. If there are nevertheles clear homologuous sequences. Entries of the form $^{S}0$ and $^{P}0$ indicate that there is homolgous sequence which however lacks structural similarity or recognizable promoter elements. The quality of the genome assembly is marked by the following sysmbols: △ – Traces, □ – Contigs, ♢ – Scaffolds, ♠ – Chromosoms.

	Coverage	Species	U1	U2	U4	U5	U6	U11	U12	U4atac	U6atac
♢		*M. brevicollis*	0	0	0-1	0-2	1	0	0	0	0
△		*Reniera sp*	2	0-1	2	3	2	1	1	0	3
♢		*Trichoplax adhaerens*	1	1	1	1	2	1	1	1	1
♢		*N. vectensis*	2	2	4	5	3	3	3	1	2
△	7.45-8.33X	*H. magnipapillata*	4	2	5	7	4	1	1	0	2
△	0.05X	*A. millepora*	0	2	0	2	2	0	0	0	0
△	0.047X	*A. palmata*	1	0	0	0	1	0	0	0	0
♢		*S. mansoni*	3	3	1	2	9	0	1	0	0
□		*S. mediterannea*	2	$^{P}0$	3	2	2	0	0	0	0
△	13.03X	*L. gigantea*	3	8	11	2	7	2	1	0	2
△	0.05X	*B. glabrata*	$^{S}0$	2	0	1	$^{S}0$	0	0	0	0
△	0.54X	*P. lobata*	1	1	1	0	0	0	0	0	0
△	0.012X	*E. scolopes*	$^{SP}0$	1	0	0	0	0	0	0	0
△	4.48X	*A. californica*	4	2	4	10	8	1	1	0	1
♢		*C. capitata*	5	2	1	4	2	1	1	1	1
♢		*H. robusta*	6	8	4	7	4	0	1	1	1
△	0.23X	*H. bacteriophora*	2	2	0	2	1	0	0	0	0
△	11.33X	*B. malayi*	3	3	1	1	2	1	0	0	—
△	12.15X	*T. spiralis*	1	5	2	3	1	1	0	0	0
△	11.24X	*P. pacificus*	2	2	4	4	7	1	0	0	0
□		*C. brenneri*	19	19	10	19	25	0	0	0	0
□		*C. remanei*	14	11	5	13	15	0	0	0	0
△	10.18X	*C. japonica*	16	15	4	14	7	0	0	0	0
♠		*C. elegans*	10	17	4	9	15	0	0	0	0
♠		*C. briggsae*	9	10	4	10	22	0	0	0	0
△	3.29X	*D. pulex*	5	6	4	9	8	1	1	$^{PS}0$	1
△	11.81X	*P. humanus*	3	4	1	2	1	1	1	0	1
□		*N. vitripennis*	7	4	3	5	5	1	2	1	2
△	2.58X	*I. scapularis*	4	4	3	4	3	0	1	0	1
△	1.6X	*A. pisum*	2	3	0	2	3	1	1	0	1
♢		*A. mellifera*	5	3	2	3	3	1	1	1	1
♢		*B. mori*	5	6	3	5	4	1	1	1	2
△	0.75X	*T. castaneum*	5	5	2	6	3	1	1	0	1
♠		*A. gambiae*	7	7	2	5	2	2	1	1	1
♠		*D. melanogaster*	5	6	3	7	3	1	1	1	1
♠		*D. ananassae*	9	8	2	4	2	1	1	1	1
♠		*D. erecta*	8	9	3	7	4	1	1	1	1
♠		*D. grimshawi*	7	6	3	7	3	1	1	1	2
♠		*D. mojavensis*	6	8	3	6	3	1	1	1	1
♠		*D. persimilis*	7	7	3	7	3	1	1	1	1
♠		*D. pseudoobscura*	7	7	3	6	3	1	1	1	1
♠		*D. sechellia*	7	6	3	7	3	1	1	1	1
♠		*D. simulans*	8	6	3	8	3	1	1	0	1
♠		*D. virilis*	6	8	3	6	2	1	1	2	1
♠		*D. willistoni*	8	9	3	8	$^{P}0$	1	1	1	0
♠		*D. yakuba*	8	7	3	8	3	1	1	1	1

Coverage	Species	U1	U2	U4	U5	U6	U11	U12	U4atac	U6atac
♢	*S. purpuratus*	5	7	9	8	3	2	3	1	1
△ 3.77X	*S. kowalevski*	7	4	4	5	4	1	2	0	3
♢	*C. savignyi*	3	2	3	7	2	1	1	1	1
♢	*C. instestinalis*	1	1	3	5	2	1	1	1	1
△ 7.8X	*O. dioica*	1	6	2	7	4	0	0	0	0
♢	*B. floridae*	8	3	5	9	4	1	1	0	1
△ 6.19X	*P. marinus*	6	5	8	9	5	1	2	$^{PS}0$	3
♠	*D. rerio*	5	4	4	7	3	1	1	1	1
♠	*O. latipes*	4	2	2	4	4	1	1	1	1
♠	*G. aculeatus*	6	2	4	7	3	1	1	1	1
♢	*F. rubripes*	5	5	3	6	4	1	1	1	1
♠	*T. nigroviridis*	4	5	3	5	2	1	1	0	1
♢	*X. tropicalis*	5	1	3	2	5	1	1	1	2
♠	*G. gallus*	1	1	1	2	4	1	1	1	1
△ 8.34X	*T. guttata*	2	5	2	3	2	1	1	0	1
△ 8.24X	*A. carolinensis*	14	6	2	6	5	1	2	1	1
♠	*O. anatinus*	5	2	2	4	6	1	1	1	1
♠	*M. domestica*	7	4	2	5	6	1	$^{PS}0$	1	1
♠	*M. musculus*	7	5	1	6	7	1	2	1	2
♠	*R. norvegicus*	4	10	1	4	5	4	1	1	1
♠	*C. familiaris*	6	5	2	4	5	1	1	1	1
♠	*B. taurus*	7	8	2	5	6	2	1	1	1
♠	*P. tropicalis*	7	2	2	7	8	1	1	3	1
♠	*H. sapiens*	8	3	2	5	7	1	1	3	1

(3) In some cases we list a "0" in Tab. 1 even though there is recognizable sequence homology in the genome. In these cases we were not able to identify the snRNA-like promoter elements and/or the secondary does not fit the expectation. These cases marked in the table.
(4) It is conceivable that some species have lost a particular snRNA and replaced it by corresponding snRNA from the other spliceosome. The observation that U4 may function in both the major and minor spliceosomes [74] shows that such a replacement mechnism might indeed be evolutionarily feasible.

In our data set, we most frequently were unable to find a U4atac homolog. We cannot know, of course, whether we missed these cases due to poor sequence conservation or due to loss of the gene. For instance, we did not recover a plausible U4atac candidate for the hemichordate *Saccoglossus kowaleski* despite the fact that the U4atac sequence of the sea urchin *Strongylocentrotus purpuratus* was easily retrieved.

Surprisingly, we found neither a canonical U6 nor a canonical U6atac in *Drosophila willistoni*. A highly derived U6 homolog has no recognizable snRNA-like promoter structure and exhibits substantial deviations from the consensus structure, see section 3.5. Similarly, the U4atac candidate from *Daphnia pulex* deviates substantially from other arthropod sequences. It is possible that in some or all of these cases the snRNA is present in the genome but is not contained in the currently available genomic sequence data. This is most likely the case for the missing minor spliceosomal snRNAs of *Ixodes scapularis*, *Pediculus humanus*, or *Drosophilia willistoni*.

