Timing of Genome Duplications Relative to the Origin of the Vertebrates: Did Cyclostomes Diverge before or after?

Shigehiro Kuraku,† Axel Meyer,† and Shigeru Kuratani**

*Laboratory for Evolutionary Morphology, RIKEN Center for Developmental Biology, Kobe, Japan; and †Lehrstuhl für Zoologie und Evolutionsbiologie, Department of Biology, University of Konstanz, 78457 Konstanz, Germany

Two rounds of whole-genome duplications are thought to have played an important role in the establishment of gene repertoires in vertebrates. These events occurred during chordate evolution after the split of the urochordate and cephalochordate lineages but before the radiation of extant gnathostomes (jawed vertebrates). During this interval, diverse agnathans (jawless fishes), including cyclostomes (hagfishes and lampreys), diverged. However, there is no solid evidence for the timing of these genome duplications in relation to the divergence of cyclostomes from the gnathostome lineage. We conducted cDNA sequencing in diverse early vertebrates for members of homeobox-containing (*Dlx* and *ParaHox*) and other gene families that would serve as landmarks for genome duplications. Including these new sequences, we performed a molecular phylogenetic census using the maximum likelihood method for 55 gene families. In most of these gene families, we detected many more gene duplications before the cyclostome–gnathostome split, than after. Many of these gene families (e.g., visual opsins, *RAR*, *Notch*) have multiple paralogs in conserved, syntenic genomic regions that must have been generated by large-scale duplication events. Taken together, this indicates that the genome duplications occurred before the cyclostome–gnathostome split. We propose that the redundancy in gene repertoires possessed by all vertebrates, including hagfishes and lampreys, was introduced primarily by genome duplications. Apart from subsequent lineage-specific modifications, these ancient genome duplication events might serve generally to distinguish vertebrates from invertebrates at the genomic level.

Introduction

Phylogenetic and genomic studies have shown to near certainty that the chordates experienced genome duplications (McLysaght et al. 2002; Lundin et al. 2003; Dehal and Boore 2005; Putnam et al. 2008). These events are implicated in the evolutionary path leading to gnathostomes after the divergence from the urochordate and cephalochordate lineages but before that of chondrichthyans (Holland et al. 1994; Miyata and Suga 2001; Venkatesh et al. 2007; Hufton et al. 2008). During this interval, diverse groups of agnathans branched off successively, two of which (hagfishes and lampreys) have survived to date (Forey and Janvier 1993). Classically, these two groups of agnathans were placed in a single taxon, the Cyclostomata (Duméril 1806). However, from subsequent morphological observations, it was believed that hagfishes diverged first in vertebrate phylogeny so that lampreys are a sister group of gnathostomes (Forey 1984; Maisey 1986). Recently, reanalysis of hagfish morphology from a developmental viewpoint provided evidence against this hypothesis (Ota et al. 2007), which suggested that further developmental analyses of hagfishes might provide more confident evidence of the monophyly of Cyclostomata. Furthermore, molecular phylogenetic analyses of diverse genes have strongly supported the monophyly of the Cyclostomata (summarized in Kuraku and Kuratani 2006). Now, another question is whether cyclostomes diverged before or after the genome duplications (fig. 1), which will be critical for a deeper understanding of vertebrate evolution at the molecular level.

The idea of genome duplications, first proposed by Ohno (1970), has been strengthened by the observations that model vertebrates often have multiple (usually as many as four) duplicated genes (paralogs) corresponding to a single set of invertebrate orthologs (Holland et al. 1994; Sidow 1996; Spring 1997). This has been confirmed by the discoveries that similar arrays of genes are found on multiple chromosomes of model vertebrates (reviewed in Kasahara 2007). However, in cyclostomes, gene repertoires and the structures of landmark gene clusters, such as *Hox* and *Dlx*, tend to indicate an incomplete or degenerate state of their genomes (Pendleton et al. 1993; Sharman and Holland 1998; Neidert et al. 2001; Force et al. 2002; Irvine et al. 2002). This has prompted many researchers to believe that cyclostomes diverged in the midst of or before successive genome duplications (hypothesis B or C in fig. 1) and to attribute their "primitive" morphological traits to their possible intermediate genomic state (e.g., Donoghue and Purnell 2005). The picture is further obscured by possible gene duplications that have been imputed to the cyclostome lineage, as reported for *Hox* genes (Fried et al. 2003; Stadler et al. 2004), although it is unknown whether this represents a genome-wide phenomenon.

To date, investigations on intragenomic conserved syntenies (paralogons) have contributed to the identification of ancient genome duplications. However, signatures of genome duplications are obscured by subsequent lineage-specific gene duplications and gene losses. Therefore, simply counting the numbers of gene repertoires or conserved syntenies will not necessarily provide a reliable estimate for the scenario of genome evolution: the key to this question is the tree topology of molecular phylogenetic trees. For example, the timing of the genome duplication that occurred in the actinopterygian lineage was dated using only three protein-coding genes (Hoegg et al. 2004), and this result has been confirmed by subsequent analyses (Crow et al. 2006; Steinke et al. 2006). Hence, in principle, gene sampling from the animal groups in question and the knowledge of their phylogenetic relationships based on other independent information should lead us to a clear picture of genome evolution, even in the absence of large-scale genome sequencing. However, phylogenetic analyses of cyclostome genes have occasionally produced contradictory results. An analysis using 33 gene families that aimed

Key words: conserved synteny, hagfish, genome duplication, lamprey, 2R hypothesis.

E-mail: shigehiro.kuraku@uni-konstanz.de.

Mol. Biol. Evol. 26(1):47–59. 2009
doi:10.1093/molbev/msn222
Advance Access publication October 8, 2008

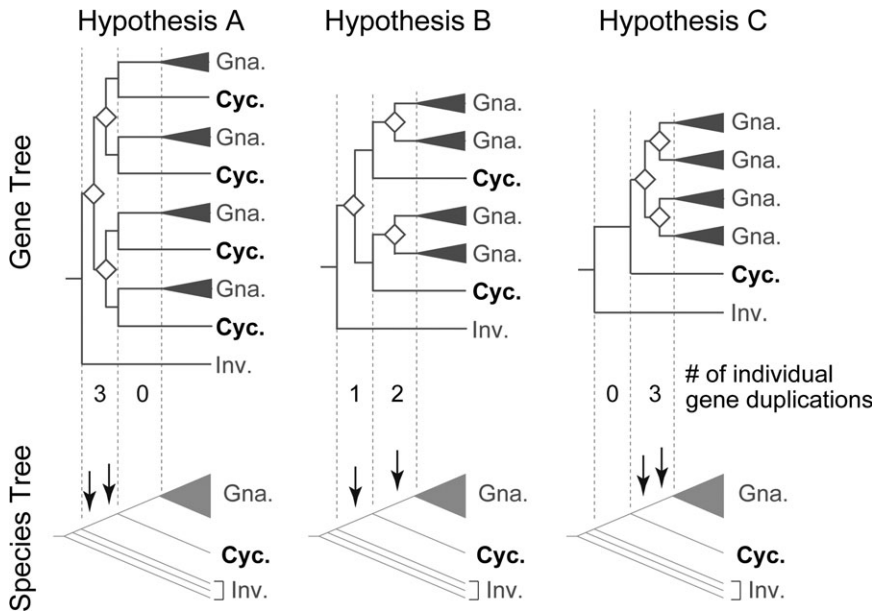

FIG. 1.—Three possible scenarios for timings of 2R genome duplications. Expected tree topologies (hypotheses A–C) for gene phylogeny are illustrated for an imaginary gene family comprising one invertebrate outgroup (Inv.) and four gnathostome (Gna.) genes with an intact set of cyclostome (Cyc.) genes. Hypothesis A has not been explicitly proposed so far, whereas there are some reports supporting hypothesis B (Pendleton et al. 1993; Sharman and Holland 1998; Escriva et al. 2002; Force et al. 2002; Stadler et al. 2004) or C (Fried et al. 2003; Furlong et al. 2007). Expected numbers of gene duplications before/after the cyclostome–gnathostome split in this gene family are indicated in the middle. Timings of genome duplications are shown as black arrows on the species phylogeny.

to resolve the timing of the genome duplications failed to provide a unified scenario but suggested that one genome duplication occurred before the cyclostome–gnathostome split and the other occurred after (hypothesis B) (Escriva et al. 2002). However, this analysis was based on a relatively unsophisticated phylogenetic method—Neighbor-Joining (NJ) method using distance matrix inferred by Poisson correction. Moreover, a recent phylogenetic analysis based on the whole-genome sequencing of *Branchiostoma floridae* could not solve this question, either (Putnam et al. 2008). Therefore, a reanalysis is necessary.

