Seminars in Cell & Developmental Biology 24 (2013) 119–127

Seminars in Cell & Developmental Biology

journal homepage: www.elsevier.com/locate/semcdb

Review

Impact of asymmetric gene repertoire between cyclostomes and gnathostomes

Shigehiro Kuraku*

Genome Resource and Analysis Unit, Center for Developmental Biology (CDB), RIKEN, 2-2-3 Minatojimaminami-machi, Chuo-ku, Kobe 650-0047, Japan

ARTICLE INFO

Article history:
Available online 3 January 2013

Keywords:
Hagfish
Lamprey
Whole genome duplication
Hidden paralogy
Secondary gene loss

ABSTRACT

Extant vertebrates are divided into the two major groups, cyclostomes and gnathostomes (jawed vertebrates). The former includes jawless fishes, hagfishes and lampreys, and the latter includes all extant jawed vertebrates. In many research fields, the phenotypic traits of the cyclostomes have been considered crucial in understanding the evolutionary process from invertebrates to vertebrates. Recent studies have suggested that the common ancestor of the extant vertebrates including hagfishes and lampreys underwent two-round of whole genome duplications, and thus the genome expansion solely does not account for phenotypic differences between cyclostomes and gnathostomes. Emerging evidence from molecular phylogeny of individual gene families indicates that the gene repertoire expanded at the common ancestor of vertebrates were later reshaped asymmetrically between the two lineages, resulting in the retention of differential gene sets. This also confuses interpretation of conserved synteny which often serves as indicator of orthology and the ploidy level. In this review, current controversy and future perspectives of cyclostome genomics are discussed with reference to evolutionary developmental biology.

Contents

1. Introduction

After agnathan (jawless fish) lineages diverged, one of them evolved into gnathostomes (jawed vertebrates) that now consists of more than 50,000 species [1]. Two of the jawless fish lineages have survived to date and are called hagfishes and lampreys, both classified into Cyclostomata [2]. As the earliest extant lineages of vertebrates, some representatives of cyclostomes, among only around 100 species, have been analyzed in various fields including immunology, endocrinology, neurobiology, and developmental biology. It is often expected that phenotypic traits of cyclostomes could be similar to the ancestral state of vertebrates. Some phenotypic traits of cyclostomes have recently been molecularly reanalyzed, and for example the vertebrae of adult hagfish were revealed to have been secondarily lost in its unique lineage [3]. This evokes caution in interpreting the evolutionary process of phenotypic traits of cyclostomes which have been considered to be ancestral to all extant vertebrates (see [4]). Maximum parsimony, a typical strategy in evolutionary reconstruction, might not hold particularly in this case. Interestingly, recent molecular phylogenetic studies is evoking a similar caution that the maximum parsimony does not

Abbreviations: WGD, whole genome duplication; RAR, retinoic acid receptor; COI, cytochrome oxidase I; Hb, hemoglobin; Mb, myoglobin; GbY, globin Y; GbE, globin E; Cygb, cytoglobin.

* Tel.: +81 78 306 3331.
E-mail address: shigehiro-kuraku@cdb.riken.jp

1084-9521
http://dx.doi.org/10.1016/j.semcdb.2012.12.009

hold in molecular phylogeny of some gene families. This review is focused to provide a synthetic hypothesis that secondary changes introduced uniquely in the cyclostome lineages might be underlying the genetic basis of these animals. Those secondary changes may be preventing clear-cut inference of the status of cyclostomes in the genome expansion as well as orthology assignment between cyclostomes and gnathostomes at the single gene family level.

2. Molecular phylogeny as prerequisite of comparative molecular biology

Molecular phylogenetics is one of the fundamental tools in molecular biology to set up the basis for cross-species comparison – it tells us what gene in one species should be compared with what gene in another species. In search of the molecular basis of homology in a phenotypic trait, it is expected that a homologous molecular program, consisting of the set of orthologs conserved between the species, is underlying. On the other hand, in the quest of the molecular trigger of phenotypic novelty, one is prompted to identify the pair of orthologs whose function is associated with the altered phenotype and then scan it for any alteration in them. Alteration of *cis*-regulation has been the typical target of such investigation [5], while some phenotypic changes could be explained by alteration in protein-coding regions [6]. In both cases, phenotypic evolution is assumed to be triggered by a molecular change involving no gain and loss of genes. Whole genome sequencing is, however, now expanding the possibility to change this classical norm. Gene duplication and loss are rampant during evolution. Before genome-wide sequence information became available, it was technically impossible to identify critical changes in gene number responsible for phenotypic evolution. Nowadays, although still challenging, accumulating sequence information is paving the way to comprehensive genome-wide comparison of gene repertoire between species. For example, recent careful assessment in diverse species is revealing many probable cases of gene loss even within the gnathostome lineage [7–9].

3. Hidden paralogy

Misunderstanding of paralogy as orthology is referred to as 'hidden paralogy' (Fig. 1; [10,11]). Examples documented so far include the zebrafish *Emx3* gene (reviewed in [12]): this gene was originally recognized as the zebrafish ortholog of mammalian *Emx1* genes, but later turned out to be the ortholog of the *Emx3* gene [13,14]. In general, there are several typical factors that can lead to hidden paralogy. First, gene duplication serves as the source of multiple genes between which one should infer orthology or paralogy (Fig. 1). The second factor is the effect of incomplete sequence data set, resulted from secondary gene loss or incomplete sequencing of relevant genes. In an old-fashioned targeted PCR-based strategy, genes with particular features, such as rapid-evolving genes with low sequence conservation, could often escape from the search. Moreover, it is not trivial to identify genes with low expression levels and limited spatiotemporal signatures. This issue remains challenging even with modern approaches involving deep sequencing. Third, incomplete taxon sampling (namely, inclusion of a limited number of species) can increase the chance to cause misidentification of orthology and paralogy.

A hypothetical illustration of hidden paralogy regarding the teleost gene phylogeny is presented in Fig. 2. Phylogenetic relationships between the major teleost fish lineages have been shown with a number of genes [15–17]. It is known that whole genome duplication (WGD) occurred in the basal teleost fish lineage, and thus many single human genes have two orthologs in teleost fish genomes. This event is termed 'third round (3R) whole genome duplication'

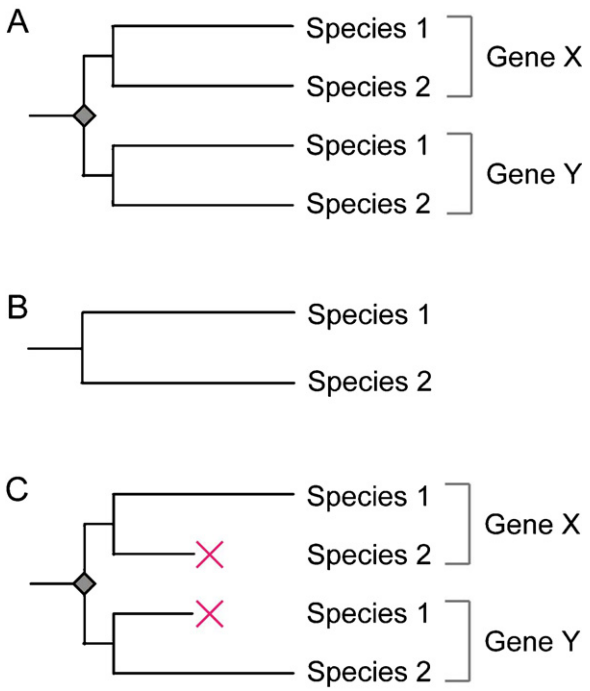

Fig. 1. Schematic illustration of hidden paralogy. (A) Hypothetical situation in which two species (Species 1 and 2) have the same set of genes (Gene X and Y) that were duplicated before the speciation between the two species. (B) Phylogenetic tree without any obvious gene duplication. If only one gene is sampled from each species without exhaustive sampling, they might not be orthologous to each other. (C) Possible explanation of the tree topology in B. Misidentification or loss of Gene Y of Species 1 and Gene X of Species 2 occurred, and thus the situation B represents paralogy between Gene X of Species 1 and Gene Y of Species 2.

