Chapter One

Molecular Evolution in the Lamprey Genomes and Its Relevance to the Timing of Whole Genome Duplications

Tereza Manousaki, Huan Qiu,
Miyuki Noro, Falk Hildebrand,
Axel Meyer and Shigehiro Kuraku

Background

The genomes of two lamprey species have been sequenced, and this has provided the basis for genome-wide comparison of molecular evolution between jawless fishes and the rest of vertebrates. Molecular phylogenetic analyses of jawless fish genes increased our knowledge of the evolutionary time scale of diversification of hagfishes and lampreys, as well as of gene redundancy in their genomes. It was shown that the ancestor of jawed vertebrates experienced two rounds of whole genome duplications (Dehal & Boore, 2005). However, it has been controversial whether this event occurred before or after the ancestors of extant jawless fishes diverged from the lineage which gave rise to jawed vertebrates (e.g., Escriva et al., 2002; reviewed in Kuraku, 2013). Recent molecular phylogenetic studies showed that the whole genome duplications occurred before the radiation of all extant vertebrates including hagfishes and lampreys (Kuraku et al., 2009a), and this scenario has been confirmed by later studies including the genomic analysis of *Petromyzon marinus* (sea lamprey) (Hoffmann et al., 2010; Smith et al., 2013). In this chapter, we analyze peculiar characteristics of the lamprey genomes, focusing mainly on protein-coding regions, to propose potential factors that act as barriers to the understanding of the timing of whole genome duplications.

Jawless fish in molecular phylogenetics

Extant jawless fishes, also called cyclostomes, are comprised of hagfishes and lampreys and diverged from the stem lineage that gave rise to jawed vertebrates (gnathostomes) about 600-500 million years ago (Kuraku et al., 2009b; Figure 1-1). They have been proven to be a monophyletic group, based on molecular phylogenies of both mitochondrial and nuclear genes (reviewed in Kuraku, 2008; Figure 1-1). So far, the so-called 'phylogenomics' approach, involving more than hundreds of genes on nuclear genomes (Kumar et al., 2012), has not been applied to elucidate the relationships between hagfish, lamprey and jawed vertebrates in such a high resolution as demonstrated for other long-standing questions in animal phylogeny (e.g. Misof et al., 2014). This is mainly due to the lack of large-scale sequence information for hagfish (see Delsuc et al., 2006).

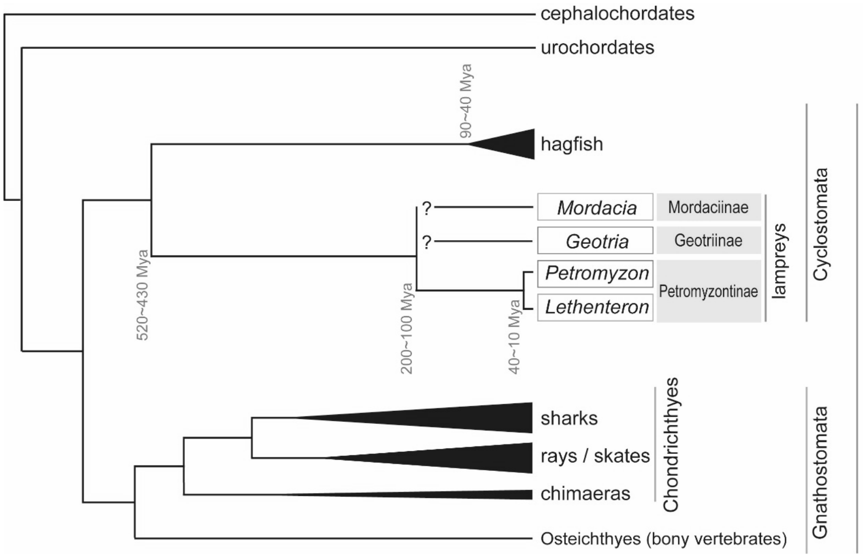

Figure 1-1. Overview of the lamprey phylogeny based on molecular data. Evolutionary time scale within the cyclostome lineage is based on previous literature (Kuraku & Kuratani, 2006; Kuraku et al., 2009b). Branch lengths in the other lineages roughly correspond to evolutionary times inferred from molecular data (Hedges et al., 2006). The phylogenetic relationships among the Mordaciinae, Geotriinae and Petromyzontinae remain to be carefully analyzed with multiple genes (reviewed in Kuraku, 2008).