In this study, to assess the timing of genome duplications in early vertebrate evolution, we performed a set of molecular phylogenetic analyses with all annotated cyclostome genes available in public databases in addition to ones sequenced in this study. By optimizing data sets and phylogenetic methods, together with thorough taxon sampling, we obtained consistent results suggesting that the two rounds of genome duplication occurred before the divergence between ancestors of cyclostomes and gnathostomes.

Materials and Methods

General Molecular Phylogenetics

Sequences were retrieved from the NCBI Entrez Protein database (URL: http://www.ncbi.nlm.nih.gov/sites/entrez?db = Protein) and Ensembl (URL: http://www.ensembl.org/). Multiple alignments of amino acid sequences were constructed using MAFFT (Katoh et al. 2005) followed by manual improvement. Unambiguously aligned regions without gaps were used for tree inferences (for the alignments used in this study, see supplementary data set S1, Supplementary Material online). To select gene families, we used the NJ method (Saitou and Nei 1987) and the quartet-puzzling algorithm implemented in Tree-Puzzle 5.2 (Schmidt et al. 2002). We excluded teleost fishes and *Xenopus laevis*, which have undergone additional genome duplications in independent lineages (Bisbee et al. 1977; Hoegg et al. 2004; Crow et al. 2006), and instead included nonteleost actinopterygians and chondrichthyans, when available. Multiple invertebrate sequences with ordinary evolutionary rates, absence of many unique gaps in alignment and robust support of orthologies, were rigorously selected for outgroups. The maximum likelihood (ML) analysis for selected gene families was performed using Tree-Puzzle 5.2, by inputting all possible tree topologies in "user-defined trees" mode, assuming the JTT + F + Γ_8 model. Phylogenetic relationships within gnathostomes and outgroups were constrained based on a generally accepted species phylogeny (Meyer and Zardoya 2003; Cracraft and Donoghue 2004). Bootstrap probabilities were calculated with 1,000 replicates. Bayesian posterior probabilities were calculated using MrBayes 3.1 (Ronquist and Huelsenbeck 2003). Shimodaira–Hasegawa (SH) test and approximately unbiased (AU) test were performed using CONSEL (Shimodaira and Hasegawa 2001). For counting discrete numbers of gene duplications, NJ trees were also inferred with the data sets used in the ML analysis, assuming the JTT + Γ_4 model.

Probabilistic Counts of Gene Duplications

After performing ML analyses, the probability of each tree topology was calculated with the resampling of estimated log-likelihoods (RELL) approximation (Kishino

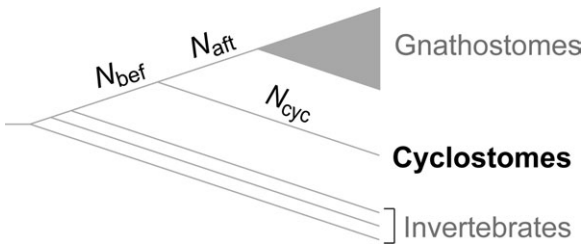

FIG. 2.—Indices used for counting gene duplications. N_{bef}, N_{aft}, and N_{cyc} represent the numbers of gene duplications estimated to occur in each branch indicated. N_{unk} represents the estimated number of gene duplications whose timings were not successfully assigned to any of N_{bef}, N_{aft}, or N_{cyc}.

et al. 1990) using CONSEL (Shimodaira and Hasegawa 2001). For each gene family, the number of gene duplications before the cyclostome–gnathostome split (N_{bef}) was calculated as

$$N_{bef} = \sum_i Pi \cdot n_{befi}, \quad (1)$$

where n_{befi} is the number of gene duplications on that branch in a given tree topology (topology i) and P_i is the RELL probability of the topology (fig. 2). Standard error was calculated based on bootstrap resampling of log-likelihood produced above by the ML method (1,000 replicates). Similarly, the number of gene duplications after the cyclostome–gnathostome split (N_{aft}), the number of gene duplications in the cyclostome lineage (N_{cyc}), and the number of gene duplications unassigned to any branch (N_{unk}) were calculated as follows:

$$N_{aft} = \sum_i Pi \cdot n_{afti}, \quad (2)$$

$$N_{cyc} = \sum_i Pi \cdot n_{cyci}, \quad (3)$$

$$N_{unk} = \sum_i Pi \cdot n_{unki}. \quad (4)$$

Assumption of the Two-Round Pattern

In gene families with three or four gnathostome paralogs, among all possible tree topologies, those representing the two-round (2R) pattern of gene duplication—for example, ((*a*, *b*), (*c*, *d*)), where *a*–*d* represent four gnathostome paralogs—were selected, and they were categorized in hypotheses A–C (supplementary table S1, Supplementary Material online). When the topology did not allow us to distinguish between hypotheses A and B, the result was categorized as "hypothesis A/B." In gene families with only two gnathostome paralogs, a tree topology with a gene duplication before the cyclostome–gnathostome split is shown as "hypothesis A/B," whereas one with a gene duplication after that split is shown as "hypothesis B/C." Probabilities of these hypotheses were calculated by summing RELL probabilities of tree topologies that supported each hypothesis.

Selection of Gene Families

For 3,207 entries for protein-coding hagfish and lamprey genes in the NCBI Protein database, we surveyed gene families for further analysis. Our criteria were the length of any unambiguous alignment (≥150 amino acids), the availability of invertebrate orthologs as outgroups (e.g., orthologies of cyclostome genes to gnathostome genes are sometimes discussed carelessly in the absence of appropriate outgroups [Kawakoshi et al. 2006; Haitina et al. 2007]), the number of gnathostome paralogs corresponding to a single invertebrate ortholog (>2 paralogs) and the resolvability of gnathostome phylogeny (for genes excluded, see supplementary data set S2, Supplementary Material online). To reduce orthology/paralogy misidentification, we excluded gene families that are prone to gene duplications (i.e., gene families with more than two gene duplications in the lineage leading to humans during gnathostome evolution; e.g., globins, neurofilaments, and olfactory receptors). We also excluded gene families in which more than two paralogs are found on a single human chromosome because this study was intended to focus exclusively on inter- rather than intrachromosomal, gene duplications (note that, even under this criterion, tandem duplications cannot be completely excluded from our data set; see Discussion). Finally, we selected 55 gene families that contained at least one cyclostome gene (see supplementary table S2, Supplementary Material online). Of these, 21 had two gnathostome paralogs, whereas 34 had more than three gnathostome paralogs. These two family categories were analyzed separately. In gene families with two and three gnathostome paralogs, two and one duplicates were assumed to have been lost secondarily following the genome duplication event, respectively. Using the human genome as a reference, members of these gene families, totaling 150 genes, covered all the 22 autosomes and the X chromosome (supplementary fig. S1, Supplementary Material online).

Identification of Synteny Groups

Amino acid sequences of human protein-coding genes, sorted in physical order, were retrieved for each chromosome from Ensembl via the EnsMart interface. We performed pairwise BlastP searches (Altschul et al. 1997) reciprocally between chromosomes harboring members of each gene family. For interchromosomal gene pairs that showed high "bits" scores (>100), we inferred molecular phylogenetic trees using amino acid sequences of other animals as above. We excluded gene families in which family members were revealed to have duplicated before the split between invertebrate chordates and vertebrate lineages or within gnathostome lineages, even if those family members were located on the set of chromosomes in question.

