SYMPOSIUM

Palaeophylogenomics of the Vertebrate Ancestor—Impact of Hidden Paralogy on Hagfish and Lamprey Gene Phylogeny

Shigehiro Kuraku¹

Laboratory for Zoology and Evolutionary Biology, Department of Biology, University of Konstanz, Universitätsstrasse 10, 78464 Konstanz, Germany


¹E-mail: shigehiro.kuraku@uni-konstanz.de

Synopsis

In dissecting the transition from invertebrates to vertebrates at the molecular level, whole-genome duplications are recognized as a key event. This gave rise to more copies of genes in jawed vertebrates (gnathostomes), such as the four Hox clusters in the human, compared to the single ancestral cluster in invertebrates. To date, as the most early-branching lineages in vertebrates, cyclostomes (hagfishes and lampreys) have been used for comparative analyses of gene regulations and functions. However, assignment of orthology/paralogy for cyclostomes’ genes is not unambiguously demonstrated. Thus, there is a high degree of incongruence in tree topologies between gene families, although whole genome duplications postulate uniform patterns in gene phylogeny. In this review, we demonstrate how expansion of an ancient genome before the cyclostome–gnathostome split, followed by reciprocal gene loss, can cause this incongruence. This is sometimes referred to as ‘hidden paralogy’.

Introduction

In understanding the evolutionary processes at the advent of vertebrates, extant jawless fishes, namely cyclostomes (hagfishes and lampreys) occupy crucial positions (Kuraku et al. 2009b). Recent applications of advanced molecular approaches have been tackling this exciting frontier in various biological fields (e.g., Shigetani et al. 2002; Pancer et al. 2005; Bridgham et al. 2006; Freitas et al. 2006; McCauley and Bronner-Fraser 2006; Ota et al. 2007). However, genomic expansion that probably occurred at about the time of the split between cyclostome and gnathostome lineages hinders clear-cut identification of genes in cyclostomes and their comparison with counterparts in gnathostomes (reviewed by Kuraku 2008). This crucial event is often called two-round whole-genome duplications (2R-WGDs), and is a major source of multiple subtypes of regulatory genes in model vertebrates, such as mice and chickens (reviewed by Kasahara 2007).

Under the assumption that the expansion of gene families was a genome-wide event, one can expect a uniform scenario of patterns and timings of gene duplications across gene families. Based on observations of a number of gene families, it was suggested that invertebrate deuterostomes (namely, echinoderms, hemichordates, cephalochordates, and urochordates) diverged before the genomic expansion, and that gnathostomes (jawed vertebrates), including cartilaginous fishes, diverged after that (reviewed by Miyata and Suga 2001). This suggestion has been confirmed by genome-wide evidence (Venkatesh et al. 2007; Putnam et al. 2008). However, once we include cyclostome sequences in a molecular phylogenetic tree, an attempt to date the 2R-WGDs relative to the cyclostome–gnathostome split often result in ambiguity: the best tree given by a phylogenetic program is hardly interpretable without assuming additional lineage-specific gains and losses of genes, and in many cases a considerable number of tree topologies are supported at similarly high probabilities (see Kuraku et al. 2009a). Extreme cases are described for Hox gene clusters, for which molecular phylogenetic trees don’t provide sufficient resolution,
and numbers of Hox clusters in cyclostomes still remain unknown (Pendleton et al. 1993; Force et al. 2002; Irvine et al. 2002; Stadler et al. 2004; reviewed in Kuraku and Meyer 2009). The author recently published a study supporting the view that both of the two rounds of WGDs occurred before the cyclostome–gnathostome split (Kuraku et al. 2009a). On one hand, for example, the visual opsin gene family provides unequivocal evidence supporting this timing, but on the other there are many gene families supporting different evolutionary scenarios. Explanations to reconcile such incongruence have not been proposed.

In this review, I focus on the phenomenon ‘hidden paralogy’ (see just below) as a possible source of incongruence between evolutionary scenarios drawn from gene-family trees. This can also account for controversy about the phylogenetic relationships between hagfishes, lampreys, and gnathostomes (see Near 2009).

**Hidden paralogy**

‘Hidden paralogy’ was first introduced as a term in the molecular phylogenetics of bacteria (Daubin et al. 2001; Gribaldo and Philippe 2002; see also Doolittle 1999). Molecular phylogenetic trees sometimes suggest an orthology for a particular pair of genes which will later turn out to be paralogous (or which will never have been revisited in unfortunate cases). This is often caused by secondary losses or delayed identifications of members of gene families.