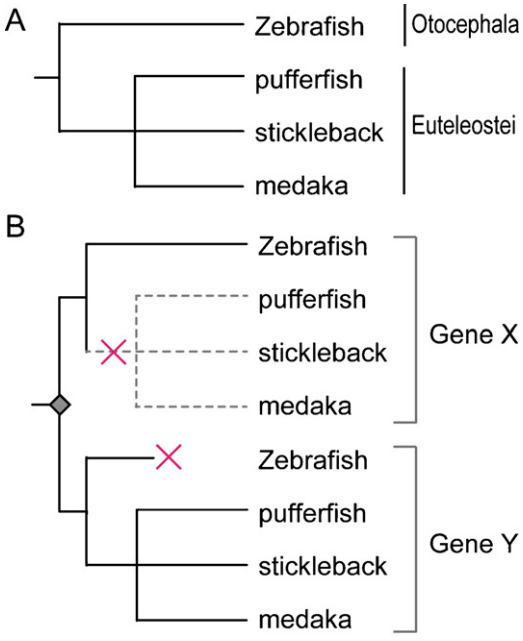

Fig. 2. Possible hidden paralogy in teleost fish gene phylogeny. (A) This gene tree, containing four selected teleost species, is consistent with their relationships in the species tree [15–17]. (B) The hypothetical tree showing the possibility that gene loss occurred in an asymmetric pattern between Otocephala and Euteleostei, and that the zebrafish gene is paralogous to the genes from the other three species in Euteleostei.

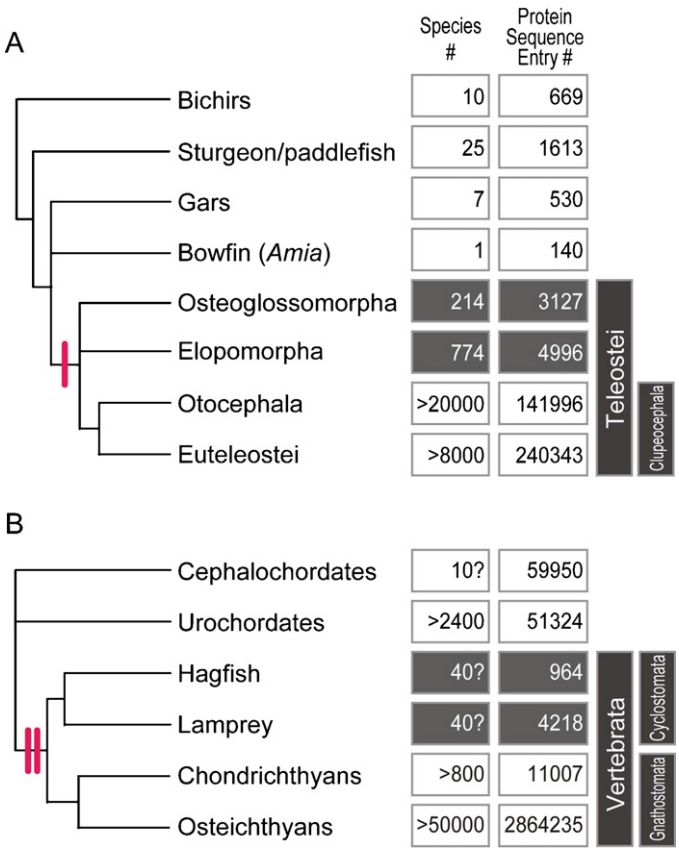

Fig. 3. Two study systems in vertebrates involving whole genome duplication. (A) Actinopterygian fish evolution with the third round (3R) genome duplication in the teleost fish lineage. Phylogenetic relationships of the included fish groups are based on previous publications [20,21,60,61]. (B) Chordate evolution with the two-round (2R) genome duplications early in the vertebrate lineage. Phylogenetic relationships of the included species are based on previous publications [61–64]. Species numbers included in (A) and (B) are based on [1]. The numbers of the protein-coding sequence entries are based on NCBI Protein database (as of November 1, 2012). In both A and B, lineages that diverged after the WGD with less molecular sequence information are shown in grey. Molecular phylogenetic analyses including genes from these lineages require special attention to possible hidden paralogy (see text).

or 'teleost-specific genome duplication (TSGD)' [18,19]. Given that molecular phylogeny is inferred with genes on the nuclear genome that are prone to duplications, its tree topology which happens to be consistent with the relationship among the species included (Fig. 2A) should not be readily interpreted as evidence of orthology. It is because an asymmetric pattern of post-duplication gene loss between the Otocephala (including zebrafish) and Euteleostei (including medaka, stickleback and pufferfish) might have resulted in the tree topology consistent with their species phylogeny (Fig. 2B). In reality, hidden paralogy could not frequently occur among this set of species, for there was likely a relatively long evolutionary period (>50 million years) between the 3R genome duplication event and the Otocephala–Euteleostei split [15]. It is likely that most genes duplicated in the 3R genome duplication could have fixed their fates before the Otocephala–Euteleostei split, and that the genes that survived until this time point should likely have been retained in both lineages. In fact, there are a few extant lineages (Osteoglossomorpha and Elopomorpha) that diverged before the Otocephala–Euteleostei split [20,21]. These lineages might exhibit unique patterns of gene retention in comparison with the Clupeocephala lineage (Fig. 3A). This study system including teleost fishes serves as a simplified model, analogous to the system involving cyclostomes. However, the latter offers more challenging questions as the genome expansion was achieved through two-round process and occurred in an older age.

4. Factors making cyclostome gene phylogeny problematic

4.1. 'Duplication before speciation' at the genomic scale

After the invertebrate chordate lineages, namely cephalochordates and urochordates diverged, two round (2R) of whole genome duplications occurred in the early vertebrate lineage ([22,23]; also see papers accompanying in this issue). It is widely accepted that this genome expansion was completed before the chondrichthyan lineage diverged [24,25]. The major controversy is whether it was already completed before the cyclostome lineage diverged. In 2009, the scenario supporting 'post-2R cyclostome' (Fig. 3B) was proposed based on a systematic molecular phylogenetic analysis on 55 individual gene families [26]. Of those, the color opsin gene family and retinoic acid receptor (RAR) gene family provided evidence of the 'post-2R cyclostome' scenario. It was remarkable that gene sampling in cyclostome species was thoroughly performed for these gene families, which yielded more information for the phylogenetic analysis at higher resolution.