In some cases, however, we failed to identify all four minor spliceosomal snRNAs. Consistent with previous work [61] we found no convincing homologs of the minor spliceosomal snRNAs U11, U12, U4atac, or U6atac in any of the nematode genomes, suggesting that the minor spliceosome was lost early in the nematode lineage. Nevertheless, we find some `blast` hits for minor spliceosomal snRNAs in some nematode genomes.

Our analysis furthermore suggests the possible loss of the minor spliceosome in *Oikopleura dioica*, while a complete complement of minor spliceosomal snRNAs was found in the genus *Ciona*. It is unclear, however, whether this is an artifact due to limiations of available shotgun traces.

Our survey provides evidence that most metazoan clades for which genomic sequences are available have retained the minor spliceosome. For many groups, such as Annelida or Cnidaria, we are not aware of earlier references to the existence of minor spliceosome.

3.2 Specific Upstream Elements

The classical snRNA-specific PSE and TATA elements that have been described in detail for several vertebrates [26, 14] are highly conserved. This appers to be an exception rather than the rule, however: the snRNA upstream elements are highly diverse across metazoa. Our analysis agrees with the recent observation that in Drosophilids there is a rapid turnover in the upstream sequences. Even though the PSE is fairly well-conserved within Drosophilids, it already differs substantially between the major insect groups [53]. Sim-

ilarly, within the nematodes conservation of upstream elements is limited to the genus level. In general, the PSE of U11, U12 and U4atac is much less conserved than their counterpart in major spliceosomal snRNA genes. For the purpose of this study, the relatively well-conserved elements were used to discriminate functional snRNAs from likely pseudogenes. We concentrated on PSE and TATA elements for this purpose because other snRNA-associated upstream elements, such as SPH, OCT, CAAT-box, GC-box, -35-element and *Inr* are even less well conserved:

A GC-box was identified in *Caenorhabditis* at a non-canonical position (about -68nt). These elements are different for each single snNRA class: U1 `GGACGG` (44/52 sites), U2 `TGGCCG` (38/60 sites) and for U5 `CGGCCG` (39/46 sites). However, also among a single snRNA this element varies a lot: insects have a U1 GC-box `GCGCTG` at about -75nt (15/39 sites). About half of the U6 sequences of basal deuterostomes show the CAAT-box motif `TGCCAAGAA` at the known position of -70nt. Interestingly, we find related motifs in the upstream region of Drosophilids U11 (`GACCAATAT`, -33nt) and other insects U5 snRNA (`TTCCAATCA`, -28nt) and . The Octamer motif (OCT, `ATTTGCAC`) was found in 6 of 7 sequences of basal deuterostomes at the known position of -54nt upstream of U6atac. However, in 12 of 14 Drosophilids sequences, the closely related motif `ATTTGCTT` was found at position -33nt. About 35nt upstream of U11 and U12 snRNAs of teleosts we found the motif `GTGACA` and `TGCACA`, respectively. The *Inr* element of U1 snRNA was found in each species. For teleost fishes and Drosophilids we found a complete set of this element for all snRNAs. However, the element show substantial sequence variations both between different genes in the same species and between homologous genes in different species. We refer to the Electronic Supplement for further details and lists of identified sequence elements.

3.3 Clusters of snRNA genes

In Mammalia, we observe linkeage of tandem copies of U2 snRNAs, see also [41,62], while there there are no clusters of distinct snRNAs. In *Drosophila*, there are surprisingly constant patterns of snRNA clusters: (a) U2-U5 clusters are observed 4-6 times per genome, (b) there are one or two U1-U2 clusters, and (c) 3-9 tandem copies of snRNAs. Two species deviated therefrom. In *D. ananassae*, we find no U2-U5 cluster, but instead 7 U1-U2, one U4-U5 cluster and 4 other tandem copies, while the *D. willistoni* lacks the U4-U5 cluster but contains 10 U2-U5 pairs and 6 tandem copies. Teleost fishes also have a common pattern: there are one or two U1-U2 pairs and 2-6 tandem copies. In general, however, snRNA do not appear in clusters throughout metazoan genomes.

In several species, linkeage of snRNAs with 5S rRNA has been observed [42,40,16,63,11,46]. We found only one further example of this type: in *Daphnia pulex* 5S and U5 snRNA are separated by only 308bp.

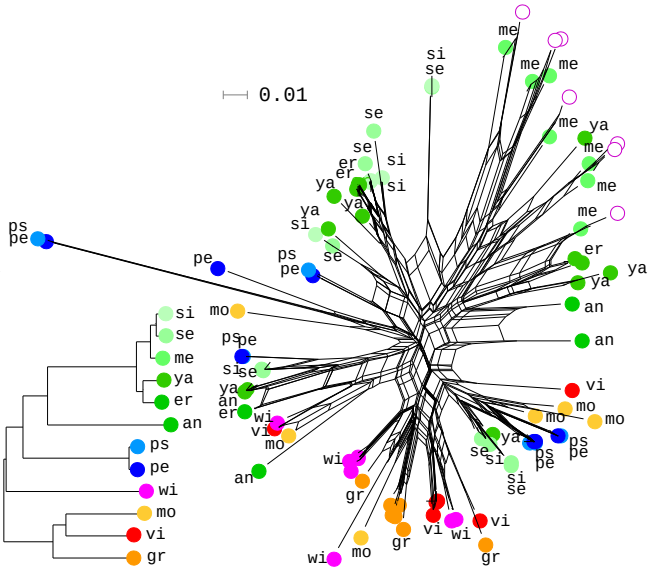


Fig. 1 Phylogenetic network of Drosophilid U5 snRNAs. The eight U5 snRNA reported by [8] are shown by white dots. me – *D. melanogaster*, er – *D. erecta*, si – *D. simulans*, se – *D. sechellia*, ya – *D. yakuba*, wi – *D. willistoni*, gr – *D. grimshawi*, mo – *D. mojavensis*, vi – *D. virilis*, pe – *D. persimilis*, ps – *D. pseudoobscura*, an – *D. ananassae*. The phylogenetic tree is adapted from ref. [15].

3.4 Phylogenetic Analysis and Paralogs

Like ribosomal RNAs, spliceosomal RNAs are subject to *concerted evolution* [28,68,21], i.e., one observes that paralogous sequences in the same species are more similar than orthologous sequences of different species. Multiple molecular mechanisms may account for this phenomenon: gene conversion, repeated unequal crossover, and gene amplification (frequent duplications and losses within family), see [40] for a review. In some cases, however, paralogs can escape from the concerted evolution mechanisms as exemplified by the two paralog groups of SSU rRNA in Chaetogatha [60].

Distinguishable snRNA paralogs that are often differentially expressed have previously been reported for a diverse collection of major spliceosmal snRNAs including U1 snRNAs in insects [43,64,77], Xenopus [12], and human [39], U2 snRNAs in *Dictyostelium* [29], sea urchin [80] and silk moth [77], U5 snRNAs in human [79], sea urchin [52], and Drosophilids [8], U6 snRNAs in silk moth [78] and human [85,14].