The synteny group containing the opsin gene family, one of the gene families on which we focused in this study, was previously reported (Nordstrom et al. 2004). However,

Table 1
ML Analysis for the Homeobox-Containing Genes

		2R Pattern Not Assumed						2R Pattern Assumed[a]			
				Probabilistic Count of Gene Duplications				Supported Hypothesis with the Highest Probabilities			
Gene Family	Length (aa)	No. of Tree Topologies within 1σ of log*L*[b]	Supported Hypothesis in ML Tree[c]	N_{bef}	N_{aft}	N_{cyc}[d]	N_{unk}	A	A/B	B	C
HoxA13-D13	64	252 (945)	A/B	2.53	0.18	0.01	0.77	0.68	0.00	0.31	0.01
Dlx1/4/6	63	381 (945)	B	1.70	1.39	0.17	0.15	0.06	0.03	0.45	0.47
Dlx2/3/5	65	208 (945)	A	2.42	0.32	0.81	0.12	0.53	0.01	0.42	0.04
Cdx1/2/4(ParaHox)	64	27 (105)	A/B	1.81	0.86	0.06	0.15	0.24	0.03	0.33	0.40
Gsh1/2 (ParaHox)	59	3 (3)	A/B	0.91	0.10	N/A	0.00	N/A	N/A	N/A	N/A

NOTE.—2R, two-round genome duplications; aa, amino acid; log*L*, log-likelihood; N/A, not applicable.
[a] Only gene families with three or four gnathostome paralogs were analyzed.
[b] Numbers of all possible tree topologies are shown in parentheses.
[c] See figure 2 for details.
[d] This does not apply to the Gsh1/2 family in which only one cyclostome paralog is available.

their results erroneously contained gene families whose members are located on more than two of human chromosome (HSA) 1, 3, 7, and X but were duplicated in a more ancient period of animal evolution, rather than the period of early vertebrate evolution. For example, in their study, plexin A2, B1, and B3 were proposed as components of these syntenic regions, although subfamilies A and B in the plexin family diverged in the early phase of animal evolution: plexin A2, B1, and B3 should not be regarded as the components of these synteny groups. In addition, some gene families in the synteny group they proposed, such as *CLSTN1/2*, *KCNAB1/2*, and *STAG1/3*, have another member that was duplicated concomitantly early in vertebrate evolution but is not located on either of HSA 1, 3, 7, and X.

In identifying the retinoic acid receptor (RAR)–related synteny group comprising HSA 3–12–17, we also detected syntenic regions on similar sets of HSAs, that is, HSA 2–7–12–17, 1–3–7–12, and 1–3–12–X. As described previously (McLysaght et al. 2002), these quadruplets of syntenic regions are juxtaposed in an interdigitated manner, possibly because of secondary interchromosomal rearrangements that occurred after the genome duplications.

Results

Analysis on Homeobox-Containing Gene Families

We first focused on the homeobox-containing gene families, *Hox*, *Dlx*, and *ParaHox*. In the *Hox* gene family, we selected paralogous group 13 (*Hox13*) because of the presence of an intact set (*HoxA13*, *B13*, *C13*, and *D13*) of gnathostome paralogs and a high level of divergence compared with other paralogous groups (Ferrier et al. 2000). So far, two *Hox13* genes that cover an entire homeobox have been reported for cyclostomes (*LjHox13α* [GenBank accession number AB293597] and *LjHox13β* [AB293598]; Kuraku et al. 2008). In the *Dlx* family, four genes (*DlxA–DlxD*; AY010116–AY010119) were identified previously in the sea lamprey *Petromyzon marinus* (Neidert et al. 2001). In the Japanese lamprey *Lethenteron japonicum*, we performed reverse transcription–polymerase chain reaction (RT–PCR) (supplementary text S1, Supplementary Material online) and isolated two novel *Dlx* genes (*LjDlxE* and *LjDlxF*) as well as orthologs of the four genes (*LjDlxA–LjDlxD*) reported previously for *P. marinus* (AB292628–AB292633; supplementary figs. S2 and S3, Supplementary Material online). In the *ParaHox* gene family, we focused on the *Cdx* and *Gsh* subfamilies because the *Xlox* subfamily, the third member of the *ParaHox* gene family, has only one gene in most gnathostomes. For each of the *Cdx* and *Gsh* families, in addition to the hagfish homologues previously reported (Furlong et al. 2007), we isolated one gene in *L. japonicum*, designated as *LjCdxA* (AB368821) and *LjGshA* (AB368822), respectively (supplementary text S1, Supplementary Material online).

To examine the molecular phylogenies of these cyclostome genes, we performed ML tree inferences. The *Dlx* family was divided into two subgroups to be analyzed separately—*Dlx1/4/6* and *Dlx2/3/5*—because each of these has a urochordate ortholog (Stock et al. 1996; Irvine et al. 2007). In the rest of this article, the phrase "gene family" denotes a single set of invertebrate orthologs plus a group of vertebrate homologues that have derived from the ancestral invertebrate ortholog.

In all these five gene families (*Hox13*, *Dlx1/4/6*, *Dlx2/3/5*, *Cdx*, and *Gsh*), the results did not support a specific tree topology with high confidence, leaving multiple tree topologies within 1σ of log-likelihood compared with the ML tree (table 1). In addition, support values for each node were low (fig. 3). We estimated the timings of gene duplications, based on the probabilistic method for counting gene duplications for each branch (N_{bef}, N_{aft}, and N_{cyc}; see Materials and Methods). For all these five families, we detected more gene duplications before the cyclostome–gnathostome split than after ($N_{bef} > N_{aft}$; table 1). We also analyzed which of the three possible hypotheses would be supported under the assumption of 2R gene duplication pattern (see Materials and Methods). None of these families supported a single hypothesis with high confidence ($P > 0.70$; table 1).

We also performed ML analyses for all these families without constraints on the phylogenetic relationships within gnathostomes and outgroups. In the ML trees of all families except for *Dlx2/3/5*, the relationships within gnathostomes were not properly reconstructed (supplementary fig. S4*A–E*,

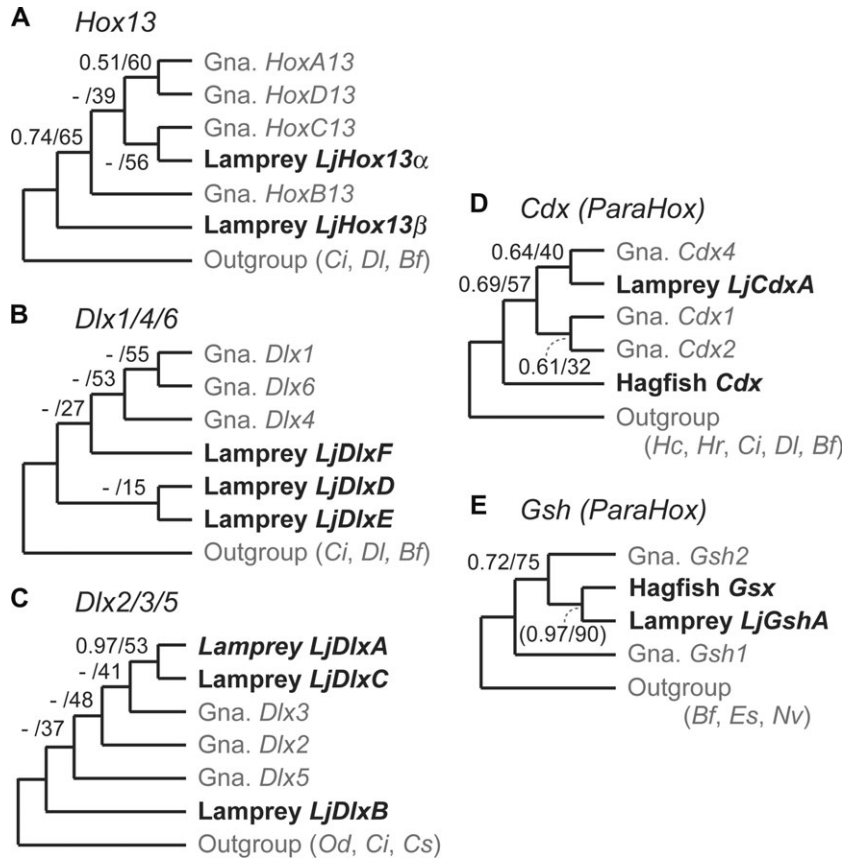

FIG. 3.—The ML trees for selected homeobox-containing gene families. Tree topologies supported by ML analysis are shown for *Hox13* (A), *Dlx1/4/6* (B), *Dlx2/3/5* (C), *Cdx* (D), and *Gsh* (E). For simplicity, relationships within outgroups, gnathostome paralogs (Gna.), and two hagfish *Gsx* genes (see Furlong et al. 2007, for details) are not displayed. Numbers at nodes indicate Bayesian posterior probabilities (left) and intact bootstrap probabilities from ML analysis (right). A hyphen indicates that the phylogenetic relationship at the node was not fully resolved in the Bayesian analysis. Numbers in parentheses indicate supporting values obtained in analyses in which the hagfish–lamprey relationship is not constrained. Abbreviations for species names: *Bf*, *Branchiostoma floridae*; *Ci*, *Ciona intestinalis*; *Dl*, *Diplosoma listerianum*; *Od*, *Oikopleura dioica*; *Cs*, *Ciona savignyii*; *Hc*, *Herdmania curvata*; *Hr*, *Halocynthia roretzi*; *Es*, *Euprymna scolopes*; *Nv*, *Nereis virens*.