Regardless of the exact timing of the genome expansion, the expansion of gene repertoire was genome-wide, and this study system thus provides an enormously intricate natural experiment in which simultaneous modifications in a number of individual gene families could have contributed to phenotypic evolution. One technical issue in dissecting the effect of the whole genome duplication is the erroneous inclusion of concomitant small-scale duplications (e.g., tandem duplication and retroposition). In the previous study [26], gene families including multiple members that duplicated in early vertebrate evolution and are located on the same chromosome were excluded from the data set. Considering chromosome rearrangements in a long subsequent evolutionary period, traces of tandem duplications might have been erased for many genes. As our attention can only reach the traces retained in the genome of extant species, it is crucial to detect secondary changes introduced after the genome expansion and adequately reconstruct the evolutionary process.

4.2. Differential gene loss and gain resulting in asymmetric gene repertoire

The second factor introduced above as typical cause of hidden paralogy is the contraction and further expansion of gene repertoire once expanded in the ancient duplication. Although loss of a duplicate gene has been formulated as one of the typical fates after gene duplication [27], there are not many reports about the detection of gene loss because it requires fine-scale genome-wide confirmation and cross-species comparison. Currently, the limited sequence resources for cyclostomes do not allow conclusive studies. Genome sequencing for the sea lamprey (*Petromyzon marinus*) was performed but did not use the germ-line genomic DNA which should be free from the programmed loss of DNA recently documented [28]. Moreover, the output of this genome sequencing project, publicly available at the Ensembl database, covers only about 10,000 protein-coding genes. This extremely small number for a vertebrate genome (see [29] for typical gene numbers) indicates that the divergent nature of protein-coding genes in this genome [30] possibly caused a massive number of false negatives in the automated gene prediction pipeline.

However, currently available information for some gene families provides marked tendency. In the globin gene family, gene sampling for both hagfishes and lampreys have been performed intensively since the dawn of molecular evolutionary studies

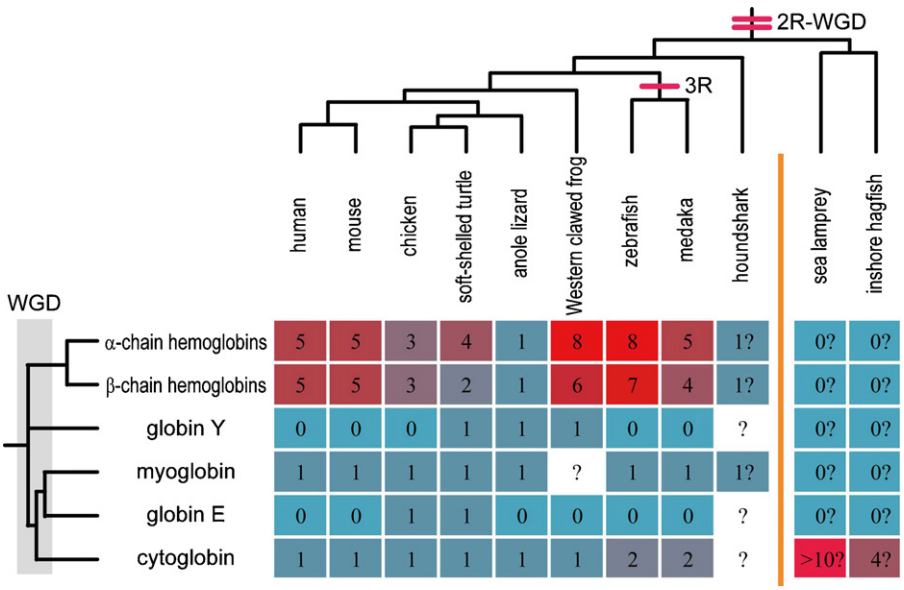

Fig. 4. Asymmetric repertoire of vertebrate globin genes. The relationship between the different globin subtypes is shown on the left, following the previous publication [65]. Based on the molecular phylogeny inferred in the previous study by Hoffmann et al. [38], cyclostome genes identified to date are considered orthologous to gnathostome cytoglobins. Numbers of the genes categorized in different globin subtypes shown in the colored boxes are based on the information in Ensembl Genome Browser (http://www.ensembl.org/), except for the inshore hagfish and houndshark whose sequence entries are found only in GenBank. The boxes are colored with intensities proportional to the gene number. It should be noted that the split between α- and β-hemoglobins was caused by tandem duplication after the 2R whole genome duplication whose trace is retained in the genomes of amphibians and actinopterygian fishes [66]. *Abbreviations*: WGD; whole genome duplication; 2R; two-round; 3R; third round WGD.

[31–33]. The members of this gene family identified in gnathostomes are divided into hemoglobins (Hb; oxygen transporter in red blood cells), myoglobin (Mb; oxygen storage protein in muscle), and other subtypes including cytoglobin (Cygb), globin Y (GbY) and globin E (GbE). Based on their chromosomal location in some gnathostome genomes, it was suggested that the 2R whole genome duplication resulted in four synteny blocks, three of which contain Hb, Mb, and Cygb, respectively [34]. The family members identified for hagfish and lamprey were initially recognized as orthologous to gnathostome Hb [35,36] or Mb [32], but were later suggested to be paralogous to both of them [37], presenting another example of hidden paralogy. To date, four hagfish globin genes and six lamprey ones at most in a single species have been identified, and are placed most closely to gnathostome Cygb [38] (Fig. 4). The Ensembl database reports more than 10 lamprey genes that seem to have been duplicated uniquely in this lineage (Fig. 4). It seems that the cyclostome globin genes are orthologous to gnathostome Cygb genes and underwent explosive lineage-specific gene duplication after the cyclostome–gnathostome split. Although not conclusive because of the incomplete genome sequencing, there might be no ortholog of Hb, Mb, GbE, or GbY retained in the cyclostome genomes (Fig. 4). In summary, while information for cartilaginous fish is still limited, the globin gene family exhibits remarkable asymmetry of gene repertoire between gnathostomes and cyclostomes. The pattern marked with secondary changes of gene repertoire in the cyclostome lineage is seen in some other gene families (e.g., [39]).

In contrast to the globin gene family, the opsin and RAR gene families mentioned above does not exhibit marked asymmetry. In these families, gene sampling from cyclostome species was similarly intensively performed, but each cyclostome duplicate tend to show one-to-one orthology to a gnathostome duplicate [26,40]. It is a subject of future investigation to reveal how frequently the asymmetric pattern of duplicate retention is observed when reliable genome-wide comparison between cyclostome and gnathostome becomes feasible.

4.3. Poor diversity of extant cyclostome species

Poor taxon sampling was raised above as the third factor that can confuse orthology inference. In other words, inclusion of more species could allow retrieval of the genes that later underwent loss or duplication in another particular lineage. In the gnathostome lineage, to overview the entire protein-coding landscape, we have an access to a mass of information from dozens of species with sequenced genomes. In contrast, cyclostomes inherently have much lower species diversity (≈100 species; Fig. 3), only one of which was subjected to genome sequencing. There is no cosmopolitan cyclostome species, and therefore molecular sequence information concentrates on several species endemic to the Northern Hemisphere. Fig. 5 shows a molecular phylogenetic tree of the mitochondrial *cytochrome oxidase I* (*COI*) gene for cyclostomes. The only hagfish species endemic to the Southern Hemisphere in this tree, *Notomyxine tridentiger* and *Myxine australis* (southern hagfish), are placed inside the clade comprised of Northern Hemisphere species of the genera *Eptatretus* and *Myxine* (Fig. 5). Importantly, the entire diversity of modern hagfishes is estimated to have been achieved within only 100 million years. Among lampreys, Southern Hemisphere species of the genera *Mordacia* and *Geotria* are placed phylogenetically outside the Northern Hemisphere clade (Fig. 5), as previously shown [41]. Lamprey species endemic to the Northern Hemisphere seem to have diversified later in a short period of time, as suggested in a previous study ([41]; Fig. 5). Overall, although some key species, especially those in the Southern Hemisphere, remain to be explored for more molecular data, the species diversity in extant cyclostomes and its evolutionary time scale seem to leave long branches within which no intermediate state can be captured by any extant species for both hagfish and lamprey lineages.