A phylogenetic analysis of the individual snRNA families nevertheless does not show widely separated paralog groups that are stable throughout larger clades. Fig. 1, for example shows that the U5 variants described by [8] do not form clear paralog groups beyond the closest relatives of *Drosophila melanogaster*. On the other hand, there is some evidence for distinguishable paralogs outside the melanogaster subgroup. The situation is much clearer for the Drosophilid U4 snRNAs, where three paralog groups can be distinguished, see Fig. 2. One group is well separated from the other two

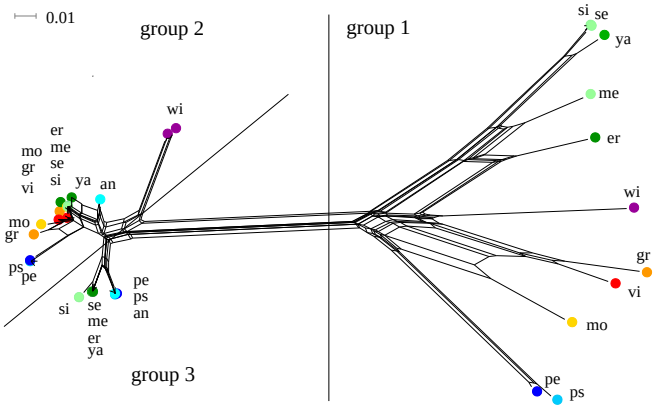


Fig. 2 Phylogenetic tree of insect U4 snRNAs. In this case we can distinguish three paralog groups within the Drosophilids. me – *D. melanogaster*, er – *D. erecta*, si – *D. simulans*, se – *D. sechellia*, ya – *D. yakuba*, wi – *D. willistoni*, gr – *D. grimshawi*, mo – *D. mojavensis*, vi – *D. virilis*, pe – *D. persimilis*, ps – *D. pseudoobscura*, an – *D. ananassae*.

Table 2 Paralog groups of major splicesomal snRNAs recognizable within major animal clades. The symbol •denotes clearly distinguishable paralog groups and refers to the supplemental material for details, ? indicates ambigous cases, = means that all paralogous genes have identical sequences.

Clade	U1	U2	U4	U5	U6
Annelids	–	–	–	–	=
Nematods	–	–	–	–	=
Caenorhabditis	–	–	–	•	=
Insects	–	–	–	–	=
Drosophilids	?	–	Fig.2	[8]	=
Teleosts	–	Fig.3a	Fig.3b	Fig.3c	–
Tetrapoda	–	–	–	–	–
Mammalia	–	–	–	•	–

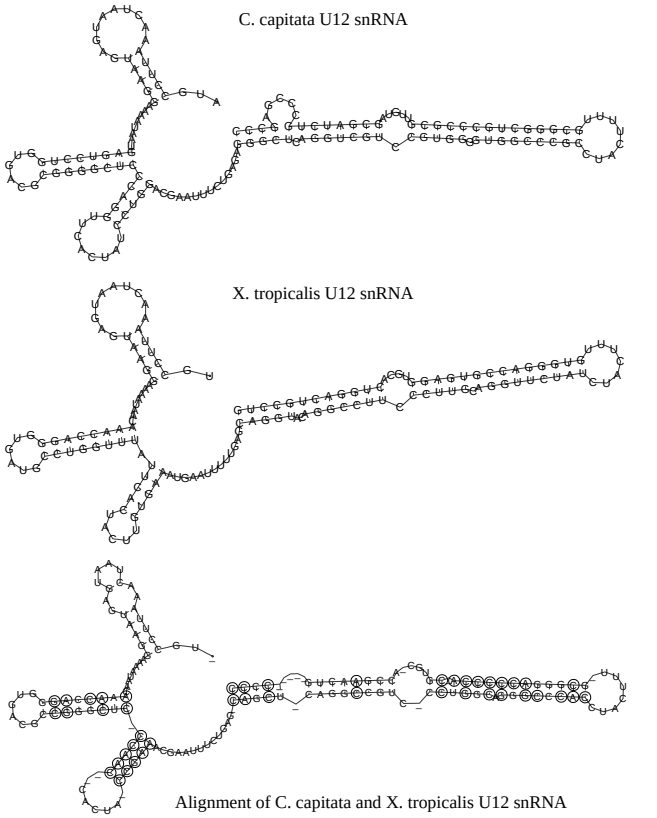


Fig. 4 Predicted secondary structures of *Capitella capitata*, *Xenopus tropicalis* and an alignment created with `RNAalifold` of both. Circles represent different bases and therewith compensatory mutations.

and internally rather diverse. The other two groups are very clear distinguishable for the melanogaster and obscura group (see [15]). For *D. virilis*, *D. mojavensis*, *D. grimshawi* and *D. willistoni* we have two nearly identical copies instead of two different groups of genes.

Table 2 summarized the presence of recognizable paralog groups within major animal groups. Within the genus *Caenorhabitis* we find evidence for the formation of U5 paralog groups in *C. remanei*, *C. brenneri*, and *C. briggsae* to the exclusion of *C. elegans* and *C. japonica*. Evidence for paralog groups of U1 snRNA in Drosophilids remains ambiguous due to the small sequence differences.

In teleost fishes we find clearly recognizable paralog groups for U2, U4, and U5 snRNAs. Surprisingly, the medaka *Oryzias latipes* has only a single group of closely related sequences, despite the fact that for U4, the split of the paralogs appear to predate the last comman ancestor of zebrafish and fugu, Fig. 3.

Neither the two rounds of genome duplications at the root of the vertebrates nor the teleost-specific genome duplication has lead to recognizable paralog groups of snRNAs. In particular, minor snRNA genes are single-copy genes in teleosts.

3.5 Secondary Structures

The spliceosomal snRNAs have evolutionarily well-conserved secondary structures [73]. These structures have received substantial interest in the past, as explified by the following non-exhaustive list of references covering a diverse set of animal species: *Homo sapiens* U1 [54], U2 [25], U4 [37], U5 [6,79], U6 [25], U11 [66,51,82], U12 [66,51,82] and U4atac [72], *Rattus norvegicus* U1 [37], U4 [37], U5 [37], *Gallus gallus* U4 [37], U5 [6], *Xenopus laevis* U1 [18], U2 [47], *Caenorhabditis elegans* U1, U2, U5, U4/U6 [84], *Drosophila melanogaster* U1 [54,56], U2 [56], U4 [56], U5 [56], U4atac/U6atac, U6atac/U12 [59], *Bombyx mori* U1 [76], U2 [75], *Asselus aquaticus* U1 [3], *Ascaris lumbricoides* U1, U2, U5, U4/U6 [70]. Large changes in snRNA structures over evolutionary time were recently reported for hemiascomycetous yeasts [50]. The comprehensive survey of snRNA sequences throughout metazoa set the stage for a comparably detailed analysis of metazoan snRNA structures. In order to asses structural variations, we contructed structure annotated sequence alignments of all snRNA families. These are provided as part of the electronic supplement.

In general we find that snRNA sequences vary more in paired regions than in the loops. The sequence variations almost exclusively comprises compensatory mutations that