Supplementary Material> online). Additionally, in the *Dlx1/4/6*, *Cdx*, and *Gsh* families, the invertebrate genes that should be regarded as an outgroup did not form a single cluster (supplementary fig. S4*B*, *D*, and *E*, Supplementary Material online). For the *Gsh* and *Cdx* families, we performed further analyses by adding our lamprey sequences to the data sets used previously (Furlong et al. 2007). The ML trees did not support hypothesis C but hypothesis A (see supplementary fig. S4*F* and *G*, Supplementary Material online). Again, the relationships within the gnathostomes and outgroups were not properly reconstructed.

ML Analysis for the 55 Gene Families without the Assumption of the "2R" Duplication Tree Topology

For all the 55 gene families selected with rigorous criteria (see Materials and Methods), we performed ML analyses (see supplementary table S3, Supplementary Material online). In 48 of these gene families, there were more than two tree topologies within 1σ of log-likelihood compared with the ML tree (for details of log-likelihood, SH test, and AU test, see supplementary fig. S5, Supplementary Material online). In the ML trees of gene families with three or four gnathostome paralogs ($n = 33$), 13 families supported tree topologies with no gene duplication after the cyclostome–gnathostome split (hypothesis A; table 2A). Five families supported tree topologies with no gene duplication before the cyclostome–gnathostome split (hypothesis C), whereas 11 supported tree topologies with gene duplications both before and after the cyclostome–gnathostome split (hypothesis B; table 2A). Topologies of ML trees in four of the gene families did not allow us to distinguish between hypothesis A or B (designated hypothesis A/B). On the other hand, eight families rejected hypothesis C at a significant σ level for log-likelihood, whereas only three (protein tyrosine phosphatase [PTP] R2, PTP RM/RK/RT/RL, and PACSIN1/2) and one (PTP R2) rejected hypotheses A and B, respectively (table 2A). In the ML trees of gene families with only two gnathostome paralogs ($n = 21$), 18 supported the tree topology with a gene duplication before the cyclostome–gnathostome split (hypothesis A/B), whereas only three supported a tree topology with a gene duplication after the cyclostome–gnathostome split (hypothesis B/C; table 2B). By contrast, none of these families rejected hypothesis A/B at a significant level of σ for log-likelihood, whereas six rejected hypothesis B/C (table 2B).

Table 2
Summary of ML Analysis

(A) Gene Families with Three or Four Gnathostome Paralogs (n= 33)

	No. of Gene Families Supporting Each Hypothesis								No. of Gene Families Rejecting Each Hypothesis					
	2R Pattern Not Assumed[a]				2R Pattern Assumed[b]				2R Pattern Not Assumed[c]			2R Pattern Assumed[d]		
Average Length (aa)	A	A/B	B	C	A	A/B	B	C	A	B	C	A	B	C
288.4	13	4	11	5	15	3	9	6	3	1	8	2	0	10

(B) Gene Families with Two Gnathostome Paralogs (n = 21)

Average Length (aa)	No. of Gene Families Supporting Each Hypothesis[a]		No. of Gene Families Rejecting Each Hypothesis[c]	
	A/B	B/C	A/B	B/C
298.9	18	3	0	6

NOTE.—See supplementary table S3 (Supplementary Material online) for details of each gene family. 2R, two-round genome duplications; aa, amino acids.
[a] Based on tree topology of the ML tree.
[b] Hypotheses supported with the highest probabilities are shown.
[c] Rejected within 1σ of log-likelihood.
[d] Probability less than 0.05 was regarded as rejected.

Probabilistic Count of Gene Duplications for the 55 Gene Families

We calculated the numbers of gene duplications for each branch using the probabilistic method (see Materials and Methods; supplementary tables S4 and S5, Supplementary Material online). The total number of N_{bef} for all 55 families was approximately double the value for N_{aft} (N_{bef} = 67.4 ± 6.4; N_{aft} = 33.1 ± 10.2; table 3). Counts of discrete numbers of gene duplications in the ML and NJ trees produced similar results (table 3). The tendency of much higher N_{bef} than N_{aft} was observed in total counts for both family categories, regardless of which cyclostome taxon (hagfishes and/or lampreys) was used (table 3).

To further characterize this tendency, we analyzed the distribution of N_{bef} and N_{aft} among all families (fig. 4). For both family categories, the N_{aft} value distributed mostly between 0 and 1.0 (fig. 4*B* and *D*). By contrast, the peak in the distribution of N_{bef} located at 0.5–1.0 for gene families with two gnathostome paralogs (fig. 4*A*), whereas it located at 1.0–2.0 for gene families with three to five gnathostome paralogs (fig. 4*C*).

The total number of gene duplications in the cyclostome lineage (N_{cyc}) was estimated at 10.2 when data for both hagfishes and lampreys were used (table 3). However, this value should be an underestimate because N_{cyc} is applicable only for gene families where multiple cyclostome paralogs are available (only 19 out of the 55 gene families). In the ML trees of eight of these 19 families, N_{cyc} was more than 0.5, suggesting high possibilities of cyclostome lineage–specific gene duplications (for phylogenetic trees of these gene families, see supplementary fig. S6, Supplementary Material online; for further discussion, see supplementary text S2, Supplementary Material online).

ML Analysis for the 55 Gene Families under the Assumption of the "2R" Duplication Tree Topology

Assumption of the 2R pattern of gene duplications is applicable only to gene families with three or four gnathostome paralogs (for discussion over relevance of 2R genome duplications, see supplementary text S2, Supplementary Material online). For these gene families (n = 33), we evaluated the probabilities of hypotheses A–C by summing the probabilities of tree topologies that were categorized in each hypothesis (see Materials and Methods). The hagfish–lamprey clusters

Table 3
Total Estimated Numbers of Gene Duplications

Family Category	Count Method	Cyclostome Used		No. of Gene Families Analyzed	Total Count of Gene Duplications			
		Hagfish	Lamprey		N_{bef}	N_{aft}	N_{cyc}	N_{unk}
All families (two to five gnathostome paralogs)	NJ discrete	+	+	55	72	27	11	11
	ML discrete	+	+	55	73	30	10	11
	ML Probabilistic	+	+	55	67.4	33.1	10.2	9.1
		+	−	23	24.1	17.1	3.3	3.1
		−	+	47	54.4	30.5	8.7	7.7
		+	+	21	17.6	5.6	4.9	0.0
Families with only two gnathostome paralogs		+	−	7	5.2	2.7	0.1	0.0
		−	+	16	12.6	3.8	4.7	0.0
		+	+	34	49.8	27.5	5.3	9.1
Families with more than three gnathostome paralogs		+	−	15	18.9	14.4	3.2	3.1
		−	+	31	41.8	26.7	3.9	7.7

NOTE.—See supplementary tables S4 and S5 (Supplementary Material online) for details for each gene family analyzed here.