Interestingly, relatively small numbers of extant species and molecular sequences available in sequence databases are also the case with the early-branching teleost lineages, Osteoglossomorpha and Elopomorpha (Fig. 3A). This aspect, together with their divergence shortly after the 3R WGD, should make it similarly difficult to

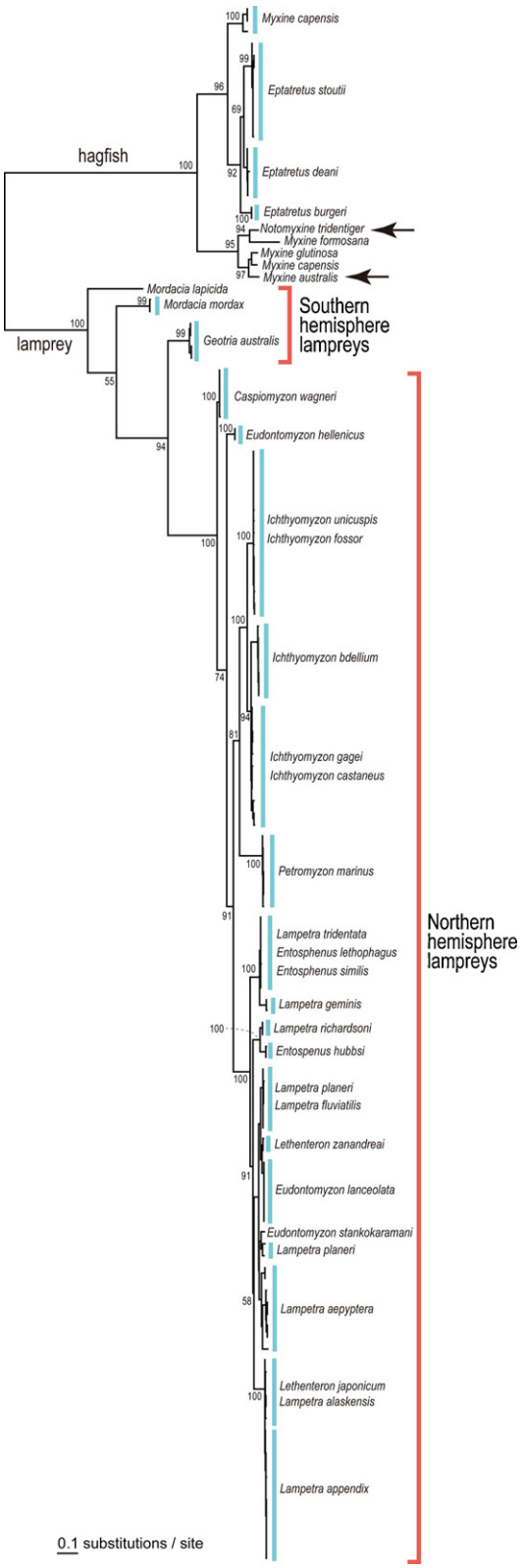

Fig. 5. Molecular phylogeny of *cytochrome oxidase I* (*COI*) gene covering the entire cyclostome clade. The tree was inferred based on a nucleotide sequence dataset (616 nt) with the maximum-likelihood method using the program PhyML [67] assuming the GTR + Γ_4 model. The monophyletic groups of hagfish and lamprey are rooted by each other. The included sequences were derived from several DNA barcoding projects [68,69] and retrieved from the NCBI Nucleotide database. Sequences with ambiguous nucleotide bases or with largely truncated ends were excluded from the alignment built with MAFFT [70]. A blue bar indicates a monophyletic group. When a monophyletic group contains multiple species intermingled with each other, the species names are listed. It should be noted that species identification is not necessarily correct, and thus that it possibly causes polyphyly of some species (e.g., *Ichthyomyzon gagei*, *Lampetra planeri*). *Lethenteron japonicum* has been renamed into *Lethenteron camtschaticum* in the NCBI Taxonomy database. One finding in this tree is that *Myxine australis* and *Notomyxine tridentiger*, two Southern Hemisphere hagfish species with no other molecular sequence data (indicated with arrows), is nested within the clade consisting of two Northern Hemisphere genera, *Myxine* and *Eptatretus*.

confidently infer orthology of genes from these two teleost lineages (e.g., [42,43]).

5. Impact on the relationship between hagfish, lamprey and jawed vertebrates

In relation to whole genome duplication, this review is originally aimed to discuss molecular phylogeny of gene families, rather than species phylogeny. Still, the discussion is tightly associated with the dispute over the relationships between hagfish, lamprey and gnathostomes. Morphological examination usually supported the closer relationship of lampreys with gnathostomes [44,45] (Fig. 6A), whereas molecular data tend to support the hagfish-lamprey grouping (Fig. 6C; reviewed in [46]). Given that genes in the nuclear genome are employed in the analysis aimed to address the question regarding the species phylogeny, hidden paralogy can mislead our inference. For example, the gene tree supporting the hagfish-lamprey grouping (Fig. 6C) does not readily support the monophyly of cyclostomes in the species tree, taking into account the possibility of differential loss of duplicates, based on the different species tree (Fig. 6D). As the majority of gene families have suggested the tree topology in Fig. 6C, the results of some gene families supporting the topology in Fig. 6A should be explained by the differential loss of duplicates (Fig. 6B).

The modern strategy to infer species phylogeny at the genomic scale, so-called 'phylogenomics', have been applied to address some longstanding questions regarding species phylogeny (e.g., [47]). Because of the lack of large-scale sequence resource particularly of hagfish, the dispute over the relationship of cyclostomes has not been a main subject in such a study (see [48]). Even when the phylogenomic data set becomes available, the closest attention should be paid to hidden paralogy in it, in order not to mislead the conclusion.

6. Impact on synteny analysis: how can the whole genome sequencing help?

Patterns of gene location on duplicated genomic regions (termed 'conserved synteny') often provide clear-cut traces of whole genome duplication (e.g., [49–51]): simple count of duplicated genomic regions indicates the ploidy level (see [52]). For example, conserved synteny in teleost fish genomes provided firm evidence of additional WGD in its basal lineage [19]. Moreover, conserved synteny can provide crucial information for orthology assignment (e.g., [53]): the similarity in the pattern of gene retention in a synteny block conserved between species is expected to indicate orthology. For example, the orthology between mouse HoxB cluster and chicken HoxB cluster is inferred by their common lack of paralog groups 10, 11 and 12 [54]. Therefore, whole genome sequencing, producing a huge amount of synteny information, has been expected to address the long-standing question regarding the timing of WGD and cyclostome–gnathostome gene orthology in individual gene families.

However, cyclostomes could offer problems in practicing this genome-scale expectation as well as in the small-scale demand of addressing molecular phylogeny. Possible results of synteny analysis involving a cyclostome genome are illustrated in Figs. 7 and 8.