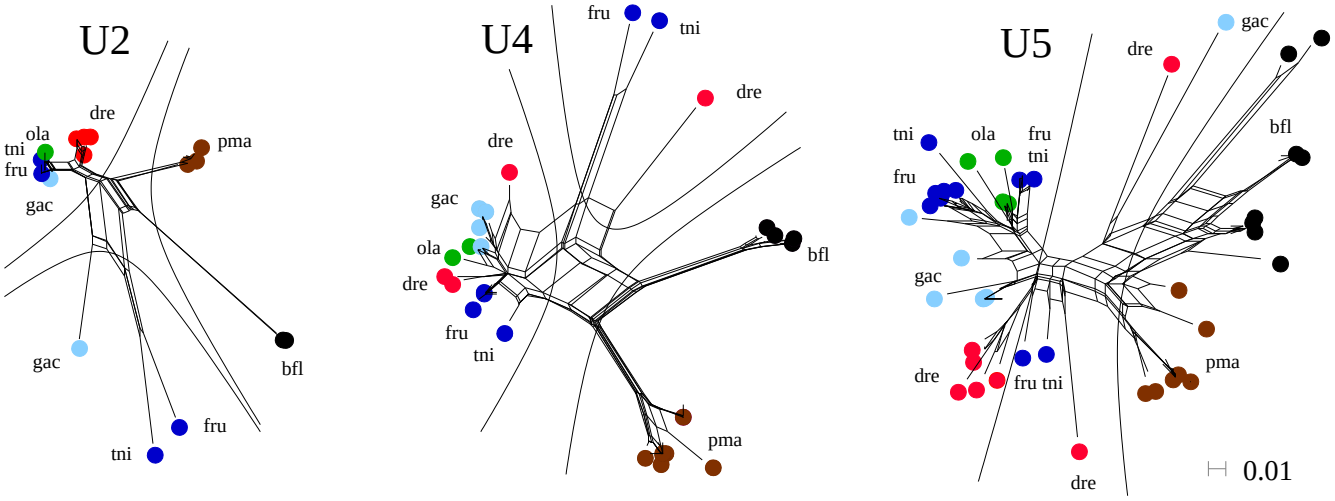


Fig. 3 Phylogenetic networks of teleost fish snRNAs. Species abbreviations: fru – *Fugu rubripes*, tni – *Tetraodon nigrovidis*, gac – *Gasterosteus aculeatus*, ola – *Oryzias latipes*, dre – *Danio rerio*, pma – *Petromyzon marinus*, bfl – *Branchiostoma floridae*.

leave the secondary structures intact. As an example, Fig. 4 shows the structures of the U12 snRNA of *Xenopus tropicalis* and *Capitella capitata*. The sequences have few paired nucleotides in common.

Structural variations are typically limited. In Fig. 5 we use the U1 snRNAs as a typical example for the evolutionary variation of snRNAs across the metazoa. Overall the structures are extremely well conserved with small variations in the length of the individual stems. With several notable exceptions this is true for all metazoan snRNAs.

As reported previously [8], the second stem of U5 snRNA shows some variations. More interestingly, the minor spliceosomal snRNAs tend to be derived in insects. This has been reported previously in particular for U11 in Drosophilids [69, 53]. We found substantial structural variations also for drosophilid U12 snRNAs: there are massive insertions in and after Stem III, while Stem I and II show mispairings. Furthermore, Stem II of U6atac is completely deleted in all examined insects. Details are compiled in the electronic supplement.

Most surprisingly, *Acyrthosiphon pisum* exhibits highly derived structures for all four minor spliceosomal snRNAs, Fig. 6.

The U2 snRNA of *Schmidtea mediterannea* does fit well to the structural alignment of the other U2 snRNAs. In *Schistosoma mansoni* we found a canonical U12 snRNA, while the sequences of the candidates for minor spliceosomal snRNAs do not fit well to the consensus secondary structure models. Details can be found in the Electronic Supplement.

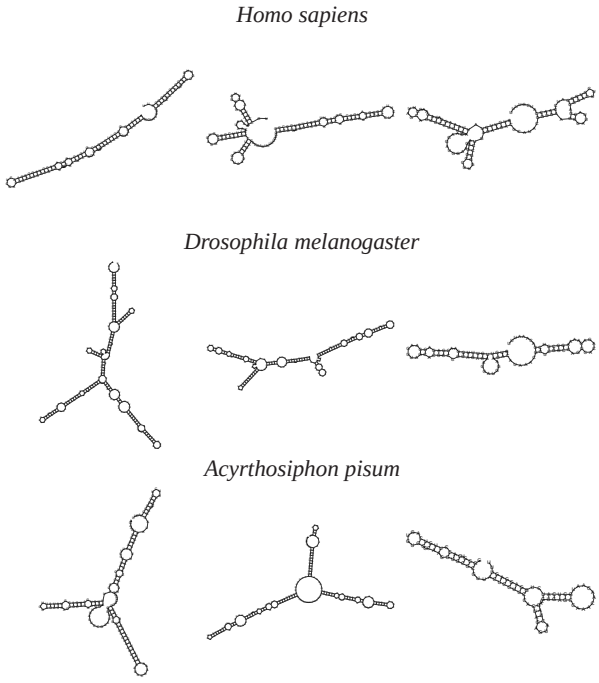


Fig. 6 Secondary structures of U11 (left), U12 (center), U6atac (right) in *Acyrthosiphon pisum*, *Drosophila melanogaster* and *Homo sapiens*. Drosophilids derived far from all other minor spliceosome structures (e.g. human). Moreover, *Acyrthosiphon pisum* built an autonomous structure group for all minor snRNAs.

3.6 Syntenic Conservation

In order to assess the conservation of the genomic positions of the snRNAs we retrieved the protein coding genes adjacent to the 31 human snRNAs (8 U1, 3 U2, 2 U4, 5 U5, 7 U6, 1 U11, 1 U12, 3 U4atac and 1 U6atac) and compared the position of their homologs in 14 vertebrate genomes (teleosts, frog, chicken, platypus, opossum, rodents, cow, dog, and chimp) with the 234 snRNA genes that were found in these genomes. We found syntenic conservation of snRNA and flanking genes in only 36 cases, of which 20 belong to the human-chimp comparison and 9 pairs are conserved between

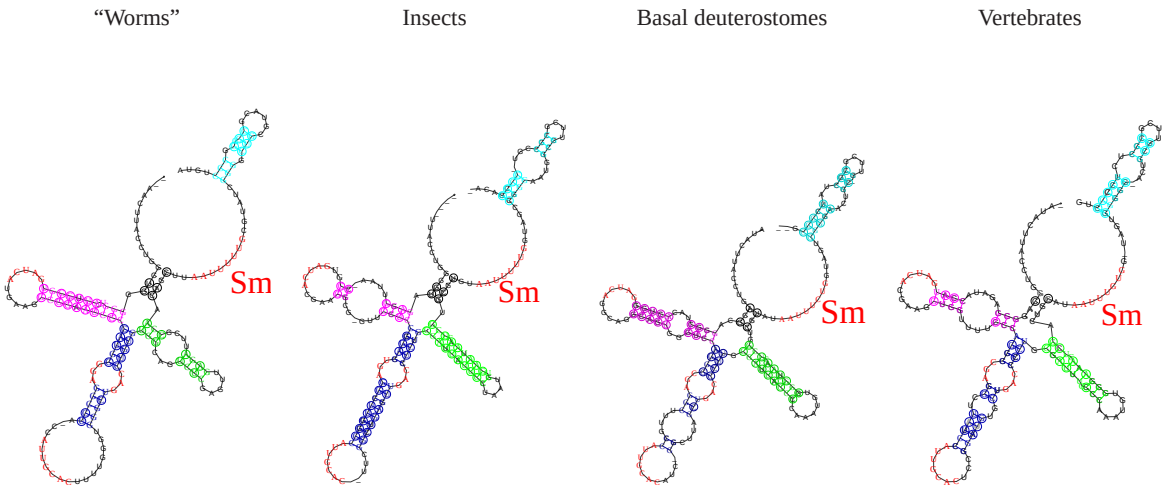


Fig. 5 Secondary structure prediction of U1 snRNA, folded by `RNAalifold`. From left to right: protostomia without insects, insects, deuterostomes without vertebrates, vertebrates. Red: Conserved sequences in all organisms, which possibly bind to proteins. Sm binding site marked separately.

human and mouse. Only a single pair is conserved between human and opossum and no syntenic conservation can be traced back further in evolutionary history. Including the pseudogenes increases the numbers of conserved pairs to 499 of 1609. Again most of these (453) are human/chimp pairs. The data clearly show that snRNA locations are not syntenically conserved, i.e., snRNA behave like mobile elements in their genomic context.

3.7 Pseudogenes

As mentioned above, snRNAs are frequently the founders of families of pseudogenes. This is a property that they share with most other small RNA classes such as 7SL RNA, Y RNA, tRNAs etc. Such families of pseudogenes are easily recognized as a by-product of `blast`-based homology searches as a large set of hits with intermediate *E*-values. Fig. 7 summarizes such data, more details are provided in the Electronic Supplement.