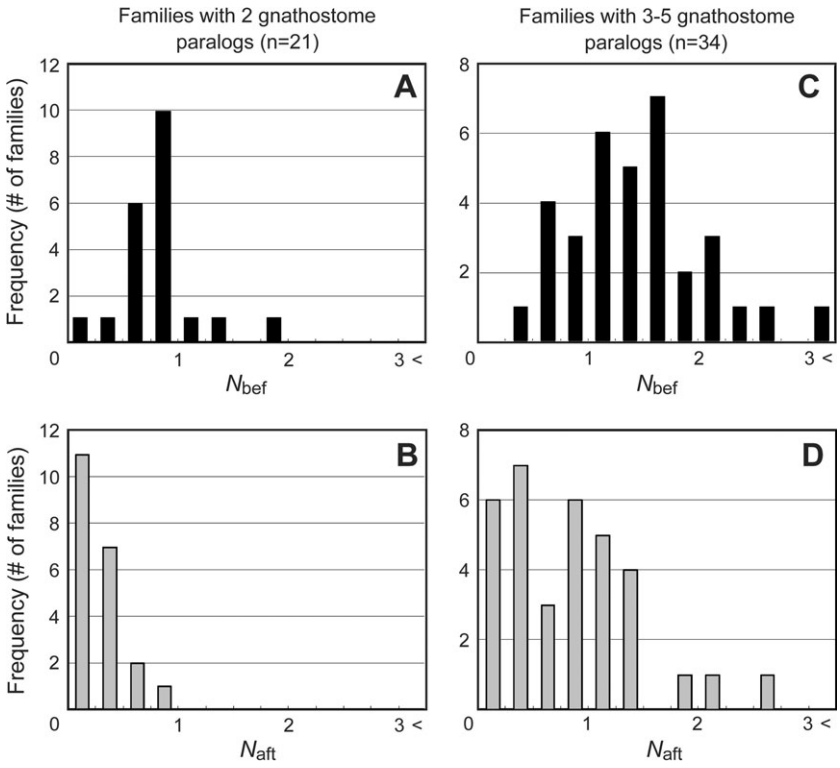

FIG. 4.—Distributions of estimated numbers of gene duplications before/after the cyclostome–gnathostome split. (A, C) Histograms of the distribution of gene duplications before the cyclostome–gnathostome split. (B, D) Histograms of the distribution of gene duplications after the cyclostome–gnathostome split. (A, B) Results for gene families with two gnathostome paralogs. (C, D) Results for gene families with three to five gnathostome paralogs. For details of N_{bef} and N_{aft}, see Materials and Methods.

supported by the ML trees were constrained (see supplementary text S2 and table S6, Supplementary Material online). Hypothesis A was supported by 15 gene families with the highest probabilities (table 2A). Hypotheses B and C were supported by nine and six gene families, respectively (table 2A). If a hypothesis with $P < 0.05$ was regarded as rejected, hypotheses A and C were rejected by two and ten gene families, respectively, whereas none of these families rejected hypothesis B (table 2A).

Analysis of Gene Families with Intensive Taxon Sampling

To gain a more definitive picture for cyclostome gene phylogenies, we focused on the gene families in which multiple cyclostome paralogs have been identified by intensive survey, namely the visual opsin and RAR families (fig. 5). For the former, five visual opsin paralogs are reported for lampreys (Hisatomi et al. 1991; Collin et al. 2003; Koyanagi et al. 2004), whereas gnathostomes generally possess five paralogs, including two paralogs (green and blue opsins) that have been lost secondarily in the mammalian lineage (Jacobs 1993). Our molecular phylogenetic study strongly suggested that each lamprey gene clusters with one of those gnathostome paralogs (bootstrap probabilities, 68–100%) (fig. 5*A*), suggesting that all the gene duplications occurred before the cyclostome–gnathostome split. We found that *RHO*, *OPN1SW*, and *OPN1LW* are located in syntenic regions comprising HSA 3, 7, and X (fig. 5*B*; see Materials and Methods). In addition, the chicken green opsin gene constitutes another syntenic region that includes the orthologs of *PLXNA2*, *NFASC*, *WNT2B*, *LRRN5*, *MAPKAPK2*, and *KCND3*, all of which are located within a 2.7-Mb region on chicken chromosome 26. This suggests that HSA 1, which harbors orthologs of the genes, is the fourth member of this synteny group.

In the RAR family, three human genes, α, β, and γ, form a synteny group comprising HSA 17, 3, and 12, respectively (fig. 5*D*; see Materials and Methods). In this family, only one gene has been reported for cyclostomes (*P. marinus RAR* gene; U93433, AY861455) (Escriva et al. 2006). We conducted an intensive RT–PCR survey of various tissues from cartilaginous fishes (the plownose chimaera, *Callorhinchus callorhynchus*, and the cloudy catshark, *Scyliorhinus torazame*), from a nonteleost actinopterygian fish (the Florida spotted gar, *Lepisosteus platyrhincus*), and from three cyclostome species (the inshore hagfish *Eptatretus burgeri*, the Japanese lamprey *L. japonicum*, and the short-headed lamprey *Mordacia mordax*; see supplementary text S1, Supplementary Material online). As with gnathostomes, we identified three paralogs for each cyclostome species, which were designated as *RAR1*, *RAR2*, and *RAR3* (AB292622–AB292624). In the ML tree, cyclostomes *RAR1* and *RAR3* clustered with gnathostomes *RAR*γ and *RAR*α, respectively (fig. 5*C*). A clustering of cyclostome *RAR2* with gnathostome *RAR*β was not supported in this tree but was supported in the second ML tree ($\Delta \log L = 2.03 \pm 4.16$). Our probabilistic count of gene duplications detected a much higher N_{bef} than N_{aft} ($N_{bef} = 1.82 \pm 0.32$, $N_{aft} = 0.03 \pm 0.18$; table 4).

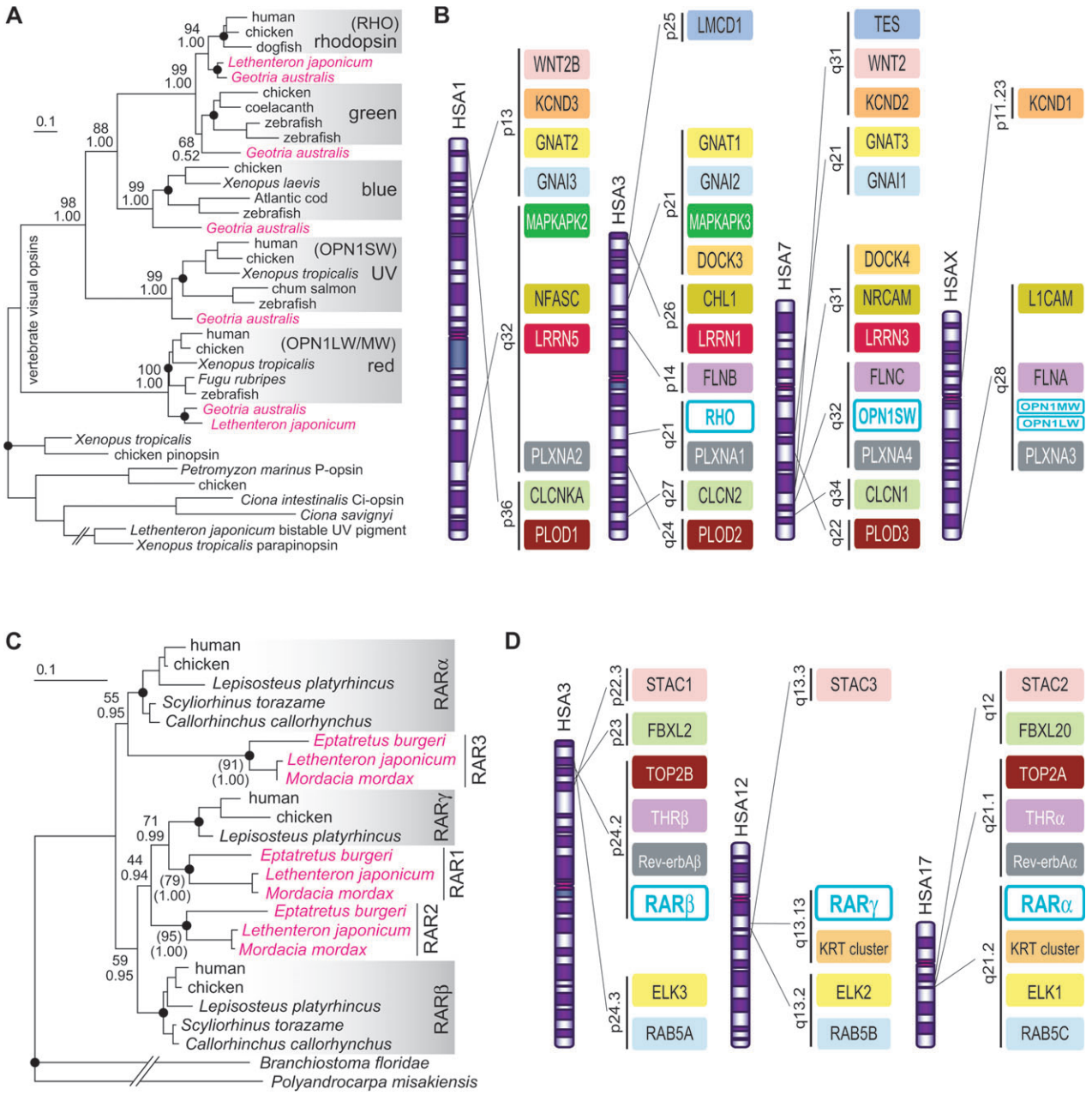

FIG. 5.—Phylogenies and synteny groups of vertebrate visual opsin and RAR genes. (A) The ML tree of vertebrate visual opsin genes (280 amino acid sites; shape parameter for gamma distribution $\alpha = 0.92$). (B) The visual opsin–related synteny group. (C) The ML tree of *RAR* genes (292 amino acid sites; $\alpha = 0.49$). (D) The RAR-related synteny group. In (A) and (C), cyclostome genes are shown in red. Scale bar indicates 0.1 substitutions per site. Numbers at nodes indicate intact bootstrap probabilities (upper) and Bayesian posterior probabilities (lower). Relationships within the nodes shown with closed circles were constrained. Numbers in parentheses indicate supporting values obtained in analyses in which the hagfish–lamprey relationship was not constrained. In (B) and (D), human homologues derived from a single invertebrate gene are aligned horizontally and shown in matching colors.