Lamprey genome sequencing

The first jawless fish species for which whole genome sequencing was started was the sea lamprey *Petromyzon marinus*. Its genome was sequenced with the so-called Sanger method using DNA extracted from the adult liver. An early version of the genome assembly was made public in 2007 at the UCSC Genome Browser (http://genome.ucsc.edu/), which is still available there as version 3 (petMar1; http://hgdownload.soe.ucsc.edu/goldenPath/petMar1/bigZips/). Later, an improved genome assembly, designated version 7 (petMar2), was generated and used as the final product in the genome-wide analysis by the genome consortium (Smith et al., 2013). In the meantime, it was reported that the sea lamprey experiences programmed genomic rearrangement (PGR) in somatic cell lineages (Smith et al., 2009; Smith et al., 2012). It is thus likely that DNA extracted from the source material for the whole genome sequencing was incomplete and heterogeneous, derived from a mixture of somatic cells with differentially rearranged genomes. In 2013, the genome assembly of another northern hemisphere species, the Arctic lamprey *Lethenteron camtschaticum* (formerly known as *L. japonicum*), based on Roche 454 sequencing platform, was also released (Mehta et al., 2013; http://jlampreygenome.imcb.a-star.edu.sg/). This *L. camtschaticum* project employed genomic DNA extracted from the mature testis, which could have contributed to a larger size and higher continuity of the genome assembly because germline cells possess the intact genome (Table 1-1). This project focused on the evolutionary history of *Hox* gene clusters – a long-standing theme regarding cyclostome gene phylogeny (reviewed in Kuraku, 2011; Kuraku and Meyer, 2009). To provide a comparison of assembly statistics between these two genomes, we recomputed basic metrics (see Bradnam et al., 2013) using the latest genome scaffold sequences downloaded from Ensembl (for *P. marinus*) and NCBI (for *L. camtschaticum*) (Table 1-1).

Table 1-1. Assembly statistics of the two lamprey genomes.

Species	**Assembly ID**	**Total # of bases (Gbp)**	**# of scaffolds**	**Scaffold N50 (Kbp)**	**% N**	**Max scaffold length (Kbp)**	**Min scaffold length (Kbp)**
Petromyzon marinus	petMar2	0.886	25,006	79.7	26.8	3,631	0.201
Lethenteron camtschaticum	LetJap1	1.031	86,125	923.6	17.2	11,640	0.867

The statistics in this table are based on computations using all publicly available scaffolds and are partly different from those in the respective publications reporting the genome analyses.

To analyze completeness of protein-coding landscape of the two lamprey genomes, we used a program pipeline CEGMA which reports the number of detected genes among 248 conserved genes (CEG, core eukaryotic genes) (Parra et al., 2009). As a result, the *L. camtschaticum* genome assembly was shown to cover 80% (199/248) of the CEGs, while 69% was detected in the *P. marinus* genome (172/248) (Table 1-2). This suggests that the *L. camtschaticum* genome assembly covers more protein-coding genes. The genome consortia of *P. marinus* and *L. camtschaticum* reported 26,046 and 17,829 protein-coding genes, respectively (Smith et al., 2013; http://jlampreygenome.imcb.a-star.edu.sg/). These two species are thought to have diverged relatively recently, i.e., 40-10 million years ago (Kuraku & Kuratani, 2006) and possess very similar karyotypic features (reviewed in Caputo Barucchi et al., 2013). It is unlikely that the genomic contents and protein-coding landscape largely differs between the two genomes. Thus, the difference in the number of predicted genes is likely caused by the difference in the completeness of genome assemblies or the difference in gene prediction methods.

Table 1-2. Protein-coding landscape in the two lamprey genome assemblies.