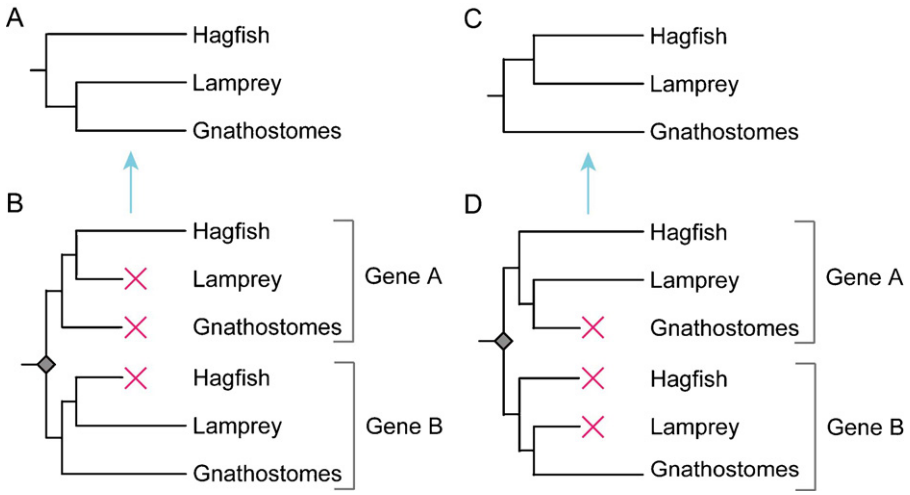

Fig. 6. Impact of hidden paralogy on the relationships between hagfish, lamprey and gnahostomes. This hypothetical figure depicts the caution in interpreting the molecular phylogeny of gene families prone to gene loss. (A) Phylogenetic tree suggesting the closer relationship of lamprey to gnathostomes than to hagfish. (B) The tree topology in (A) may be explained by hidden paralogy. Cyclostomes show monophyletic relationship, and each of the hagfish, lamprey and gnathostome lineages lost one gene (indicated with 'X') after gene duplication (grey diamond). (C) Tree suggesting the closer relationship of lamprey to hagfish than to gnathostomes, that is, monophyly of cyclostomes. (D) The tree topology in C may be explained by asymmetric loss of the duplicates, on the basis of the paraphyletic relationships of cyclostomes. As a consequence of the asymmetric gene loss, the resultant tree topology erroneously suggests the relationship inconsistent with species tree. In order to avoid such misidentification, it is advised to perform thorough gene sampling, at least in gnathostome species whose sequence resources are abundant, including cartilaginous fish which diverged before the relevant genes get lost.

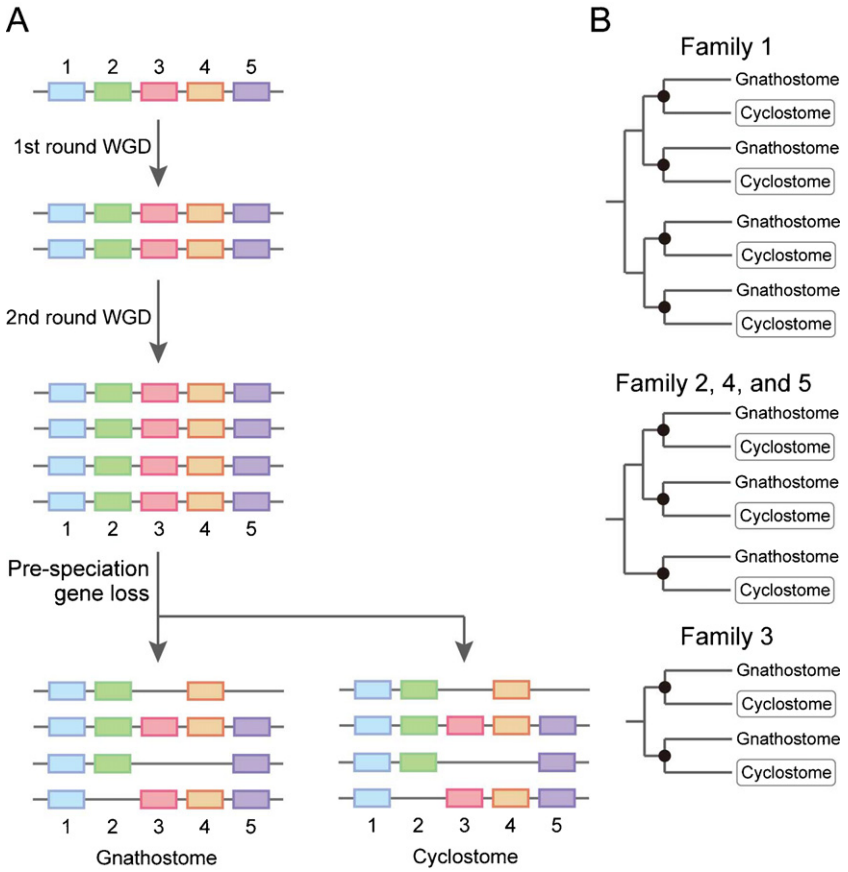

Fig. 7. Hypothesized situation in which synteny analysis can help orthology assignment. (A) Evolutionary scenario in which the cyclostome lineage diverged after 2R WGDs and following gene loss. Consequently, gene repertoire and thus synteny compositions are identical between gnathostome and cyclostome. In this case, synteny analysis can address the question about the timing of whole genome duplication and helps orthology identification of gene arrays and also individual genes, between gnathostome and cyclostome. (B) Deduced molecular phylogenetic trees for Family 1–5 in A. These trees show pairs of one-to-one orthologs (black circles on the node) between gnathostome and cyclostome, which is seldom observed in reality.