It is worth noting that no evident clusterings of multiple cyclostome paralogs were detected in the ML trees of these families, as indicated by extremely small N_{cyc} values ($N_{cyc} = 0.00 \pm 0.00$ and $N_{cyc} = 0.04 \pm 0.20$, for the visual opsin and RAR families, respectively).

Analysis of Gene Families in Well-Studied Synteny Groups

Our data set contained gene families harbored by other well-characterized synteny groups: the NOTCH and bromodomain-containing BRD families in the major histocompatibility complex (MHC)-related synteny group (Kasahara et al. 1997) (table 4) and the amphiphysin and solute carrier family SLC4A families in the *Hox*-related synteny group (Larhammar et al. 2002) (table 4). The latter also contains the *Dlx* clusters. These synteny groups, conserved between chromosomes, were generated by an ancient genome duplication event (Kasahara et al. 1997; Larhammar et al. 2002). In the ML trees of these gene families, each cyclostome gene clusters with a gnathostome paralog (see supplementary fig. S7*A* and *B*, Supplementary Material online),

Table 4
Timing of Gene Duplications in Conserved Synteny Groups

Gene Family	Alignment Length (aa)	Chromosomal Location in Human	Supported Hypothesis[a]	Probabilistic Count of Gene Duplications			
				N_{bef}	N_{aft}	N_{cyc}[b]	N_{unk}
MHC-related synteny group							
Complement component C3/C4/C5	642	6, 9, 19	A	2.00	0.16	0.07	0.26
Bromodomain-containing BRD T/2/3/4	265	1, 6, 9, 19	A	2.04	0.48	N/A	0.49
Notch 1/2/3/4	347	1, 6, 9, 19	A	1.55	0.81	N/A	0.63
Protein tyrosine phosphatase LAR/σ/RD	452	1, 9, 19	C	0.40	1.26	1.81	0.18
Retinoid X receptor (RXR) α/β/γ	276	1, 6, 9	A/B	1.13	0.65	0.84	0.36
Calcium channel voltage-dependent α 1 A/B/E	401	1, 9,19	A	1.11	0.77	N/A	0.12
Hox-related synteny group							
Solute carrier family 4 (SLC4A) 1/2/3	424	2, 7, 17	A	1.71	0.20	N/A	0.09
Amphiphysin AMPH/BIN1/BIN2	208	2, 7, 12	A	1.61	0.31	N/A	0.09
AMP-activated protein kinase γ 1/2/3	267	2, 7, 12	A	1.76	0.42	0.09	0.15
RAR-related synteny group							
RAR α/β/γ	292	3, 12, 17	A	1.82	0.03	0.04	0.11
Nuclear receptors NR1I 1/2/3 (VDR/PXR/CAR)	259	1, 3, 12	A/B	1.07	0.03	N/A	0.90
Glucose transporter (SLC2A) 1/2/3/4	417	1, 3, 12, 17	A	1.57	0.64	N/A	0.79
Enolase α/β/γ	332	1, 12, 17	C	1.47	1.41	0.03	0.08

NOTE.—Gene families with more than three gnathostome paralogs are shown. N/A, not applicable; aa, amino acids.
[a] Hypotheses with the highest probabilities in the ML analysis under the assumption of the 2R pattern are shown.
[b] For gene families with only one cyclostome gene, N_{cyc} is indicated as 'N/A'.

supporting hypothesis A. Similar results were obtained for members of the RAR-related synteny group, which is extended up to an entire chromosomal level, namely the nuclear receptor–encoding NR1I family and the glucose transporter SLC2A family (table 4). Especially in the NR1I family, the *P. marinus* vitamin D receptor (VDR) gene clustered with gnathostome VDR genes with very high confidence (bootstrap probability, 98%; see supplementary fig. S7*C*, Supplementary Material online). Although the tree topology did not allow us to distinguish between hypotheses A and B, this gene family explicitly rejected hypothesis C at a significant level of σ for log-likelihood (see supplementary fig. S7 and table S3, Supplementary Material online). In these gene families, although the bootstrap probabilities for the clusters consisting of a gnathostome paralog and a cyclostome paralog are not uniformly high (39–98%; see supplementary fig. S7, Supplementary Material online), our probabilistic count of gene duplications provided much higher values for N_{bef} than N_{aft} (table 4), with the exceptions of the PTP LAR/σ/RD and enolase α/β/γ families.

Discussion

Are Homeobox-Containing Genes Reliable Landmarks for Dating Genome Duplications?

Our current understanding of the timing of genome duplication events largely depends on analyses of gene repertoires in cyclostomes, especially the *Hox*, *Dlx*, and *ParaHox* genes. Our identification of six *Dlx* genes in the lamprey, as in other nonteleost gnathostomes (Stock et al. 1996; Stock 2005), shows that it is misleading to conclude that cyclostomes have a preduplication genomic state, simply based on the fewer number of *Dlx* genes known previously. In the *Gsh* family, we identified one lamprey gene, *LjGshA*, that clusters firmly with the hagfish genes (fig. 3*E*), suggesting that these cyclostome genes are orthologous to each other. In the *Cdx* family, our lamprey gene, *LjCdxA*, was placed close to one of the gnathostome paralogs (fig. 3*D*). This indicates that the hagfish *Cdx* gene has evolved too rapidly to reconstruct the clustering with an ortholog in the lamprey or that the lamprey *LjCdxA* gene is paralogous to the reported hagfish *Cdx* gene. If the latter is true, lampreys are expected to possess another *ParaHox* cluster or its remnant in addition to the one already reported for two hagfish species (Furlong et al. 2007).

An advantage of studying these landmark clusters in a phylogenetic context is that orthologies of harbored genes are guaranteed by their conserved genomic structures. Thus, we constrained phylogenetic relationships within gnathostome genes that reside in orthologous gene clusters, to focus only on the relationships between gnathostome and cyclostome paralogs. In our phylogenetic tree inferences including the newly identified lamprey genes, none of the five homeobox-containing gene families supported the hypothesis that genome duplications occurred after the cyclostome–gnathostome split (hypothesis C) (table 1 and fig. 3), which was recently supported by *ParaHox* genes (Furlong et al. 2007). In fact, our analysis has shown that these homeobox-containing gene families do not provide enough resolution to reconstruct the phylogenetic relationships even among gnathostomes and outgroups (supplementary fig. S3, Supplementary Material online). This indicates that conclusions on the timing of genome duplications drawn by analyses of these gene families and from other *Hox* paralogous groups (Force et al. 2002; Irvine et al. 2002; Fried et al. 2003; Stadler et al. 2004; Furlong et al. 2007), which are less divergent than *Hox13* (Ferrier et al. 2000), cannot have sufficient level of reliability. This point strongly demonstrates the need for studies using a more reliable data set.

Evaluation of Phylogenetic Trees

In this study, we adopted the ML method, which facilitates "exhaustive" evaluation of every possible tree

topology in a solid statistical framework (note that the Bayesian method, which has become increasingly popular, is powerful only in "heuristic" tree search; Felsenstein 1981; Holder and Lewis 2003). The ML method is relatively robust even when sequence data contain some degree of rate heterogeneity among lineages (Felsenstein 1978; Philippe et al. 2005), as might be expected in cases involving cyclostome genes, which have often evolved rapidly (Kuraku and Kuratani 2006). To estimate the relative timing of gene duplications, we applied a strategy for counting gene duplications in a probabilistic framework (see Materials and Methods). Unlike deterministic "best tree" approaches (e.g., Escriva et al. 2002; Putnam et al. 2008), from which information concerning alternative tree topologies is discarded, our approach is expected to yield a robust estimate that is not distorted by slight variations in taxon sampling, alignment construction, and model selection and thereby should allow us to extract much more phylogenetic information from the input data. This aspect was especially advantageous in our study where multiple tree topologies tended to be supported with similar log-likelihood values.