Spliceosomal snRNA pseudogenes families are very unevenly distributed across distinct phylogenetic groups and have clearly arisen in independent burst multiple times across animal evolution. Within deuterostomes, almost all sequenced genomes, whith the notable exception of teleosts and chicken, contain at least one large family of snRNA-derived pseudogenes.

The genus *Caenorhabditis* shows no pseudogenes, whereas other nematods show nearly such a high number of pseudogenes as primates. Annelids, molluscs and plathelminths behave similarly. The *Trichoplax adhaerens* genome, on the other hand, contains a single copy of each of the nine spliceosomal snRNAs.

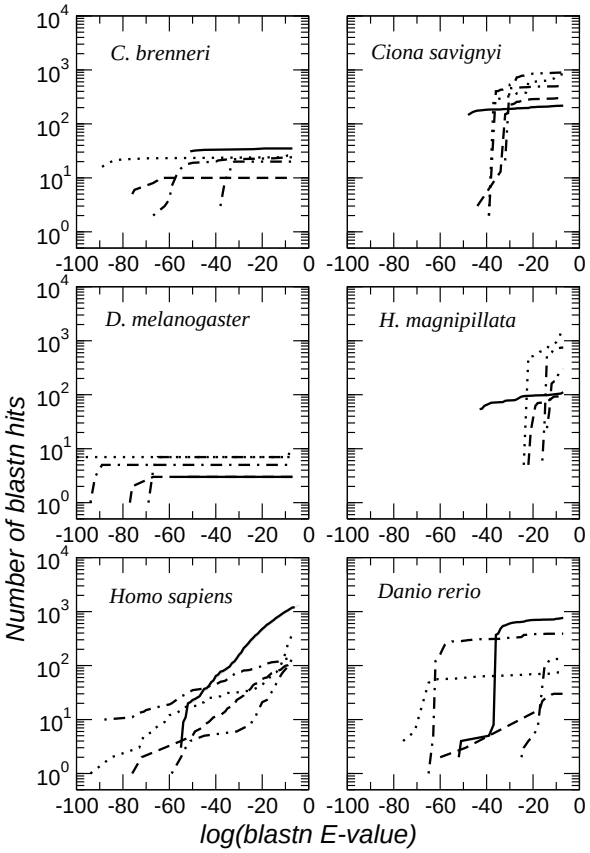


Fig. 7 Double-logarithmic plot of the number of blast hits versus cutoff *E*-value for 6 different genomes. Pseudogene families appear as a slowly increasing curve, while genes without a "cloud" of pseudogene have a flat distribution for $E < 10^{-5}$. Dashdotted line – U1; dotted line –U2; dashed line – U4; dashdotdotted line – U5; continuous line – U6.

4 Discussion

We have reported here on a comprehensive computational survey of spliceosomal snRNA in all currently available metazoan genomes. We thus provide a comparable and nearly complete collection of animal snRNA sequences. The dense taxon sampling allowed us to verify homology of candidate sequences. Both the major and the minor spliceosome are present in almost all metazoan clades, nematodes (and possibly *Oikopleura*) being the only notable exception. For many of the metazoan families we report here the first evidence on their spliceosomal RNAs.

Using restrictive filtering of the candidates by both secondary structure and canonical promoter structure leaves us with a high-quality data set that was then used to construct secondary structure models. This is useful in particular for the snRNAs of the minor spliceosome for which very few sequences are reported in databases; indeed, the `Rfam 7.0` [23] lists only the U11 and U12 families with a meager set of seed sequences from few model organisms. The sequence and secondary structure data compiled in this study provide a substantially improved databasis and set the stage for systematic searches of even more distant homologs.

The analysis of the genomic distribution of snRNAs reveals that discernible paralogs are not uncommon within genera or families. However, no dramatically different paralogs have been found. Spliceosomal snRNAs are prone to spawning large pseudogene families, which arose independently in many species. They behave like mobile genetic elements in that they barely appear in syntenic positions as measured by their flanking genes. While in some genomes snRNAs appear in tandem and/or associated with with 5S rRNA genes, these clusters are not conserved over longer evolutionary time-scales. Taken together, the data are consistent with a dominating duplication-deletion mechanism of concerted evolution for the genomic evolution and proliferation of snRNA.

Acknowledgments

This work was supported in part by the *Graduierten-Kolleg Wissensrepräsentation* and by the Bioinformatics Initiative of the *Deutsche Forschungs-Gemeinschaft* (DFG). Special thanks to Petra Pregel and Jens Steuck for making work much easier.

References

1. Bandelt, H.J., Dress, A.W.M.: A canonical decomposition theory for metrics on a finite set. Adv. Math. **92**, 47 (1992)
2. Bark, C., Weller, P., Zabielski, J., Pettersson, U.: Genes for human U4 small nuclear RNA. Gene **50**, 333–344 (1986)
3. Barzotti, R., Pelliccia, F., Rocchi, A.: Identification and characterization of U1 small nuclear RNA genes from two crustacean isopod species. Chromosome Res **11**, 365–373 (2003)
4. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L.: `GenBank`. Nucleic Acids Res. **35**, D21–D25 (2007)
5. Bhathal, H.S., Zamrod, Z., Tobaru, T., Stumph, W.E.: Identification of proximal sequence element nucleotides contributing to the differential expression of variant U4 small nuclear RNA genes. J. Biol. Chem. **270**, 27,629–27,633 (1995)
6. Branlant, C., Krol, A., Lazar, E., Haendler, B., Jacob, M., Galego-Dias, L., Pousada, C.: High evolutionary conservation of the secondary structure and of certain nucleotide sequences of U5 RNA. Nucleic Acids Res **11**, 8359–8367 (1983)
7. Bryant, D., Moulton, V.: Neighbor-net: An agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. **21**, 255–265 (2004)
8. Chen, L., Lullo, D.J., Ma, E., Celniker, S.E., Rio, D.C., Doudna, J.A.: Identification and analysis of U5 snRNA variants in Drosophila. RNA **11**, 1473–1477 (2005)
9. Collins, L., Penny, D.: Complex spliceosomal organization ancestral to extant eukaryotes. Mol. Biol. Evol. **22**, 1053–1066 (2005)
10. Collins, L.J., Macke, T.J., Penny, D.: Searching for ncRNAs in eukaryotic genomes: maximizing biological input with `RNAmotif`. J. Integ. Bioinf. **1**, 2004–08–04 (2004). URL `http://journal.imbio.de/index.php?paper_id56`
11. Cross, I., Rebordinos, L.: 5S rDNA and U2 snRNA are linked in the genome of *Crassostrea angulata* and *Crassostrea gigas* oysters: does the $(ct)_n.(ga)_n$ microsatellite stabilize this novel linkage of large tandem arrays? Genome **48**, 1116–1119 (2005)
12. Dahlberg, J.E., Lund, E.: The genes and transcription of the major small nuclear RNAs. In: M.L. Birnstiel (ed.) Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles, pp. 38–70. Springer-Verlag, Berlin (1988)
13. Denison, R.A., Van Arsdell, S.W., Bernstein, L.B., Weiner, A.M.: Abundant pseudogenes for small nuclear RNAs are dispersed in the human genome. Proc. Natl. Acad. Sci. USA **78**, 810–814 (1981)
14. Domitrovich, A.M., Kunkel, G.R.: Multiple, dispersed human U6 small nuclear RNA genes with varied transcriptional efficiencies. Nucleic Acids Res. **31**, 2344–2352 (2003)
15. Drosophila 12 Genomes Consortium: Evolution of genes and genomes on the Drosophila phylogeny. Nature **450**, 203–218 (2007)
16. Ebel, C., Frantz, C., Paulus, F., Imbault, P.: Trans-splicing and cis-splicing in the colourless euglenoid, *Entosiphon sulcatum*. Curr Genet **35**, 542–550 (1999)
17. Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S.C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Gräf, S., Haider, S., Hammond, M., Holland, R., Howe, K.L., Howe, K., Johnson, N., Jenkinson, A., Kähäri, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A.J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Hubbard, T.J.P., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., Searle, S.: Ensembl 2008. Nucleic Acids Res. **36**, D707–D714 (2008)
18. Forbes, D.J., Kirschner, M.W., Caput, D., Dahlberg, J.E., Lund, E.: Differential expression of multiple U1 small nuclear RNAs in oocytes and embryos of *Xenopus laevis*. Cell **38**, 681–689 (1984)
19. Gautheret, D., Lambert, A.: Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. J. Mol. Biol. **313**, 1003–1011 (2001)
20. Giribet, G., Edgecombe, G.D., Wheeler, W.C.: Arthropod phylogeny based on eight molecular loci and morphology. Nature **413**, 157–161 (2001)
21. Gonzalez, I.L., Sylvester, J.E.: Human rDNA: Evolutionary patterns within the genes and tandem arrays derived from multiple chromosomes. Genomics **73**, 255–263 (2001)
22. Griffiths-Jones, S.: `RALEE`—rna alignment editor in `Emacs`. Bioinformatics **21**, 257–259 (2005)
23. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., Bateman, A.: `Rfam`: annotating non-coding RNAs in complete genomes. Nucleic Acids Res **33**, D121–D124 (2005)