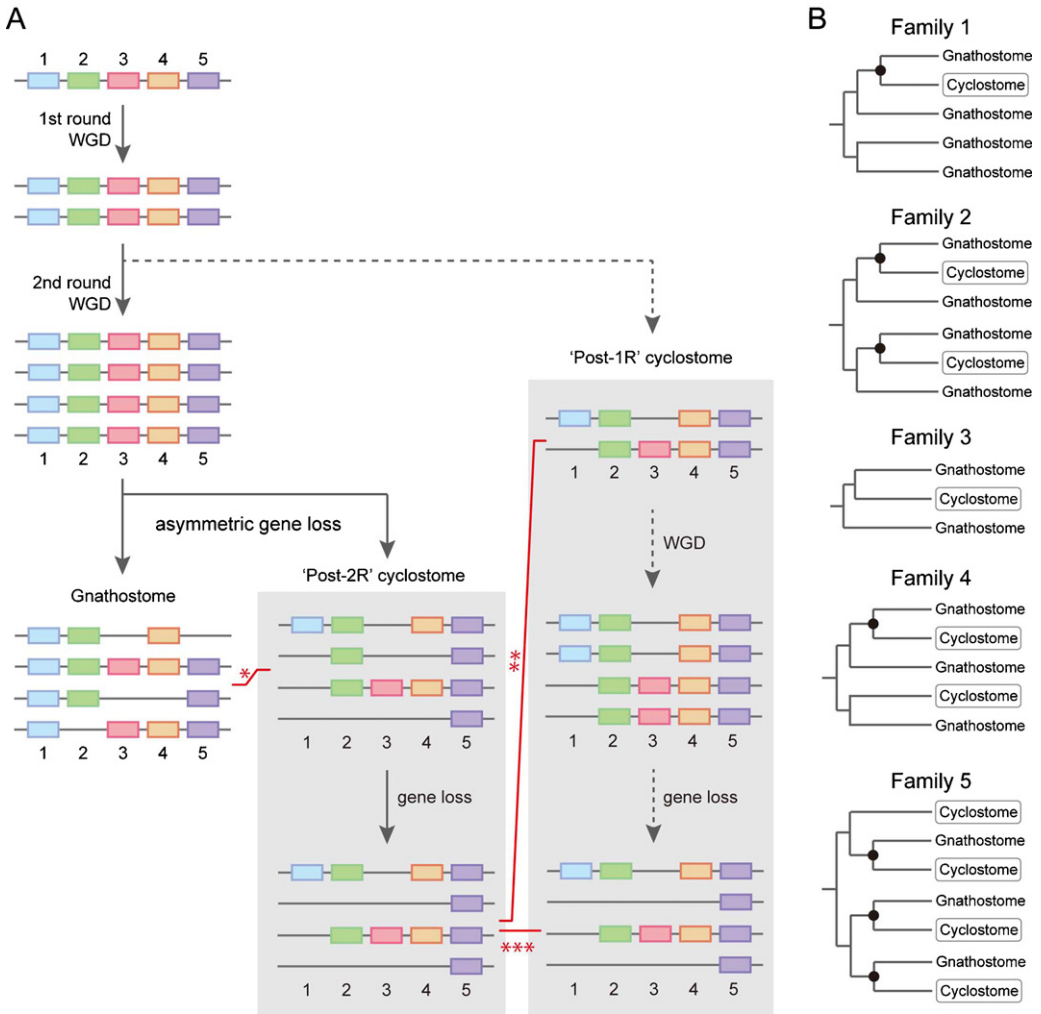

Fig. 8. Hypothesized situation in which asymmetric gene retention confuses orthology assignment. (A) Alternative evolutionary scenarios about synteny conservation according to different timings of the cyclostome divergence. If the cyclostome lineage diverged after 2R WGD but before the fate of the duplicates become fixed ('Post-2R cyclostome'), the cyclostome synteny blocks would have little similarity in gene composition to gnathostome counterparts ('*'), because asymmetric gene loss between the two subsequent lineages could have erased the signatures of orthologous gene arrays. Remarkably, an alternative scenario that the cyclostome diverged between the first and second round of WGD ('Post-1R cyclostome') could result in a similar or identical pattern to the consequence of the former scenario ('Post-2R cyclostome'), as indicated by the comparisons with '**' or '***'. The only clue to distinguishing these two would be the gene repertoire and their phylogenetic relationships in Family 5 which shows little asymmetry in gene repertoire. This indicates that under the assumption of asymmetric gene loss it is difficult to infer the timing of the divergence of the cyclostome lineage and reliably assign orthology between multiple syntenic regions of gnathostome and cyclostome. (B) Deduced molecular phylogenetic trees of Family 1–5 in (A). It is of note that the cyclostome–gnathostome clades without black circles on the node are occurring because of hidden paralogy (see Section 3). It is also remarkable that the sampled number of cyclostome genes is smaller than that of gnathostome counterparts that could be sampled in far more diverse species. In reality, the difficulty would be even further enhanced than illustrated in this figure by secondary rearrangement of gene order and incompatibility of conventional phylogenetic reconstruction method to lineage-specific sequence properties, for example, in lampreys [30].

If the cyclostome–gnathostome split occurred after the pattern of gene loss is established (Fig. 7A), the gene composition should be identical or similar between cyclostome and gnathostome genomes for each syntenic region. This should likely allow clear-cut assignment of orthologous clusters. The deduced topologies of molecular phylogenetic trees for the hypothetical gene families involved in this case are depicted in Fig. 7B, but in reality it is rare to be able to retrieve one-to-one relationships between cyclostome and gnathostome, to the author's knowledge, except for the opsin and RAR gene families [26].

In contrast, even under the 'post-2R cyclostome' condition, if the cyclostome and gnathostome lineages underwent asymmetric gene loss (Fig. 8A), it could be difficult to detect similarity in gene composition in any pair of cyclostome and gnathostome synteny regions (an asterisk in Fig. 8A). Especially if the cyclostome lineage was to undergo more frequent gene loss, its resultant genome could even exhibit similar gene composition to what is expected with the cyclostome divergence before the 2R WGD (double asterisks in Fig. 8A). The similarity could further increase, if additional WGD, previously suggested [55–57], occurred uniquely in the cyclostome lineage, followed by loss of the duplicates (triple asterisks in Fig. 8A). Whichever hypothetical scenario is closer to the real evolutionary history, the distinction between these alternative scenarios in Fig. 8 is not evident. Importantly, the key to inferring the ploidy level and identifying orthologous syntenic regions should be exhaustive gene sampling, as indicated by 'Family 5' in Fig. 8 with the almost full set of the duplicates (also see below).

As an attempt to infer orthology using synteny information, the globin phylogeny, mentioned above (see Section 4.2 and Fig. 4), is here revisited. Molecular phylogenetic analysis previously suggested orthology of the cyclostome globins to the gnathostome Cygb ([38]; Fig. 4), but can this be confirmed by synteny information? A typical procedure is to identify genes localized closely to

Cygb in multiple gnathostome genomes, search for the orthologs of the neighboring genes in a cyclostome genome, and verify the linkage between them and the cyclostome globin genes, the possible orthologs of the gnathostome Cygb. This attempt, however, seems infeasible, primarily because the identification of cyclostome orthologs of the neighboring genes would not be unambiguously performed for the reason mentioned above concerning asymmetric gene loss (see Section 4.2). Because the expansion of the gene repertoire encompassed the whole genome, the ambiguity of orthology caused by the possible asymmetric gene retention (illustrated in Fig. 8) might also be genome-wide and thus synteny block-wide. Obviously, the typical strategy to refer to orthology of neighboring genes should not be reliable, if the orthology of the neighboring genes itself is ambiguous. Moreover, the multiplicity of the cyclostome orthologs of Cygb would not allow clear-cut comparison of syntenic regions, and the linkage of the genes in the different gene families could not be verified because of secondary chromosomal rearrangement in the cyclostome lineage or incomplete sequencing of the genomic region of interest.

In general, synteny analyses have provided reliable clues to molecular phylogeny, but in this particular case involving hitherto unseen challenges, we are urged to adopt an alternative strategy. In the previous study, molecular phylogeny of the color opsin gene family served as the major evidence for the 'post-2R cyclostome' scenario [26]. As mentioned above, this gene family allowed exhaustive gene sampling in lampreys and also exhibited their robust one-to-one orthology to the gnathostome counterparts (exactly as 'Family 5′ in Fig. 8). From the viewpoint of comparative physiology, it is intriguing to understand why this gene family exhibits the clear-cut cyclostome–gnathostome orthology [58]. It is possible that the different opsin subtype genes, duplicated in WGD, rapidly became functionally differentiated. They probably accumulated subtype-specific signatures in their sequences, shortly after the WGD, which now serves as distinct signals in phylogenetic reconstruction. Those genes should likely be retained in the genome for a long time because of their critically differentiated roles. In extracting significant information from whole genome sequencing of cyclostomes, a promising strategy, alternative to synteny analysis, could be the genome-wide search of more gene families with such distinct phylogenetic signals.