Species	# CEGs detected by CEGMA			# of predicted genes
	Complete	Partial	All	
Petromyzon marinus	140	32	172	26.046
Lethenteron camtschaticum	141	58	199	17.829

See Parra et al. (2009) for details of the criteria for categorizing genes detected in genome assembly into ‘complete’ and ‘partial’.

GC-content

Peculiarity of lamprey genes in terms of GC-content was already reported before the whole genome sequence of lampreys became available (Kuraku & Kuratani, 2006; reviewed in Kuraku, 2008). In the comprehensive analysis of the *P. marinus* genome consortium, we performed an intensive investigation of its base composition (Smith et al., 2013). The *P. marinus* genome exhibited relatively high overall GC-content (45.9%), and protein-coding regions, particularly synonymous nucleotide sites, especially had high GC-content (Supplementary Figure 6 in Smith et al., 2013; also see Figure 1-2).

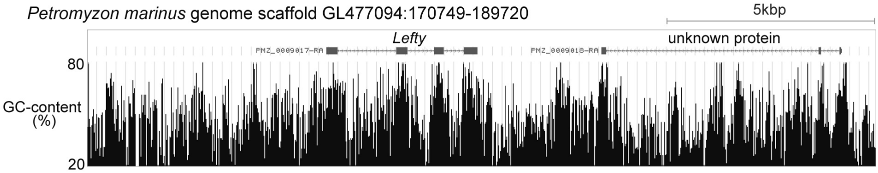

Figure 1-2. Browser view of GC-content in a selected region in the *Petromyzon marinus* genome. The graph of GC-content was obtained as GC-percent track at the UCSC Genome Browser, for the *P. marinus* genomic scaffold GL477094 (base position 170749-189720) containing a homolog of the *Lefty* gene (PMZ_0009017-RA). Note that the exons of this gene tend to have high GC-content (70-80%). The other 'unknown' gene in this view (PMZ_0009018-RA on the right) does not have any obvious homolog in other species and might be a lamprey-specific gene. In such a case, GC-content might serve as an indicator of protein-coding nature of genomic sequences.

Here we have analyzed the *Lethenteron camtschaticum* genome and compared some characteristics about GC-content with other vertebrates including *P. marinus* (Table 1-3). The *L. camtschaticum* genome exhibited markedly higher overall GC-content (48.0%) than the *P. marinus* genome (45.9%) (Figure 1-3, centerfold page i). Similarly, overall GC-content of protein-coding regions showed a comparable difference between the two species (Table 1-3).

The difference of global GC-content in the whole genome sequences of the two lampreys might be caused by either the respective choices of DNA source tissue (liver versus testis, in light of programmed genomic rearrangement), sequencing platform (Sanger versus Roche 454) or assembly methods (Arachne versus Newbler), rather than reflecting the genuine genome compositions. This might also hold for the difference in GC-content of protein-coding regions described above. The lamprey genomes would provide an interesting system to study how epigenetic information is organized in the genome with exceptional GC compartmentalization between coding (GC-rich) and non-coding (GC-poor) regions, as little is known about epigenetic regulation of this group of animals (see Tweedie et al., 1997; Covelo-Soto et al., 2014).

Table 1-3. Global and protein-coding GC-content in the two lamprey genomes.

Species	Genome		Overall GC % of coding regions
	Overall GC %	GC % of 10Kbp non-overlapping windows	
Petromyzon marinus	45.9	45 ± 3	56.3
Lethenteron camtschaticum	48.0	47 ± 4	59.6