24. Hastings, K.E.: SL trans-splicing: easy come or easy go? Trends Genet. **21**, 240–247 (2005)
25. Hausner, T.P., Giglio, L.M., Weiner, A.M.: Evidence for base-pairing between mammalian U2 and U6 small nuclear ribonucleoprotein particles. Genes Dev **4**, 2146–2156 (1990)
26. Hernandez, N.: Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. J. Biol. Chem. **276**, 26,733–26,736 (2001)
27. Hillier, L.W., Miller, W., Birney, E., *** authors: Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432**, 695–716 (2004)
28. Hillis, D.M., Dixon, M.T.: Ribosomal DNA: molecular evolution and phylogenetic inference. Q. Rev. Biol. **66**, 411–453 (1991)
29. Hinas, A., Larsson, P., Avesson, L., Kirsebom, L.A., Virtanen, A., Söderbom, F.: Identification of the major spliceosomal RNAs in *Dictyostelium discoideum* reveals developmentally regulated U2 variants and polyadenylated snRNAs. Eukaryotic Cell **5**, 924–934 (2006)
30. Hofacker, I.L., Fekete, M., Stadler, P.F.: Secondary structure prediction for aligned RNA sequences. J. Mol. Biol. **319**, 1059–1066 (2002)
31. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., Schuster, P.: Fast folding and comparison of RNA secondary structures. Monatsh. Chem. **125**, 167–188 (1994)
32. Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X.M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., Birney, E.: `Ensembl` 2005. Nucleic Acids Res. **33**, D447–D453 (2005)
33. Huson, D.H., Bryant, D.: Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. **23**, 254–267 (2006)
34. Kirsten, T., Rahm, E.: `BioFuice`: Mapping-based data intergation in bioinformatics. In: U. Leser, F. Naumann, B. Eckman (eds.) Proceedings of the 3rd International Workshop on Data Integration in the Life Sciences (DILS), vol. 4075, pp. 124–135. Springer Verlag, Berlin, Heidelberg (2006)
35. König, H., Matter, N., Bader, R., Thiele, W., Müller, F.: Splicing segregation: the minor spliceosome acts outside the nucleus and controls cell proliferation. Cell **131**, 718–729 (2007)
36. Korf, G.M., Stumph, W.E.: Chicken U2 and U1 RNA genes are found in very different genomic environments but have similar promoter structures. Biochemistry **25**, 2041–2047 (1986)
37. Krol, A., Branlant, C., Lazar, E., Gallinaro, H., Jacob, M.: Primary and secondary structures of chicken, rat and man nuclear U4 RNAs. Homologies with U1 and U5 RNAs. Nucleic Acids Res **9**, 2699–2716 (1981)
38. Kunkel, G.R., Pederson, T.: Upstream elements required for efficient transcription of a human U6 RNA gene resemble those of U1 and U2 genes even though a different polymerase is used. Genes Dev **2**, 196–204 (1988)
39. Kyriakopoulou, C., Larsson, P., Liu, L., Schuster, J., Söderbom, F., Kirsebom, L.A., Virtanen, A.: U1-like snRNAs lacking complementarity to canonical 5' splice sites. RNA **12**, 1603–1611 (2006)
40. Liao, D.: Concerted evolution: molecular mechanism and biological implications. Am J Hum Genet **64**, 24–30 (1999)
41. Liao, D., Pavelitz, T., Kidd, J.R., Kidd, K.K., Weiner, A.M.: Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the RNU2 locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene conversion. EMBO J. **16**, 588–598 (1997)
42. Liao, D., Weiner, A.M.: Concerted evolution of the tandemly repeated genes encoding primate U2 small nuclear RNA (the RNU2 locus) does not prevent rapid diversification of the $(CT)_n.(GA)_n$ microsatellite embedded within the U2 repeat unit. Genomics **30**, 583–593 (1995)
43. Lo, P.C., Mount, S.M.: *Drosophila melanogaster* genes for U1 snRNA variants and their expression during development. Nucleic Acids Res **18**, 6971–6979 (1990)
44. Lorković, Z.J., Lehner, R., Forstner, C., Barta, A.: Evolutionary conservation of minor U12-type spliceosome between plants and humans. RNA **11**, 1095–1107 (2005)
45. Macke, T.J., Ecker, D.J., Gutell, R.R., Gautheret, D., Case, D.A., Sampath, R.: RNAMotif, an RNA secondary structure definition and search algorithm. Nucl. Acids Res. **29**(22), 4724–4735 (2001)
46. Manchado, M., Zuasti, E., Cross, I., Merlo, A., Infante, C., Rebordinos, L.: Molecular characterization and chromosomal mapping of the 5S rRNA gene in *Solea senegalensis*: a new linkage to the U1, U2, and U5 small nuclear RNA genes. Genome **49**, 79–86 (2006)
47. Mattaj, I.W., Zeller, R.: *Xenopus laevis* U2 snRNA genes: tandemly repeated transcription units sharing 5' and 3' flanking homology with other RNA polymerase II transcribed genes. EMBO J **2**, 1883–1891 (1983)
48. Missal, K., Rose, D., Stadler, P.F.: Non-coding RNAs in *Ciona intestinalis*. Bioinformatics **21 S2**, i77–i78 (2005)
49. Missal, K., Zhu, X., Rose, D., Deng, W., Skogerbø, G., Chen, R., Stadler, P.F.: Prediction of structured non-coding RNAs in the genome of the nematode *Caenorhabitis elegans*. J. Exp. Zool.: Mol. Dev. Evol. **306B**, 379–392 (2006)
50. Mitrovich, Q.M., Guthrie, C.: Evolution of small nuclear RNAs in *S. cerevisiae*, *C. albicans*, and other hemiascomycetous yeasts. RNA **13**, 2066–2080 (2007)
51. Montzka, K.A., Steitz, J.A.: Additional low-abundance human small nuclear ribonucleoproteins: U11, U12, etc. Proc Natl Acad Sci U S A **85**, 8885–8889 (1988)
52. Morales, J., Borrero, M., Sumerel, J., C., S.: Identification of developmentally regulated sea urchin U5 snRNA genes. DNA Seq. **7**, 243–259 (1997)
53. Mount, S.M., Gotea, V., Lin, C.F., Hernandez, K., Makałowski, W.: Spliceosomal small nuclear RNA genes in 11 insect genomes. RNA **13**, 5–14 (2007)
54. Mount, S.M., Steitz, J.A.: Sequence of U1 RNA from *Drosophila melanogaster*: implications for U1 secondary structure and possible involvement in splicing. Nucleic Acids Res **9**, 6351–6368 (1981)
55. Myslinksi, E., Krol, A., Carbon, P.: Characterization of snRNA and snRNA-type genes in the pufferfish *Fugu rubripes*. Gene **330**, 149–158 (2004)
56. Myslinski, E., Branlant, C., Wieben, E.D., Pederson, T.: The small nuclear RNAs of *Drosophila*. J. Mol. Biol. **180**, 927–945 (1984)
57. Nei, M., Rooney, A.P.: Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. **39**, 121–152 (2005)
58. Nilsen, T.W.: The spliceosome: the most complex macromolecular machine in the cell? Bioessays **25**, 1147–1149 (2003)
59. Otake, L.R., Scamborova, P., Hashimoto, C., Steitz, J.A.: The divergent U12-type spliceosome is required for pre-mRNA splicing and is essential for development in *Drosophila*. Mol Cell **9**, 439–446 (2002)
60. Papillon, D., Perez, Y., Caubit, X., Le Parco, Y.: Systematics of chaetognatha under the light of molecular data, using duplicated ribosomal 18S DNA sequences. Mol Phylogenet Evol. **38**, 621–634 (2006)
61. Patel, A.A., Steitz, J.A.: Splicing double: insights from the second spliceosome. Nat. Rev. Mol. Cell Biol. **4**, 960–970 (2003)
62. Pavelitz, T., Liao, D., Weiner, A.M.: Concerted evolution of the tandem array encoding primate U2 snRNA (the RNU2 locus) is accompanied by dramatic remodeling of the junctions with flanking chromosomal sequences. EMBO J **18**, 3783–3792 (1999)
63. Pelliccia, F., Barzotti, R., Bucciarelli, E., Rocchi, A.: 5S ribosomal and U1 small nuclear RNA genes: a new linkage type in the genome of a crustacean that has three different tandemly repeated units containing 5S ribosomal DNA sequences. Genome **44**, 331–335 (2001)
64. Pereira-Simon, S., Sierra-Montes, J.M., Ayesh, K., Martinez, L., Socorro, A., Herrera, R.J.: Variants of U1 small nuclear RNA assemble into spliceosomal complexes. Insect Molecular Biology **13**, 189–194 (2004)