Our probabilistic approach produced results similar to those obtained in the total count of discrete numbers of gene duplications in the ML tree and NJ tree (table 3). However—as for each family separately—our approach provides a more realistic overview for the timing of gene duplications on a probabilistic basis. For example, in the ML tree of *RAR* (fig. 5*C*), we must assume three gene duplications before the cyclostome–gnathostome split, instead of two (plus a secondary loss of the gnathostome ortholog of cyclostome *RAR2* and another secondary loss or unidentification of the cyclostome ortholog of gnathostome *RARβ*), to produce three gnathostome paralogs. However, our probabilistic count estimated N_{bef} at 1.82 ± 0.32. In the probabilistic count, the sums for N_{bef}, N_{aft}, and N_{unk} for each gene family resulted in values close to the numbers that can produce the sets of gnathostome paralogs with minimum numbers of gene duplications (supplementary table S4, Supplementary Material online), which fits better parsimonious estimation.

Gene Repertoires in Cyclostome Genomes: How Many Paralogs?

In more than half of the gene families analyzed here (36 out of 55), only one reported cyclostome gene was available. This is thought to be mainly because of incomplete identification of cyclostome genes. Our cDNA survey did not only identify the first members for cyclostomes in some families (*NOTCH* and *fringe*) but also identify additional members to already reported cyclostome genes in other families (*Cdx*, *Gsh*, *Dlx*, *RAR*, *RXR*, and *SLC2A*). Among these genes, we focused on the *RAR* family, where we identified three paralogs in *L. japonicum* (*LjRAR1–LjRAR3*) and performed intensive survey of additional paralogs and their orthologs in *M. mordax* and *E. burgeri*. With all the efforts on these three cyclostome species and different kinds of cDNA examined, we ended up with only these three paralogs (fig. 5*C*). Together with the identification of six *Dlx* genes in *L. japonicum* (supplementary fig. S2, Supplementary Material online), our survey has proved that cyclostomes have more redundant gene repertoires than previously thought. This compels a reconsideration of an idea that attributes fewer gene repertoires in cyclostomes to their primitive genomic status.

Timing of Genome Duplications: before the Cyclostome–Gnathostome Split or after?

Results of our ML analysis for 55 families again showed that these genome duplication events are not easily resolved, as suggested by numbers of tree topologies that could not be rejected by statistical analysis (supplementary fig. S5 and table S3, Supplementary Material online). Our analysis using these gene families provided results consistent with the monophyly of cyclostomes, the 2R pattern of genome duplications, and cyclostome lineage–specific gene duplications (discussed in supplementary text S2, Supplementary Material online).

The majority of the 55 families supported hypothesis A with the highest probabilities, regardless of the assumption of the 2R pattern of genome duplications (table 2). It is notable that hypothesis C was rejected with the highest frequency (table 2A). In this respect, gene families with only two gnathostome paralogs showed a clear result: none of them rejected hypothesis A/B, whereas six out of 21 families rejected hypothesis B/C (table 2B). This suggests that hypothesis C is not likely.

In the probabilistic count of gene duplications, the total number of N_{bef} was significantly larger than that of N_{aft} (table 3; for details of each gene family, see supplementary table S5, Supplementary Material online); as shown in figure 1, if hypothesis B is true, more gene duplications should be detected after the cyclostome–gnathostome split than before, even if we assume that some of the duplicates have been lost secondarily. Moreover, the distributions of N_{aft} did not resemble the unimodal distributions of N_{bef} but were skewed toward zero (fig. 4*B* and *D*). This suggests that a certain portion of N_{aft} originates from statistical fluctuation. Thus, our probabilistic count of gene duplications supports hypothesis A.

Despite the large amount of support for hypothesis A, three of the 55 gene families in our data set explicitly supported hypothesis B or C and rejected hypothesis A (PACSIN 1/2, PTP R2, and PTP RM/RK/RT/RL families; see also supplementary table S3, Supplementary Material online). These might have been caused by a family-specific and small-scale (such as tandem) duplication event, which was not fully excluded at the step of selecting gene families (see Materials and Methods), or by methodological artifacts resulting from an inexhaustive data set, such as incomplete gene identification in cyclostomes. In the latter case, an unidentified cyclostome gene that has evolved less rapidly might cluster with one of multiple gnathostome paralogs. For example, in the *Cdx* family (fig. 3*D*), our lamprey sequence, *LjCdxA*, clustered with gnathostome *Cdx4*, suggesting that a time estimation of gene duplications based solely on one rapidly evolving cyclostome gene (hagfish *Cdx* genes, in this case) is sometimes misleading (for other examples, see supplementary fig. S6*C* and *E*, Supplementary Material online).

For better identification of orthology/paralogy and more precise estimation of the timing of gene duplication, we focused on two gene families, those for visual opsins and *RAR*s, where intensive gene sampling has been performed. In both of these families, we proved that multiple gnathostome paralogs were duplicated through large-scale duplication events (fig. 5*B* and *D*) and that those events occurred before the cyclostome–gnathostome split (fig. 5*A* and *C*). It is unlikely that the scattered placement of multiple cyclostome paralogs in the trees of visual opsins and RAR family were produced by artifact, such as parallel sequence evolution between a gnathostome paralog and a cyclostome paralog. This is an unmistakable evidence for hypothesis A and suggests that the immediate common ancestor of cyclostomes and gnathostomes shared a similar level of redundancy in gene repertoires introduced by a large-scale duplication event.

The gene families shown in table 4 also provide a consistent result that gene duplications occurred before the cyclostome–gnathostome split, although gene sampling in cyclostomes is not thorough. Provided that gene families in the same synteny groups have shared their evolutionary history, this line of evidence suggests that the *Hox* and *Dlx* clusters contained in these synteny groups also duplicated before the cyclostome–gnathostome split, though they themselves did not provide sufficient resolution. Under the assumption that early vertebrates experienced only two rounds of genome duplications as reported (e.g., Dehal and Boore 2005), our analysis suggests that both of these rounds occurred before the cyclostome–gnathostome split.

Conclusions

Our investigation, which exceeded previous studies in the amount of data and precision of methods, yielded no solid evidence of hypotheses previously proposed (hypothesis B and C) but provided consistent evidence that the last common ancestor of the extant vertebrates—including hagfishes and lampreys—emerged subsequent to the genome duplications that can be regarded as a genomic synapomorphy of all extant vertebrates (hypothesis A). Importantly, our approach and our conclusion have not been influenced by controversies over the monophyly of cyclostomes, the patterns of genome duplication (2R or not) and additional genome duplication(s) specific to the cyclostome lineage (see supplementary text S2, Supplementary Materials online). If our conclusion is correct, "primitive" morphological traits in cyclostomes may be attributable to secondary modification of their gene repertoires or changes in gene function and regulation introduced secondarily in either of the gnathostome or cyclostome lineages (see Shigetani et al. 2002, for an example), rather than to incomplete gene repertoires at a preduplication state of cyclostome genomes. Above all, this proposed scenario underlines the need to maintain awareness of the presence of more redundant cyclostome gene repertoires than previously thought. We also advocate caution in determining orthologies between gnathostomes and cyclostomes that might be the bases for the reasonable phylogenetic evaluation of phenotypic transitions from prevertebrates to vertebrates at the molecular level.

Supplementary Material

Supplementary text S1 and S2, data sets S1 and S2, figs. S1–S7, and tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournal.org/). Accession numbers: Sequences identified in this study are deposited in GenBank (URL: http://www.ncbi.nlm.nih.gov/Genbank) under accession numbers AB292620–AB292646, AB368820, and AB368821.

Acknowledgments

We thank S. Moriyama, A. Takahashi, M. Nozaki, H. Kawauchi, K. Tamura, H. Aono, K. G. Ota, N. Iwabe, T. Miyata, and R. Kusakabe for animal tissues or nucleic acids and K. Katoh for critical reading of manuscript and valuable discussions. Our gratitude extends to anonymous reviewers for constructive comments and suggestions.

Literature Cited

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.