An example of hidden paralogy is the case of a zebrafish homeobox-containing gene now called *emx3*. This gene was first identified during the time when *Emx1* and *Emx2* genes were successively cloned and revealed to be involved in regionalization of brain (Fig. 1A; Morita et al. 1995). Although the paralogy between this zebrafish gene (former *emx1*) and tetrapod *Emx1* was suggested later (Patarnello et al. 1997), it was not until 2002 that the third subtype (*Emx3*) of the Emx subfamily was identified (Derobert et al. 2002) and that the genuine zebrafish ortholog of *Emx1* was identified (Kawahara and Dawid 2002) (Fig. 1B). In fact, none of eutherian mammals and sauropsids seems to retain the *Emx3* ortholog. Although retention of the *Emx3* orthologs by *Xenopus tropicalis* and opossum *Monodelphis domesticus* was recently evaluated in a phylogenetic context (Viktorin et al. 2009), they still remain non-annotated as protein-coding genes in Ensembl Genome Browser (http://www.ensembl.org/; version 56). Delayed identification of the entire *Emx3* subtype, largely due to the absence of human, mouse, and chicken *Emx3* orthologs, misled the proper orthology/paralogy assignment inside this group of genes. This example does not involve genes of hagfish or lamprey (see Tank et al. 2009 for molecular phylogeny of lamprey *Emx* genes). In the following sections, a few examples involving cyclostome genes.

![Fig. 1](https://academic.oup.com/icb/article-abstract/50/1/124/733872)
are introduced. To achieve sound comparative biological analyses at the molecular level in general, hidden paralogy should be considered as a source of error in construction of gene trees as more and more losses of genes are being reported (e.g., Kuraku et al. 2008; Feiner et al. 2009).

Cdx genes

An example of hidden paralogy involving cyclostomes genes occurred in a recent publication by the present author. The Cdx gene is one of the components of the ParaHox gene cluster (Ferrier and Minguillon 2003; Ferrier et al. 2005), and the entire ParaHox cluster is known to have arisen by WGDs (reviewed by Furlong and Mulley 2008).

As a result of genomic sequencing of the ParaHox cluster in two hagfish species, Eptatretus burgeri and Myxine glutinosa, one Cdx subtype was found in E. burgeri (Furlong et al. 2007). In the molecular phylogenetic tree, placement of the hagfish Cdx gene suggested its orthology to all the three gnathostome subtypes, Cdx1, Cdx2, and Cdx4 (Fig. 2A). However, further survey identified a Cdx homolog in the Japanese lamprey Lethenteron japonicum (LjCdxA) clustering with gnathostome Cdx4 (Fig. 2B) (Kuraku et al. 2009a). Although the statistical support for this clustering is not necessarily high, orthology between this lamprey Cdx homolog and the previously identified one of hagfish was not supported (Kuraku et al. 2009a).

The tree topology shown in Fig. 2B is based solely on the maximum-likelihood (ML) tree, and there were more than 20 tree topologies that were not statistically rejected (Kuraku et al. 2009a). Even though other possibilities cannot be ruled out, in Fig. 2C, one hypothesis for explaining the tree topology in Fig. 2B is proposed. Here, the most early-branching hagfish Cdx in Fig. 2A is regarded as a representative of the fourth Cdx subtype that were lost or unidentified in gnathostomes (Fig. 2C).

Bmp2/4/16 genes

The other example of hidden paralogy also involves complicated phylogeny of regulatory genes, Bmp2 and Bmp4, responsible for many key processes in animal development (reviewed by Mizutani and Bier 2008). Conventionally, two vertebrate subtypes, Bmp2 and Bmp4, were known as orthologs of the fly decapentaplegic (dpp) gene. Recently, as the third ortholog of the fly dpp, a novel group of teleost genes, designated bmp16, was reported with its embryonic expression patterns in the zebrafish (Feiner et al. 2009). It has not been shown clearly that the gene duplications between Bmp2/4/16 were part of the 2R-WGDs, but we here discuss phylogenetic relationships of this group of genes with an assumption that they also convey traces of the 2R-WGDs.

In cyclostomes, three homologs of the sea lamprey Petromyzon marinus, PmBmp2/4-A, PmBmp2/4-B, and PmBmp2/4-C, were reported with their expression analysis (McCauley and Bronner-Fraser 2004). A molecular phylogenetic analysis therein supported the basal branching of the cluster consisting of the three lamprey homologs, suggesting their orthology...
to both of Bmp2 and Bmp4 (Fig. 3A). The following reanalysis, including the newly found teleost fish bmp16 subtype, combined it with the lamprey Bmp2/4-A (Fig. 3B) (Feiner et al. 2009). In addition, lamprey Bmp2/4-B and Bmp2/4-C remained clustered with each other without any gnathostome neighbor. Again, the ML analysis also supported other tree topologies with slightly less likelihood (Feiner et al. 2009). With all statistical fluctuation, an emerging hypothesis based on the tree topology shown in Fig. 3B is illustrated in Fig. 3C. This hypothesis pos-
tulates gene duplications between Bmp2, Bmp4, and Bmp16 before the cyclostome–gnathostome split (probably in the 2R-WGDs) and subsequent losses of the cyclostome counterparts of Bmp2 and Bmp4. The lamprey Bmp2/4-B and Bmp2/4-C are regarded as products of a gene duplication unique to the cy-
clostome lineage, whose gnathostome counterpart was also secondarily lost (Fig. 3C). Further gene sampling in the currently missing key lineage, namely hagfishes, could guide us to a more convinc-
ing reconstruction of the evolutionary history of this gene family.