65. Russell, A.G., Charette, J.M., Spencer, D.F., Gray, M.W.: An early evolutionary origin for the minor spliceosome. Nature **443**, 863–866 (2006)
66. Russell, A.G., Charette, J.M., Spencer, D.F., Gray, M.W.: An early evolutionary origin for the minor spliceosome. Nature **443**, 863–866 (2006)
67. S., L.E., H., W.R., S., C.F., *** authors: Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**, 69–87 (2005)
68. Schlötterer, C., Tautz, D.: Chromosomal homogeneity of Drosophila ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. Curr. Biol. **4**, 777–783 (1994)
69. Schneider, C., Will, C.L., Brosius, J., Frilander, M., Lührmann, R.: Identification of an evolutionarily divergent U11 small nuclear ribonucleoprotein paricle in *Drosophila*. Proc. Natl. Acad. Sci. USA **101**(26), 9584–9589 (2004)
70. Shambaugh, J.D., Hannon, G.E., Nilsen, T.W.: The spliceosomal U small nuclear RNAs of *Ascaris lumbricoides*. Mol Biochem Parasitol **64**, 349–352 (1994)
71. Sheth, N., Roca, X., Hastings, M.L., Roeder, T., Krainer, A.R., Sachidanandam, R.: Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Res **34**, 3955–3967 (2006)
72. Shukla, G.C., Cole, A.J., Dietrich, R.C., Padgett, R.A.: Domains of human U4atac snRNA required for U12-dependent splicing in vivo. Nucleic Acids Res **30**, 4650–4657 (2002)
73. Shukla, G.C., Padgett, R.A.: Conservation of functional features of U6atac and U12 snRNAs between vertebrates and higher plants. RNA **5**, 525–538 (1999)
74. Shukla, G.C., Padgett, R.A.: U4 small nuclear RNA can function in both the major and minor spliceosomes. Proc. Natl. Acad. Sci. USA **101**, 93–98 (2004)
75. Sierra-Montes, J.M., Freund, A.V., Ruiz, L.M., Szmulewicz, M.N., Rowold, D.J., Herrera, R.J.: Multiple forms of U2 snRNA coexist in the silk moth *Bombyx mori*. Insect Mol Biol **11**, 105–114 (2002)
76. Sierra-Montes, J.M., Pereira-Simon, S., Freund, A.V., Ruiz, L.M., Szmulewicz, M.N., Herrera, R.J.: A diversity of U1 small nuclear RNAs in the silk moth *Bombyx mori*. Insect Biochem Mol Biol **33**, 29–39 (2003)
77. Sierra-Montes, J.M., Pereira-Simon, S., Smail, S.S., Herrera, R.J.: The silk moth *Bombyx mori* U1 and U2 snRNA variants are differentially expressed. Gene **352**, 127–136 (2005)
78. Smail, S.S., Ayesh, K., Sierra-Montes, J.M., Herrera, R.J.: U6 snRNA variants isolated from the posterior silk gland of the silk moth *Bombyx mori*. Insect Biochem Mol Biol. **36**, 454–465 (2006)
79. Sontheimer, E.J., Steitz, J.A.: Three novel functional variants of human U5 small nuclear RNA. Mol. Cell. Biol. **12**, 734–746 (1992)
80. Stefanovic, B., Li, J.M., Sakallah, S., Marzluff, W.F.: Isolation and characterization of developmentally regulated sea urchin U2 snRNA genes. Dev Biol. **148**, 284–294 (1991)
81. Stefanovic, B., Marzluff, W.F.: Characterization of two developmentally regulated sea urchin U2 small nuclear RNA promoters: a common required TATA sequence and independent proximal and distal elements. Mol Cell Biol **12**, 650–660 (1992)
82. Tarn, W.Y., Yario, T.A., Steitz, J.A.: U12 snRNAs in vertebrates: Evolutionary conservation of 5' sequences implicated in splicing of pre-mRNAs containing a minor class of introns. RNA **1**, 644–656 (1995)
83. Telford, M.J., Holland, P.W.H.: Evolution of 28S ribosomal DNA in chaetognaths: Duplicate genes and molecular phylogeny. J. Mol. Evol. **44**, 135–144 (1997)
84. Thomas, J., Lea, K., Zucker-Aprison, E., Blumenthal, T.: The spliceosomal snRNAs of *Caenorhabditis elegans*. Nucleic Acids Res **18**, 2633–2642 (1990)
85. Tichelaar, J.W., Wieben, E.D., Reddy, R., Vrabel, A., Camacho, P.: *In vivo* expression of a variant human U6 RNA from a unique, internal promoter. Biochemistry **37**, 12,943–12,951 (1998)
86. Valadkhan, S.: snRNAs as the catalysts of pre-mRNA splicing. Curr. Op. Chem. Biol. **9**, 603–608 (2005)
87. Valadkhan, S.: The spliceosome: caught in a web of shifting interactions. Curr. Op. Struct. Biol. **17**, 310–315 (2007)
88. Valadkhan, S., Mohammadi, A., Wachtel, C., Manley, J.L.: Protein-free spliceosomal snRNAs catalyze a reaction that resembles the first step of splicing. RNA **13**, 2300–2311 (2007)
89. Will, C.L., Lührmann, R.: Splicing of a rare class of introns by the U12-dependent spliceosome. Biol. Chem. **386**, 713–724 (2005)

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-de ned 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.