7. Perspectives

To dissect problems in understanding genome evolution in early vertebrates, three factors: (1) whole genome duplication, (2) asymmetric gene repertoire and (3) poor taxon sampling, has been presented as systematic problems in cyclostome gene phylogeny. Remarkably, the situation is parallel with the other study system of early teleost fish evolution. However, addressing questions associated with the early vertebrate evolution should be more challenging, given the more ancient timing of the genome expansion and the fact that the expansion was achieved through two successive events.

The author's teams have analyzed molecular phylogeny of a number of gene families and have supported 'cyclostome monophyly' and 'post-2R cyclostome'. However, it should be admitted that some gene families exhibit incongruent results. For example, previous studies on the Emx [57] and Irx gene families [59] concluded that WGD occurred in the gnathostome lineage after the split of the cyclostome lineage. For both lines of discussion, namely regarding species tree and gene tree, it is possible that hidden paralogy accounts for the incongruences, and exhaustive gene sampling involving more cyclostome species is urgent for reanalysis.

Apart from the interpretation of molecular phylogeny, the possibly frequent asymmetric gene retention presents an interesting biological insight. First, some of the phenotypic differences between cyclostomes and gnathostomes (for example, the jaw) could be explained by differential loss of genes duplicated in WGD. This could be extended to explain the difference between hagfish and lamprey (for example, the loss of vertebrae in adult hagfish). Second, even if there is no marked difference in a particular phenotypic trait between cyclostomes and gnathostomes (for example, neural crest cell differentiation), there could be a non-orthologous set of genes which resulted from asymmetric gene retention and is responsible for such homologous phenotypic traits. One example presented above which falls into this category is oxygen transport in blood, which is achieved by paralogous genes between cyclostomes and gnathostomes [38].

As scrutinized above (see Section 6), large-scale synteny information probably obtained in cyclostome genome sequencing might not readily provide determinative clues to the timing and mode of WGDs, given that the effect of asymmetric gene repertoire is dominant throughout the cyclostome genomes. More vast effort is anticipated to fill the crucial gap of molecular studies (manifested in the amount of sequence data in Fig. 3B) for better understanding of chordate evolution.

Acknowledgements

I thank Jay Storz and Guojun Shen for insightful discussion about globin gene phylogeny. I apologize to many colleagues whose work I could not discuss due to space limitations. I am grateful to members of my team and colleagues I had in Konstanz, Germany. This study is supported by Center for Developmental Biology, RIKEN.

References

[1] Nelson JS. Fishes of the world. 4th ed. Hoboken, NJ: John Wiley; 2006.
[2] Kuraku S, Ota KG, Kuratani S. Cyclostomata. In: Kumar S, Hedges B, editors. Timetree of life. 2009.
[3] Ota KG, Fujimoto S, Oisi Y, Kuratani S. Identification of vertebra-like elements and their possible differentiation from sclerotomes in the hagfish. Nature Communications 2011;2:373.
[4] Janvier P. Evolutionary biology: born-again hagfishes. Nature 2007;446:622–3.
[5] Carroll SB. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 2008;134:25–36.
[6] Hoekstra HE, Coyne JA. The locus of evolution: evo devo and the genetics of adaptation. Evolution 2007;61:995–1016.
[7] Feiner N, Begemann G, Renz AJ, Meyer A, Kuraku S. The origin of bmp16, a novel Bmp2/4 relative, retained in teleost fish genomes. BMC Evolutionary Biology 2009;9:277.
[8] Feiner N, Ericsson R, Meyer A, Kuraku S. Revisiting the origin of the vertebrate Hox14 by including its relict sarcopterygian members. Journal of Experimental Zoology Part B: Molecular Developmental Biology 2011;316:515–25.
[9] Manousaki T, Feiner N, Begemann G, Meyer A, Kuraku S. Co-orthology of Pax4 and Pax6 to the fly eyeless gene: molecular phylogenetic, comparative genomic, and embryological analyses. Evolution and Development 2011;13:448–59.
[10] Daubin V, Gouy M, Perriere G. Bacterial molecular phylogeny using supertree approach. Genome Informatics 2001;12:155–64.
[11] Martin AP, Burg TM. Perils of paralogy: using HSP70 genes for inferring organismal phylogenies. Systematic Biology 2002;51:570–87.
[12] Kuraku S. Palaeophylogenomics of the vertebrate ancestor – impact of hidden paralogy on hagfish and lamprey gene phylogeny. Integrative and Comparative Biology 2010;50:124–9.
[13] Derobert Y, Plouhinec JL, Sauka-Spengler T, Le Mentec C, Baratte B, Jaillard D, et al. Structure and expression of three Emx genes in the dogfish *Scyliorhinus canicula*: functional and evolutionary implications. Developmental Biology 2002;247:390–404.
[14] Kawahara A, Dawid IB. Developmental expression of zebrafish emx1 during early embryogenesis. Gene Expression Patterns 2002;2:201–6.
[15] Azuma Y, Kumazawa Y, Miya M, Mabuchi K, Nishida M. Mitogenomic evaluation of the historical biogeography of cichlids toward reliable dating of teleostean divergences. BMC Evolutionary Biology 2008;8:215.
[16] Steinke D, Salzburger W, Meyer A. Novel relationships among ten fish model species revealed based on a phylogenomic analysis using ESTs. Journal of Molecular Evolution 2006;62:772–84.
[17] Li C, Lu G, Orti G. Optimal data partitioning and a test case for ray-finned fishes (Actinopterygii) based on ten nuclear loci. Systematic Biology 2008;57:519–39.
[18] Meyer A, Van de Peer Y. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays 2005;27:937–45.