Gene model

As the lamprey genomes have peculiar features in their protein-coding sequences (see below), gene prediction based on training with those features is expected to enhance its sensitivity and precision. The *Petromyzon marinus* genome consortium employed the program package MAKER (Cantarel et al., 2008) for genome-wide gene prediction, and it produced a gene typical of vertebrate genomes (Table 1-2) (Smith et al., 2013). In order to predict lamprey genes more precisely, we independently sought to implement lamprey-specific features in gene prediction. First, we built transcriptome assembly using all Sanger sequence reads of *P. marinus* available in NCBI dbEST (as of March 2008). In the assembled transcript contigs, we inferred open reading frames (ORFs) with identical lengths and high sequence conservation (≥70% positive match at the amino acid level, with a methionine corresponding to the putative start codon) in comparison with their jawed vertebrate homologs. Among 828 putative ORFs selected as above, we identified 132 ORF sequences that were contained in the *P. marinus* genome assembly petMar1 (version 3) with presumably full intronic and 2Kbp-long flanking sequences. Using them, we executed the training module of AUGUSTUS version 2.0.3 (Stanke & Waack, 2003) as instructed in its manual. The resulting parameter files for *P. marinus* gene model were passed to the developer of AUGUSTUS and are now available in the default species list (with the species identifier ‘lamprey’) of the installable program package (http://bioinf.uni-greifswald.de/augustus/ binaries/) and web server (http://bioinf.uni-greifswald.de/webaugustus/ prediction/create). This alternative gene prediction platform provides a complementary approach to exploit genomic resources of lampreys, although it remains to be carefully assessed whether the species parameters for *P. marinus* performs well for other lamprey species.

Codon usage bias and amino acid composition

Before whole genome sequences of lampreys became available, we performed analyses on codon usage bias and amino acid composition with 173 protein-coding genes of *Petromyzon marinus* that were available in GenBank (Qiu et al., 2011). In this study based on the relatively small data set, we suggested that lampreys have peculiar patterns of codon usage bias and amino acid composition. More recently, with the whole genome sequences of *P. marinus*, we performed more comprehensive analyses on those characteristics and confirmed that the peculiarity in the sequences of lamprey genes and peptides is genome-wide (Smith et al., 2013; Figure 1-4a and 1-4b, centerfold page ii). In addition, we revealed that GC-content in protein-coding regions is the major factor contributing to the peculiarity of codon usage bias and amino acid composition (Figure 1-4c and 1-4d, centerfold page ii). Our analyses did not support the relevance of codon usage bias to gene expression levels (Supplementary Figure 10 of Smith et al., 2013). It is of particular interest whether this lamprey-specific pattern is shared with other jawless fish genomes.

Homopolymeric amino acid (HPAA) tracts

More recently, we focused on homopolymeric amino acid (HPAA) tracts in peptide sequences (or single amino acid repeats, such as ‘QQQQQQQQ’; see Mularoni et al., 2010) and carried out a cross-species comparison of their frequencies (Noro et al., 2015). Our interest originated from a particular case of lamprey *Emx* genes (reviewed in Kuraku, 2010). Lampreys possess at least two *Emx* genes (*EmxA* and *EmxB*; Tank et al., 2009), and their gene products have a Q-tract and an A-tract at equivalent locations in the sequences of the two *Emx* gene products (Figure 1-5a, centrefold page iii-iv; Noro et al., 2015). A comparison with their hagfish orthologs without conspicuous HPAA tracts indicates that the insertions of the HPAA tracts occurred in the lamprey lineage after the split of the hagfish lineage (Figure 1-5a, centerfold page iii-iv). Our reanalysis confirmed the result by Tank et al. (2009) supporting lamprey lineage-specific *Emx* gene duplication, whereas the support was significantly weakened when the HPAA tracts were deleted from the multiple sequence alignment (Noro et al., 2015).

Inspired by the example of *Emx* genes, we performed a genome-wide survey of HPAA tract insertion. Our survey revealed a significant abundance of HPAA tracts in the overall protein-coding landscape of the sea lamprey genome, compared to that in the human and zebrafish (Noro et al., 2015). It also detected significant enrichment of G-tracts and Q-tracts unique to the sea lamprey (Noro et al., 2015; Figure 1-5b, centerfold page iii-iv). It is unknown what biochemical reasons underlie this species difference in HPAA tract insertion. If the trend of HPAA tract insertion reflects similarly on multiple sequences with similar property, namely paralogs, phylogenetic signals in those sequences might be weakened or erased by the secondary effects. This can result in erroneous alignments and molecular phylogeny inferences.