**Retinoic acid receptors genes**

Members of this subgroup of nuclear receptors function as receptors of retinoic acid (RA) and also as transcription factors regulating spatiotemporal ex-
pression of downstream genes (Mark et al. 2006). One retinoic acid receptors (RAR) homolog was first identified in *P. marinus* (Escriva et al. 2006; Fig. 4A). Later, three subtypes, designated RAR1, RAR2, and RAR3, were isolated from two lamprey species, *L. japonicum* and the southern lamprey *Mordacia mordax*, as well as one hagfish species *E. burgeri* (Kuraku et al. 2009a). In the human genome, three subtypes RARα, RARβ, and RARγ, are located in a large synten block on chromosomes 3, 12, and 17, suggesting that this redundancy was generated by large-scale chromosomal duplications (Kuraku et al. 2009a).

Molecular phylogenetic analysis resulted in the ML tree combining cyclostome RAR1 with gnathostome RARγ as well as cyclostome RAR3 with gnathostome RARα (Fig. 4B). Orthology between cyclostome RAR2 and gnathostome RARβ was not supported by the ML tree (Fig. 4B), but this tree topology could be reconciled most parsimoniously by loss of only one gene (Fig. 4C), or by two independent losses (Fig. 4D). In fact, the tree topology in Fig. 4C which suggests orthology between cyclostome RAR2 and gnathostome RARβ, was supported by the second ML tree (Kuraku et al. 2009a). In this scheme, we do not have to postulate any loss of genes or any unidentified subtype except a highly possible loss of the fourth subtype just after the 2R-WGDs (Fig. 4C). Although there were a large number of tree topologies that were not rejected by statistical assessment, most of those did not support the scenario that places gene duplications after the split between gnathostomes and cyclostomes (Kuraku et al. 2009a).

**Conclusions**

Comparative studies involving molecules should be based on solid assessment of orthology/paralogy. Three examples shown above very clearly illustrate potential problems involved in molecular phylogeny.
after whole genome duplications. To avoid misleading assignment of orthology/paralogy, it is crucial to employ sound molecular phylogenetic approaches after including as many relevant sequences (with large-scale or/and small-scale targeted sequencing) as possible. The illustrated scenarios hypothesized for Cdx and Bmp2/4/16 genes based on ML trees (Figs. 2C and 3C) involve reciprocal gene losses and are not consistent with the maximum-parsimony that is normally respected in evolutionary reconstruction. They are rather based on the recently suggested ‘pan-vertebrate tetraploidization (PV4)’ hypothesis which postulates 2R-WGDs before the cyclostome–gnathostome split (Kuraku et al. 2009a). To explain the pattern in gene families providing very clear phylogenetic signals (e.g., opsin; Kuraku et al. 2009a), the PV4 hypothesis, assuming both of the two rounds of WGDs before the cyclostome–gnathostome split, seems to be the only likely scenario. Gene families that are incongruent with this scheme should be reanalyzed by taking possible hidden paralogy into account. Including the controversy over monophyly of cyclostomes, key questions involving the phylogeny of cyclostome genes should probably be more clearly resolved by this paradigm.

Acknowledgments
I am grateful to Linda Z. Holland and Stacia Sower for their invitation to the symposium organized by the Division of Evolutionary Developmental Biology in the SICB annual meeting and to publication of this article. My gratitude extends to Falk Hildebrand, Axel Meyer, Huan Qiu, Tereza Manousaki, Nathalie Feiner and Adina J. Renz for valuable discussion and three anonymous reviewers for their constructive comments.

Funding
Research grant by German Research Foundation (DFG) (No. KU2669/1-1) and Young Scholar Fund by University of Konstanz.

References

Fig. 4 Alleged hidden paralogy in RAR gene phylogeny. (A) Molecular phylogenetic tree suggested at the time of identification of one sea lamprey homolog PmRAR (Escriva et al. 2006). (B) Molecular phylogenetic tree suggested at the time of recent identification of three RAR subtypes in three cyclostomes species (designated RAR1, RAR2, and RAR3; Kuraku et al. 2009a). (C) A hypothesized view of the molecular phylogeny of vertebrate RAR genes which rather respects parsimony than the tree topology shown in (B). (D) Another hypothesized view of the molecular phylogeny of vertebrate RAR genes which rather respects the tree topology shown in (B) than parsimony. Dotted lines indicate absences of relevant genes (gene losses or incomplete identification). Note that in (A) the lamprey homolog (initially designated RAR but later found orthologous to RAR1) was initially regarded as a pre-duplication prototype orthologous to all the three gnathostome subtypes or an ortholog of either of the three gnathostome subtypes. Arrows indicate gene duplications between gnathostome paralogs. Abbreviations: Gna., gnathostomes; Cyc., cyclostomes; Inv., invertebrates.