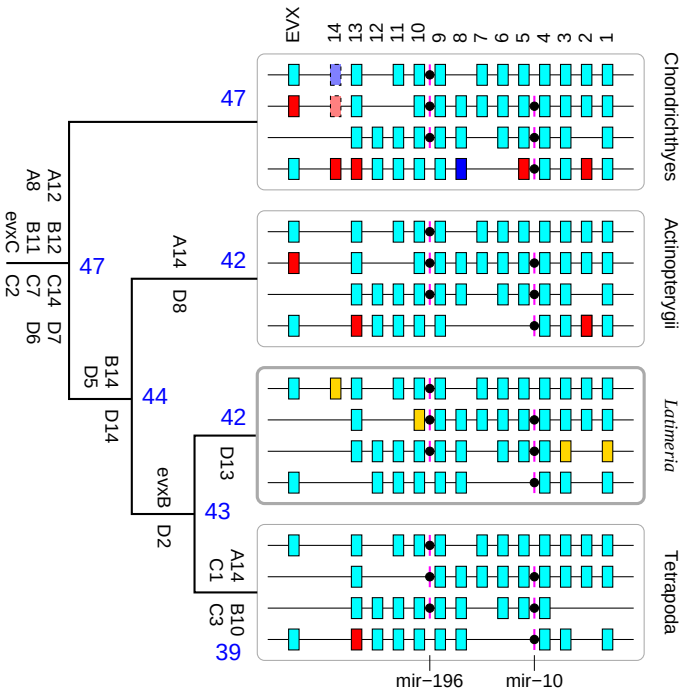


Fig. 1. Evolution of the HOX clusters in chordates. For each taxon, HOX clusters are illustrated from top to bottom, HOXA, HOXB, HOXC and HOXD. Genes shown in cyan inferred to constitute the ancestral states of the major chordate lineages. Dark blue boxes are losses in the actinopterygian stem linages; 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.

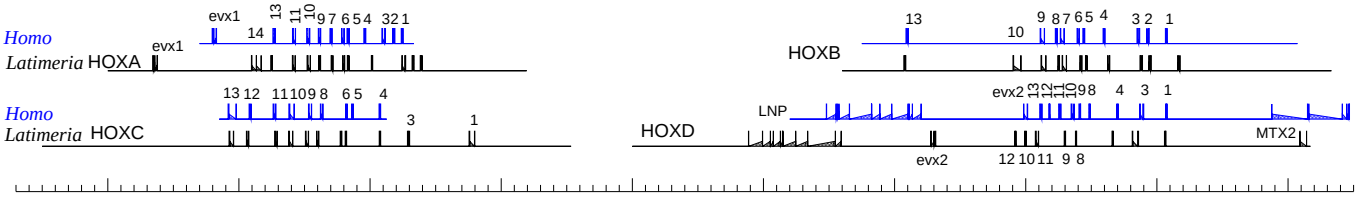


Fig. 2. Scale map of the *Latimeria menadoensis* HOX clusters compared to their human counterparts. Major tic marks are 100kb. Comparison of relative HOX cluster sizes and intergenic spacing among various vertebrates is given in Fig.S2.

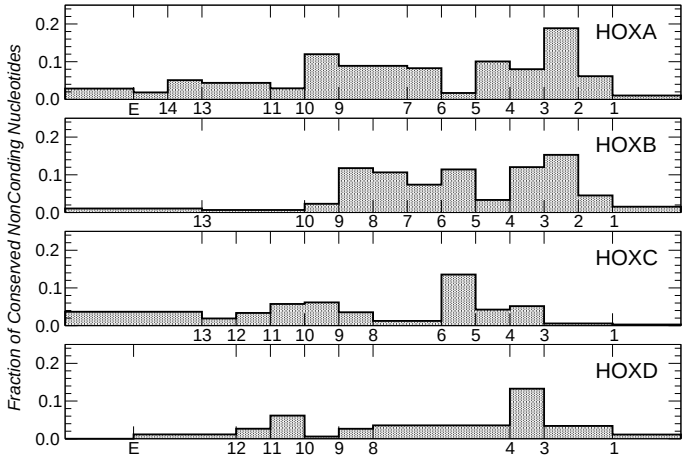


Fig. 3. Distribution of conserved non-coding DNA in intergenic regions between *Hox* genes. The figure summarizes the compilation of the conserved phylogenetic footprints as determined the `tracker` algorithm. A listing of all conserved footprints is given in the online supplement. For each intergenic 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.

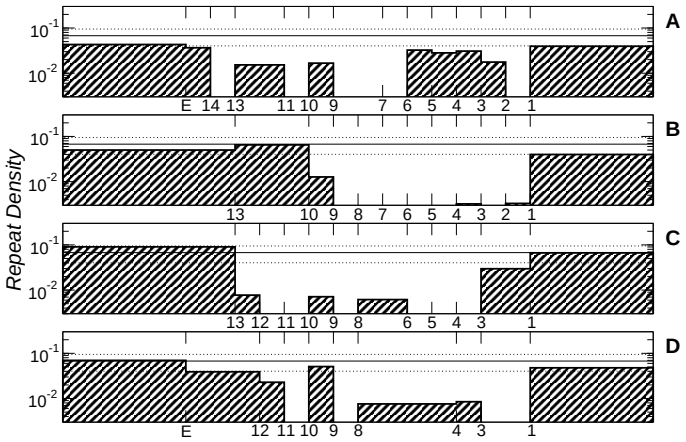


Fig. 4. Density of repetitive elements measured as the fraction of nucleotides annotated as interspersed repeats by `repeatmasker`. Numbers refer to *Hox* genens, E=*Evx*. The fraction of nucleotides in repetitive elements is shown on a log-scale for each IGR and the regions adjacent to the HOX clusters. 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.

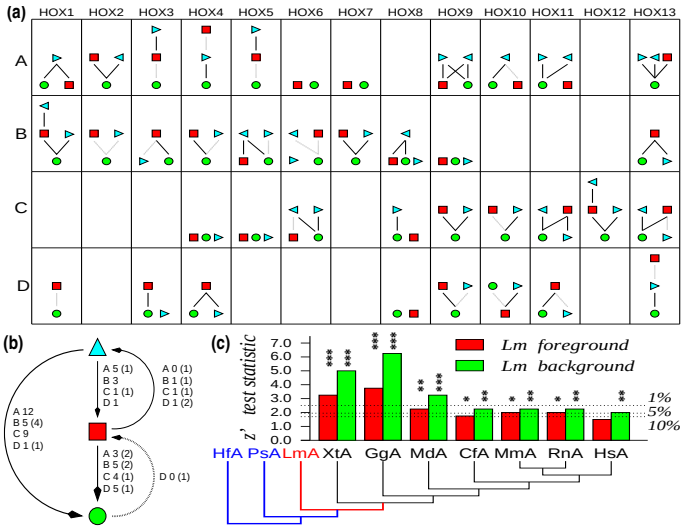


Fig. 5. Relative Rate Tests. **(a)** Summary of Tajima tests performed on Hox protein sequences using horn shark (HOXA, HOXB, HOXD) or elephant shark (HOXC) as outgroup. For each gene, a Hasse diagram shows highly significant ($p \leq 0.01$, full line) and significant ($0.01 < p \leq 0.05$, dotted line) 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).