Perspectives

Several studies have reported gene duplications in the cyclostome lineage (Fried et al., 2003; Stadler et al., 2004; Tank et al., 2009). More recently, a genome-wide analysis suggested that a duplication event at the genome scale introduced lineage-specific duplicates (Mehta et al., 2013). Our analyses have highlighted several unique aspects of molecular evolution that seem to be characteristic of the cyclostome lineage. Above all, the unique sequence property of lamprey protein-coding genes is remarkable. It possibly drove convergent sequence evolution among ancient paralogs towards unexpected similarity, resulting in erroneous proximity between the paralogs in inferred molecular phylogeny. Molecular phylogeny inference based on amino acid sequences, as often practiced, is apparently supposed to circumvent the effect of GC-content and codon usage bias but is still prone to unfavorable effect of amino acid composition. Some of the gene duplications attributed to the cyclostome lineage so far could be explained by this possible artefact, and it can mislead our interpretation of the timing of whole genome duplications. Thus, extra cautions should be exercised in analyzing gene family trees involving lamprey genes (Figure 1-6).

Our knowledge of cyclostome genome compositions is still limited to Northern Hemisphere lampreys. Evolving DNA sequencing technology has enabled economical genome sequencing, and whole genome sequencing of hagfish and Southern Hemisphere lampreys are anticipated. With those

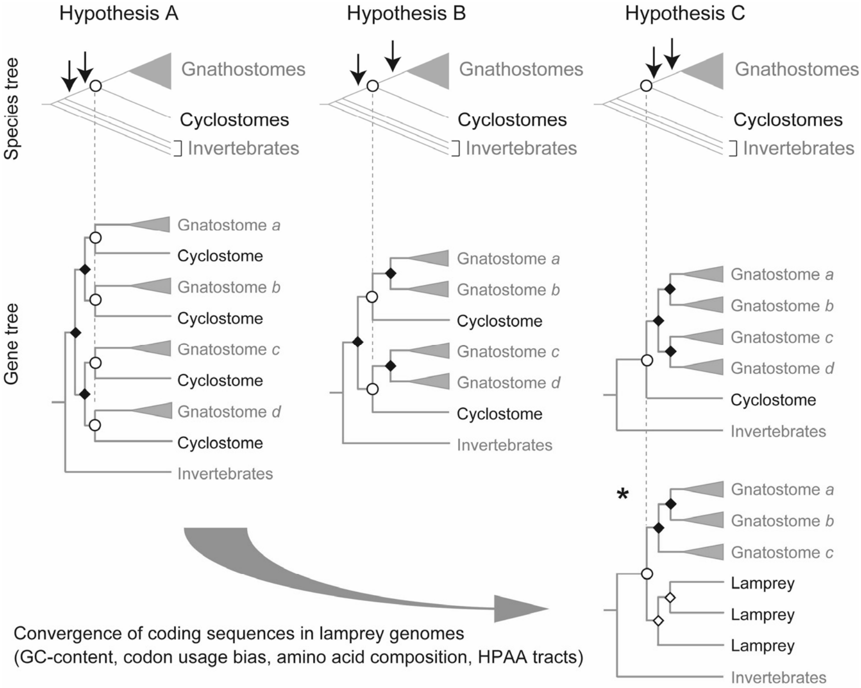

Figure 1-6. Possible causes of misinterpretation on molecular phylogeny involving lamprey genes. Alternative scenarios regarding the timing of two whole genome duplications (Hypothesis A-C) are shown with species trees and hypothetical gene trees. In the species trees, open circles indicate gnathostome-cyclostome split, and black arrows indicate the timing of whole genome duplication. In the hypothetical gene trees, a black diamond indicates gene duplication giving rise to multiple gnathostome paralogs, while a white diamond represents gene duplication giving rise to multiple cyclostome paralogs. Hypothesis A, with both rounds of whole genome duplications before the split between cyclostomes and gnathostomes, has been supported by a series of recent studies (Hoffmann et al., 2010; Kuraku et al., 2009a; Smith et al., 2013). In reality, some gene families exhibit molecular phylogeny depicted in the right bottom corner (*), with multiple cyclostome genes exclusively clustering with each other. This phylogenetic pattern, with gene duplications after the split between cyclostomes and gnathostomes, is incongruent with Hypothesis A and is rather compatible with Hypothesis C. We propose that the incongruence is, at least partly, caused by convergence of lamprey sequences discussed in this chapter.