[19] Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, et al. Zebrafish hox clusters and vertebrate genome evolution. Science 1998;282:1711–4.
[20] Inoue JG, Miya M, Tsukamoto K, Nishida M. A mitogenomic perspective on the basal teleostean phylogeny: resolving higher-level relationships with longer DNA sequences. Molecular Phylogenetics and Evolution 2001;20:275–85.
[21] Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP, et al. Resolution of ray-finned fish phylogeny and timing of diversification. Proceedings of the National Academy of Sciences of the United States of America 2012;109:13698–703.
[22] Holland PW, Garcia-Fernandez J. Hox genes and chordate evolution. Developmental Biology 1996;173:382–95.
[23] Sidow A. Gen(om)e duplications in the evolution of early vertebrates. Current Opinion in Genetics and Development 1996;6:715–22.
[24] Venkatesh B, Kirkness EF, Loh YH, Halpern AL, Lee AP, Johnson J, et al. Survey sequencing and comparative analysis of the elephant shark (*Callorhinchus milii*) genome. PLoS Biology 2007;5:e101.
[25] Miyata T, Suga H. Divergence pattern of animal gene families and relationship with the Cambrian explosion. Bioessays 2001;23:1018–27.
[26] Kuraku S, Meyer A, Kuratani S. Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after. Molecular Biology and Evolution 2009;26:47–59.
[27] Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 1999;151:1531–45.
[28] Smith JJ, Antonacci F, Eichler EE, Amemiya CT. Programmed loss of millions of base pairs from a vertebrate genome. Proceedings of the National Academy of Sciences of the United States of America 2009;106:11212–7.
[29] Pertea M, Salzberg SL. Between a chicken and a grape: estimating the number of human genes. Genome Biology 2010;11:206.
[30] Qiu H, Hildebrand F, Kuraku S, Meyer A. Unresolved orthology and peculiar coding sequence properties of lamprey genes: the KCNA gene family as test case. BMC Genomics 2011;12:325.
[31] Goodman M. Globin evolution was apparently very rapid in early vertebrates: a reasonable case against the rate-constancy hypothesis. Journal of Molecular Evolution 1981;17:114–20.
[32] Romero-Herrera AE, Lieska N, Nasser S. Characterization of the myoglobin of the lamprey *Petromyzon marinus*. Journal of Molecular Evolution 1979;14:259–66.
[33] Ohno S, Morrison M. Multiple gene loci for the monomeric hemoglobin of the hagfish (*Eptatretus stoutii*). Science 1966;154:1034–5.
[34] Hoffmann FG, Opazo JC, Storz JF. Whole-genome duplications spurred the functional diversification of the globin gene superfamily in vertebrates. Molecular Biology and Evolution 2012;29:303–12.
[35] Wald G, Riggs A. The hemoglobin of the sea lamprey, *Petromyzon marinus*. Journal of General Physics 1951;35:45–53.
[36] Bannai S, Sugita Y, Yoneyama Y. Studies on hemoglobin from the hagfish *Epatatretus burgeri*. Journal of Biological Chemistry 1972;247:505–10.
[37] Katoh K, Miyata T. Cyclostome hemoglobins are possibly paralogous to gnathostome hemoglobins. Journal of Molecular Evolution 2002;55:246–9.
[38] Hoffmann FG, Opazo JC, Storz JF. Gene cooption and convergent evolution of oxygen transport hemoglobins in jawed and jawless vertebrates. Proceedings of the National Academy of Sciences of the United States of America 2010;107:14274–9.
[39] Fujimoto S, Oisi Y, Kuraku S, Ota KG, Kuratani S. Non-parsimonious evolution of the Dlx genes in the hagfish. BMC Evolutionary Biology, in press.
[40] Collin SP, Knight MA, Davies WL, Potter IC, Hunt DM, Trezise AE. Ancient colour vision: multiple opsin genes in the ancestral vertebrates. Current Biology 2003;13:R864–5.
[41] Kuraku S, Kuratani S. Time scale for cyclostome evolution inferred with a phylogenetic diagnosis of hagfish and lamprey cDNA sequences. Zoological Science 2006;23:1053–64.
[42] Henkel CV, Burgerhout E, de Wijze DL, Dirks RP, Minegishi Y, Jansen HJ, et al. Primitive duplicate Hox clusters in the European eel's genome. PLoS ONE 2012;7:e32231.
[43] Ogino Y, Katoh H, Kuraku S, Yamada G. Evolutionary history and functional characterization of androgen receptor genes in jawed vertebrates. Endocrinology 2009;150:5415–27.
[44] Janvier P. Early vertebrates. New York: Oxford University Press; 1996.
[45] Near TJ. Conflict and resolution between phylogenies inferred from molecular and phenotypic data sets for hagfish, lampreys, and gnathostomes. Journal of Experimental Zoology Part B: Molecular Biology and Evolution 2009;312:749–61.
[46] Kuraku S. Insights into cyclostome phylogenomics: pre-2R or post-2R. Zoological Science 2008;25:960–8.
[47] Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 2008;452:745–9.
[48] Delsuc F, Brinkmann H, Chourrout D, Philippe H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 2006;439:965–8.
[49] Kim DK, Cho EB, Moon MJ, Park S, Hwang JI, Kah O, et al. Revisiting the evolution of gonadotropin-releasing hormones and their receptors in vertebrates: secrets hidden in genomes. General and Comparative Endocrinology 2011;170:68–78.
[50] Hoffmann FG, Opazo JC, Storz JF. Differential loss and retention of cytoglobin, myoglobin, and globin-E during the radiation of vertebrates. Genome Biology and Evolution 2011;3:588–600.
[51] Braasch I, Volff JN, Schartl M. The endothelin system: evolution of vertebrate-specific ligand–receptor interactions by three rounds of genome duplication. Molecular Biology and Evolution 2009;26:783–99.
[52] Kuraku S, Meyer A. Detection and phylogenetic assessment of conserved synteny derived from whole genome duplications. Methods in Molecular Biology 2012;855:385–95.
[53] Jun J, Mandoiu II, Nelson CE. Identification of mammalian orthologs using local synteny. BMC Genomics 2009;10:630.
[54] Kuraku S, Meyer A. The evolution and maintenance of Hox gene clusters in vertebrates and the teleost-specific genome duplication. International Journal of Developmental Biology 2009;53:765–73.
[55] Fried C, Prohaska SJ, Stadler PF. Independent Hox-cluster duplications in lampreys. Journal of Experimental Zoology Part B: Molecular Development and Evolution 2003;299:18–25.
[56] Stadler PF, Fried C, Prohaska SJ, Bailey WJ, Misof BY, Ruddle FH, et al. Evidence for independent Hox gene duplications in the hagfish lineage: a PCR-based gene inventory of *Eptatretus stoutii*. Molecular Phylogenetics and Evolution 2004;32:686–94.
[57] Tank EM, Dekker RG, Beauchamp K, Wilson KA, Boehmke AE, Langeland JA. Patterns and consequences of vertebrate Emx gene duplications. Evolution and Development 2009;11:343–53.
[58] Collin SP, Davies WL, Hart NS, Hunt DM. The evolution of early vertebrate photoreceptors. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 2009;364:2925–40.
[59] Jimenez-Guri E, Pujades C. An ancient mechanism of hindbrain patterning has been conserved in vertebrate evolution. Evolution and Development 2011;13:38–46.
[60] Inoue JG, Miya M, Tsukamoto K, Nishida M. Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the ancient fish. Molecular Phylogenetics and Evolution 2003;26:110–20.
[61] Kikugawa K, Katoh K, Kuraku S, Sakurai H, Ishida O, Iwabe N, et al. Basal jawed vertebrate phylogeny inferred from multiple nuclear DNA-coded genes. BMC Biology 2004;2:3.
[62] Delarbre C, Gallut C, Barriel V, Janvier P, Gachelin G. Complete mitochondrial DNA of the hagfish, *Eptatretus burgeri*: the comparative analysis of mitochondrial DNA sequences strongly supports the cyclostome monophyly. Molecular Phylogenetics and Evolution 2002;22:184–92.
[63] Kuraku S, Hoshiyama D, Katoh K, Suga H, Miyata T. Monophyly of lampreys and hagfishes supported by nuclear DNA-coded genes. Journal of Molecular Evolution 1999;49:729–35.
[64] Stock DW, Whitt GS. Evidence from 18S ribosomal RNA sequences that lampreys and hagfishes form a natural group. Science 1992;257:787–9.
[65] Storz J, Opazo JC, Hoffmann FG. Gene duplication, genome duplication, and the functioal diversification of vertebrate globins. Molecular Phylogenetics and Evolution, http://dx.doi.org/10.1016/j.ympev.2012.07.013, in press.
[66] Opazo JC, Butts GT, Nery MF, Storz JF, Hoffmann FG. Whole-genome duplication and the functional diversification of teleost fish hemoglobins. Molecular Biology and Evolution 2012.
[67] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 2010;59:307–21.
[68] Mabragana E, Diaz de Astarloa JM, Hanner R, Zhang J, Gonzalez Castro M. DNA barcoding identifies Argentine fishes from marine and brackish waters. PLoS ONE 2011;6:e28655.
[69] April J, Mayden RL, Hanner RH, Bernatchez L. Genetic calibration of species diversity among North America's freshwater fishes. Proceedings of the National Academy of Sciences of the United States of America 2011;108:10602–7.
[70] Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinformatics 2008;9:286–98.