Bisbee CA, Baker MA, Wilson AC, Haji-Azimi I, Fischberg M. 1977. Albumin phylogeny for clawed frogs (*Xenopus*). Science. 195:785–787.

Collin SP, Knight MA, Davies WL, Potter IC, Hunt DM, Trezise AE. 2003. Ancient colour vision: multiple opsin genes in the ancestral vertebrates. Curr Biol. 13:R864–R865.

Cracraft J, Donoghue MJ. 2004. Assembling the tree of life. Oxford: Oxford University Press.

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.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3:e314.

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

Duméril AMC. 1806. Zoologie analytique, ou méthode naturelle de classification des animaux, rendue plus facile a l'Aide de Tableaux synoptiques par. Paris: Allais.

Escriva H, Bertrand S, Germain P, Robinson-Rechavi M, Umbhauer M, Cartry J, Duffraisse M, Holland L, Gronemeyer H, Laudet V. 2006. Neofunctionalization in vertebrates: the example of retinoic acid receptors. PLoS Genet. 2:e102.

Escriva H, Manzon L, Youson J, Laudet V. 2002. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. Mol Biol Evol. 19:1440–1450.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool. 27: 401–410.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 17:368–376.

Ferrier DE, Minguillon C, Holland PW, Garcia-Fernandez J. 2000. The amphioxus *Hox* cluster: deuterostome posterior flexibility and Hox14. Evol Dev. 2:284–293.

Force A, Amores A, Postlethwait JH. 2002. *Hox* cluster organization in the jawless vertebrate *Petromyzon marinus*. J Exp Zool. 294:30–46.

Forey P, Janvier P. 1993. Agnathans and the origin of jawed vertebrates. Nature. 361:129–134.

Forey PL. 1984. Yet more reflections on agnathan-gnathostome relationships. J Vert Paleont. 4:330–343.
Fried C, Prohaska SJ, Stadler PF. 2003. Independent *Hox*-cluster duplications in lampreys. J Exp Zool B Mol Dev Evol. 299:18–25.
Furlong RF, Younger R, Kasahara M, Reinhardt R, Thorndyke M, Holland PW. 2007. A degenerate *ParaHox* gene cluster in a degenerate vertebrate. Mol Biol Evol. 24:2681–2686.
Haitina T, Klovins J, Takahashi A, Lowgren M, Ringholm A, Enberg J, Kawauchi H, Larson ET, Fredriksson R, Schioth HB. 2007. Functional characterization of two melanocortin (MC) receptors in lamprey showing orthology to the MC1 and MC4 receptor subtypes. BMC Evol Biol. 7:101.
Hisatomi O, Iwasa T, Tokunaga F, Yasui A. 1991. Isolation and characterization of lamprey rhodopsin cDNA. Biochem Biophys Res Commun. 174:1125–1132.
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.
Holder M, Lewis PO. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet. 4:275–284.
Holland PW, Garcia-Fernandez J, Williams NA, Sidow A. 1994. Gene duplications and the origins of vertebrate development. Dev Suppl. 1994:125–133.
Hufton AL, Groth D, Vingron H, Lehrach AJ, Poustka M, Panopoulou G. Forthcoming. 2008. Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. Genome Res. 18:1582–1591.
Irvine SQ, Cangiano MC, Millette BJ, Gutter ES. 2007. Non-overlapping expression patterns of the clustered *Dll-A/B* genes in the ascidian *Ciona intestinalis*. J Exp Zool B Mol Dev Evol. 308:428–441.
Irvine SQ, Carr JL, Bailey WJ, Kawasaki K, Shimizu N, Amemiya CT, Ruddle FH. 2002. Genomic analysis of *Hox* clusters in the sea lamprey *Petromyzon marinus*. J Exp Zool. 294:47–62.
Jacobs GH. 1993. The distribution and nature of colour vision among the mammals. Biol Rev Camb Philos Soc. 68:413–471.
Kasahara M. 2007. The 2R hypothesis: an update. Curr Opin Immunol. 19:547–552.
Kasahara M, Nakaya J, Satta Y, Takahata N. 1997. Chromosomal duplication and the emergence of the adaptive immune system. Trends Genet. 13:90–92.
Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.
Kawakoshi A, Hyodo S, Nozaki M, Takei Y. 2006. Identification of a natriuretic peptide (NP) in cyclostomes (lamprey and hagfish): cNP-4 is the ancestral gene of the NP family. Gen Comp Endocrinol. 148:41–47.
Kishino H, Miyata T, Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol. 30:151–160.
Koyanagi M, Kawano E, Kinugawa Y, Oishi T, Shichida Y, Tamotsu S, Terakita A. 2004. Bistable UV pigment in the lamprey pineal. Proc Natl Acad Sci USA. 101:6687–6691.
Kuraku S, Kuratani S. 2006. Time Scale for cyclostome evolution inferred with a phylogenetic diagnosis of hagfish and lamprey cDNA sequences. Zool Sci. 23:1053–1064.
Kuraku S, Takio Y, Tamura K, Aono H, Meyer A, Kuratani S. 2008. Noncanonical role of Hox14 revealed by its expression patterns in lamprey and shark. Proc Natl Acad Sci USA. 105:6679–6683.
Larhammar D, Lundin LG, Hallbook F. 2002. The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. Genome Res. 12: 1910–1920.
Lundin LG, Larhammar D, Hallbook F. 2003. Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. J Struct Funct Genomics. 3:53–63.
Maisey JG. 1986. Heads and tails: a chordate phylogeny. Cladistics. 2:201–256.
McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. Nat Genet. 31:200–204.
Meyer A, Zardoya R. 2003. Recent advances in the (molecular) phylogeny of vertebrates. Annu Rev Ecol Evol Syst. 34:311–338.
Miyata T, Suga H. 2001. Divergence pattern of animal gene families and relationship with the Cambrian explosion. Bioessays. 23:1018–1027.
Neidert AH, Virupannavar V, Hooker GW, Langeland JA. 2001. Lamprey *Dlx* genes and early vertebrate evolution. Proc Natl Acad Sci USA. 98:1665–1670.
Nordstrom K, Larsson TA, Larhammar D. 2004. Extensive duplications of phototransduction genes in early vertebrate evolution correlate with block (chromosome) duplications. Genomics. 83:852–872.
Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.
Ota KG, Kuraku S, Kuratani S. 2007. Hagfish embryology with reference to the evolution of the neural crest. Nature. 446:672–675.
Pendleton JW, Nagai BK, Murtha MT, Ruddle FH. 1993. Expansion of the Hox gene family and the evolution of chordates. Proc Natl Acad Sci USA. 90:6300–6304.
Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol. 5:50.
Putnam NH, Butts T, Ferrier DE, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature. 453:1064–1071 (37 co-authors).
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.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18: 502–504.
Sharman AC, Holland PW. 1998. Estimation of *Hox* gene cluster number in lampreys. Int J Dev Biol. 42:617–620.
Shigetani Y, Sugahara F, Kawakami Y, Murakami Y, Hirano S, Kuratani S. 2002. Heterotopic shift of epithelial-mesenchymal interactions in vertebrate jaw evolution. Science. 296: 1316–1319.
Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 17:1246–1247.
Sidow A. 1996. Gen(om)e duplications in the evolution of early vertebrates. Curr Opin Genet Dev. 6:715–722.
Spring J. 1997. Vertebrate evolution by interspecific hybridisation—are we polyploid? FEBS Lett. 400:2–8.
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 Phylogenet Evol. 32:686–694.
Steinke D, Hoegg S, Brinkmann H, Meyer A. 2006. Three rounds (1R/2R/3R) of genome duplications and the evolution of the glycolytic pathway in vertebrates. BMC Biol. 4:16.

Stock DW. 2005. The Dlx gene complement of the leopard shark, *Triakis semifasciata*, resembles that of mammals: implications for genomic and morphological evolution of jawed vertebrates. Genetics. 169:807–817.

Stock DW, Ellies DL, Zhao Z, Ekker M, Ruddle FH, Weiss KM. 1996. The evolution of the vertebrate *Dlx* gene family. Proc Natl Acad Sci USA. 93:10858–10863.

Venkatesh B, Kirkness EF, Loh YH, et al. (12 co-authors). 2007. Survey sequencing and comparative analysis of the elephant shark (*Callorhinchus milii*) genome. PLoS Biol. 5:e101.

Billie Swalla, Associate Editor

Accepted September 27, 2008