resources, more comprehensive comparison of genomic features is expected to provide an increased understanding of what in the genome makes the phenotypic differences between jawless fishes and other chordates.

Acknowledgements

The authors acknowledge Maria Anisimova and Christophe Dessimoz for insightful discussion and Yuichiro Hara and Mitsutaka Kadota for critical reading of the manuscript.

References

Bradnam K.R., Fass J.N., Alexandrov A., Baranay P., Bechner M., Birol I., Boisvert S., Chapman J.A., Chapuis G., Chikhi R., Chitsaz H., Chou W.C., Corbeil J., Del Fabbro C., Docking T.R., Durbin R., Earl D., Emrich S., Fedotov P., Fonseca N.A., Ganapathy G., Gibbs R.A., Gnerre S., Godzaridis E., Goldstein S., Haimel M., Hall G., Haussler D., Hiatt J.B., Ho I.Y., Howard J., Hunt M., Jackman S.D., Jaffe D.B., Jarvis E.D., Jiang H., Kazakov S., Kersey P.J., Kitzman J.O., Knight J.R., Koren S., Lam T.W., Lavenier D., Laviolette F., Li Y., Li Z., Liu B., Liu Y., Luo R., Maccallum I., Macmanes M.D., Maillet N., Melnikov S., Naquin D., Ning Z., Otto T.D., Paten B., Paulo O.S., Phillippy A.M., Pina-Martins F., Place M., Przybylski D., Qin X., Qu C., Ribeiro F.J., Richards S., Rokhsar D.S., Ruby J.G., Scalabrin S., Schatz M.C., Schwartz D.C., Sergushichev A., Sharpe T., Shaw T.I., Shendure J., Shi Y., Simpson J.T., Song H., Tsarev F., Vezzi F., Vicedomini R., Vieira B.M., Wang J., Worley K.C., Yin S., Yiu S.M., Yuan J., Zhang G., Zhang H., Zhou S. & Korf I.F. 2013. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience* 2, 10. doi 10.1186/2047-217X-2-10.

Cantarel B.L., Korf I., Robb S.M.C., Parra G., Ross E., Moore B., Holt C., Sánchez Alvarado A. & Yandell M. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* 18, 188-196. doi 10.1101/gr.6743907.

Caputo Barucchi V., Giovannotti M., Nisi Cerioni P. & Splendiani A. 2013. Genome duplication in early vertebrates: insights from agnathan cytogenetics. *Cytogenetetics & Genome Research* 141, 80-89. doi 10.1159/000354098.

Covelo-Soto L., Morán P., Pasantes J.J. & Pérez-García C. 2014. Cytogenetic evidences of genome rearrangement and differential epigenetic chromatin modification in the sea lamprey (*Petromyzon marinus*). *Genetica* 142, 545-554. doi 10.1007/s10709-014-9802-5.

Dehal P. & Boore J.L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology* 3, e314. doi 10.1371/journal.pbio.0030314.

Delsuc F., Brinkmann H., Chourrout D. & Philippe H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* 439, 965-968. doi 10.1038/nature04336.

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. *Molecular Biology and Evolution* 19, 1440-1450.

Fried C., Prohaska S.J. & Stadler P.F. 2003. Independent Hox-cluster duplications in lampreys. *Journal of Experimental Zoology Part B: Molecular Development and Evolution* 299, 18-25. doi 10.1002/jez.b.37.

Hedges S.B., Dudley J. & Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22, 2971-2972. doi 10.1093/bioinformatics/btl505.

Hoffmann F.G., Opazo J.C. & Storz J.F. 2010. 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* 107, 14274-14279. doi 10.1073/pnas.1006756107.

Kumar S., Filipski A.J., Battistuzzi F.U., Kosakovsky Pond S.L. & Tamura K. 2012. Statistics and truth in phylogenomics. *Molecular Biology and Evolution* 29, 457-472. doi 10.1093/molbev/msr202.

Kuraku S. 2008. Insights into cyclostome phylogenomics: pre-2R or post-2R. *Zoological Science* 25, 960-968. doi: 10.2108/zsj.25.960.

Kuraku S. 2010. Palaeophylogenomics of the vertebrate ancestor—impact of hidden paralogy on hagfish and lamprey gene phylogeny. *Integrative and Comparative Biology* 50, 124-129. doi 10.1093/icb/icq044.

Kuraku S. 2011. *Hox* gene clusters of early vertebrates: do they serve as reliable markers for genome evolution? *Genomics, Proteomics & Bioinformatics* 9, 97-103. doi 10.1016/S1672-0229(11)60012-0.

Kuraku S. 2013. Impact of asymmetric gene repertoire between cyclostomes and gnathostomes. 2013. *Seminars in Cell and Developmental Biology* 24, 119-127. doi

10.1016/j.semcdb.2012.12.009.

Kuraku S. & Kuratani S. 2006. Time scale for cyclostome evolution inferred with a phylogenetic diagnosis of hagfish and lamprey cDNA sequences. *Zoological Science* 23, 1053-1064. doi

http://dx.doi.org/10.2108/zsj.23.1053.

Kuraku S. & Meyer A. 2009a. The evolution and maintenance of *Hox* gene clusters in vertebrates and the teleost-specific genome duplication. *The International Journal of Developmental Biology* 53, 765-773. doi 10.1387/ijdb.072533km.

Kuraku S., Meyer A. & Kuratani S. 2009b. Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? *Molecular Biology and Evolution* 26, 47-59. doi 10.1093/molbev/msn222.

Kuraku S., Ota K.G. & Kuratani S. 2009c. Jawless fishes (Cyclostomata). In S.B. Hedges & S. Kumar (eds.): *The Timetree of Life*. Pp. 317-319. New York: Oxford University Press.

Mehta T.K., Ravi V., Yamasaki S., Lee A.P., Lian M.M., Tay B.H., Tohari S., Yanai S., Tay A., Brenner S. & Venkatesh B. 2013. Evidence for at least six Hox clusters in the Japanese lamprey (*Lethenteron japonicum*). *Proceedings of the National Academy of Sciences of the United States of America* 110, 16044-16049. doi 10.1073/pnas.1315760110.

Misof B., Liu S., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B., Ware J., Flouri T., Beutel R.G., Niehuis O., Petersen M., Izquierdo-Carrasco F., Wappler T., Rust J., Aberer A.J., Aspock U., Aspock H., Bartel D., Blanke A., Berger S., Bohm A., Buckley T.R., Calcott B., Chen J., Friedrich F., Fukui M., Fujita M., Greve C., Grobe P., Gu S., Huang Y., Jermiin L.S., Kawahara A.Y., Krogmann L., Kubiak M., Lanfear R., Letsch H., Li Y., Li Z., Li J., Lu H., Machida R., Mashimo Y., Kapli P., McKenna D.D., Meng G., Nakagaki Y., Navarrete-Heredia J.L., Ott M., Ou Y., Pass G., Podsiadlowski L., Pohl H., von Reumont B.M., Schutte K., Sekiya K., Shimizu S., Slipinski A., Stamatakis A., Song W., Su X., Szucsich N.U., Tan M., Tan X., Tang M., Tang J., Timelthaler G., Tomizuka S., Trautwein M., Tong X., Uchifune T., Walzl M.G., Wiegmann B.M., Wilbrandt J., Wipfler B., Wong T.K., Wu Q., Wu G., Xie Y., Yang S., Yang Q., Yeates D.K., Yoshizawa K., Zhang Q., Zhang R., Zhang W.,