GBE

Genome Biology and Evolution

SMBE

Early Evolution of Vertebrate Mybs: An Integrative Perspective Combining Synteny, Phylogenetic, and Gene Expression Analyses

Emeline B. Campanini[1,†], Michael W. Vandewege[2,†], Nisha E. Pillai[3], Boon-Hui Tay[3], Justin L. Jones[4], Byrappa Venkatesh[3,5,*], Federico G. Hoffmann[2,6,*]

[1]Departament of Genetics and Evolution, Federal University of São Carlos, Brazil

[2]Department of Biochemistry, Molecular Biology, Entomology, and Plant Pathology, Mississippi State University

[3]Institute of Molecular and Cell Biology, Comparative and Medical Genomics Laboratory, Agency for Science, Technology and Research, Biopolis, Singapore

[4]Department of Biological & Physical Sciences, Saint Augustine's University

[5]Departments of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[6]Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University

*Corresponding author: E-mail: mcbbv@imcb.a-star.edu.sg; federico.g.hoffmann@gmail.com.

[†]These authors contributed equally to this work.

Accepted: October 8, 2015

Abstract

The genes in the Myb superfamily encode for three related transcription factors in most vertebrates, *A-*, *B-*, and *c-Myb*, with functionally distinct roles, whereas most invertebrates have a single *Myb*. *B-Myb* plays an essential role in cell division and cell cycle progression, *c-Myb* is involved in hematopoiesis, and *A-Myb* is involved in spermatogenesis and regulating expression of pachytene PIWI interacting RNAs, a class of small RNAs involved in posttranscriptional gene regulation and the maintenance of reproductive tissues. Comparisons between teleost fish and tetrapods suggest that the emergence and functional divergence of the *Myb* genes were linked to the two rounds of whole-genome duplication early in vertebrate evolution. We combined phylogenetic, synteny, structural, and gene expression analyses of the *Myb* paralogs from elephant shark and lampreys with data from 12 bony vertebrates to reconstruct the early evolution of vertebrate *Mybs*. Phylogenetic and synteny analyses suggest that the elephant shark and Japanese lamprey have copies of the *A-*, *B-*, and *c-Myb* genes, implying their origin could be traced back to the common ancestor of lampreys and gnathostomes. However, structural and gene expression analyses suggest that their functional roles diverged between gnathostomes and cyclostomes. In particular, we did not detect *A-Myb* expression in testis suggesting that the involvement of A-Myb in the pachytene PIWI interacting RNA pathway is probably a gnathostome-specific innovation. We speculate that the secondary loss of a central domain in lamprey A-Myb underlies the functional differences between the cyclostome and gnathostome A-Myb proteins.

Key words: gene family, gnathostome, cyclostome, gene duplication, whole genome duplication.

Introduction

Gene duplications and whole-genome duplications (WGDs) are major generators of raw material for biological innovations (Ohno 1970). Following gene duplication, the functional and regulatory divergence among the resulting paralogs is considered an important source of evolutionary novelties (Ohno 1970; Zhang 2003; Nei and Rooney 2005; Lynch 2007). In the case of vertebrates, multiple analyses suggest that the emergence of several key vertebrate innovations is linked to the two rounds of WGD early in vertebrate evolution (see Van de Peer et al. 2009 and references cited therein). Some of these vertebrate-specific innovations include the endoskeleton, the neural crest and derivative cell types, neurogenic placodes, many signaling transduction pathways, a specialized system for oxygen storage and delivery, and a complex, segmented brain (Zhang and Cohn 2008; Braasch et al. 2009;

Larhammar et al. 2009; Hoffmann et al. 2012; Schwarze et al. 2014). The vertebrate members of the Myb family are thought to be one such example (Davidson et al. 2005, 2013).

The genes in the Myb superfamily encode for DNA-binding proteins that function as transcription regulators that play key roles in development. Animal Myb proteins are characterized by a highly conserved N-terminal DNA-binding domain, which consists of three tandem repeats, labeled as Myb repeats (Sakura et al. 1989), and a conserved regulatory C-terminal domain that appears to be animal specific (Oh and Reddy 1998). There are three related *Myb* genes in most vertebrates, namely *A-*, *B-*, and *c-Myb*, also known as *MybL1*, *MybL2,* and *Myb* that play critical roles in cell differentiation (Oh and Reddy 1999; Lipsick et al. 2001). In contrast, most invertebrate genomes encode a single *Myb* transcription factor gene (Davidson et al. 2005).

Experimental evidence indicates that the three vertebrate Mybs have functionally distinct roles. *B-Myb* is ubiquitously expressed and plays an essential role in cell division and cell cycle progression (Sitzmann et al. 1996; Sala 2005; Tarasov et al. 2008). c-Myb is part of a complex genetic network whose function is to specify and maintain hematopoietic progenitors and to regulate their differentiation (Mucenski et al. 1991; Soza-Ried et al. 2010). In turn, A-Myb acts in the proliferation and/or differentiation of neurogenic, spermatogenic, and B-lymphoid cells (Trauth et al. 1994) and is highly expressed in the male germ cells and breast epithelial cells of pregnant mice (Toscani et al. 1997). This protein is also the male-specific master regulator of meiosis (Bolcun-Filas et al. 2011), and in amniotes, regulates the expression of PIWI interacting RNA (piRNA) precursors at the pachytene stage of prophase 1 during spermatogenesis (Li et al. 2013). piRNAs are a class of small RNAs involved in protecting the genome against transposable elements, gene regulation, and sperm maturation (Aravin et al. 2007; Gou et al. 2014). Mice knockouts illustrate the importance of the roles played by these genes: *B-Myb* knockouts die as early embryos (Tanaka et al. 1999), *c-Myb* knockouts die as late embryos due to failures in hematopoiesis (Mucenski et al. 1991), and *A-Myb* knockouts are viable but cannot complete spermatogenesis or mammary gland development (Toscani et al. 1997).

Based on phylogenetic, structural, and synteny analyses, Davidson et al. (2005, 2013) linked the emergence of the *Myb* genes of bony vertebrates, or Eutelostomes, to segmental duplications in their common ancestor that probably correspond to the two rounds of WGD early in vertebrate evolution. In this scenario, the first duplication gave rise to the *B-Myb* gene and the *A/c-Myb* progenitor, and the second duplication gave rise to the *A-* and *c-Myb* paralogs (fig. 1*A*). Structural features support this phylogenetic arrangement, as the peptides encoded by vertebrate *A-Myb* and *c-Myb* include a central transcriptional activation domain (CTAD), which is absent in vertebrate *B-Mybs* and invertebrates Mybs. In agreement with structural similarities, functional comparisons suggest that vertebrate *B-Myb* is functionally equivalent to the single copy *Myb* of fruit flies and probably represents the ancestral functional role (Davidson et al. 2005, 2013). In the proposed model, the first of these duplications was followed by a subfunctionalization event, where the *A/c-Myb* progenitor developed a restricted pattern of expression, and a neofunctionalization event associated to the acquisition of the CTAD and novel functional roles, and the second duplication was followed by a subfunctionalization event that lead to the current roles of the *A-* and *c-Myb* paralogs (Davidson et al. 2005, 2013).

The above model was based on characterization of *Myb* genes from tetrapods and teleost. In particular, no *Myb* gene has been characterized in cartilaginous fish or jawless vertebrates, which represent the two deepest lineages of extant vertebrates. Cartilaginous fish and bony vertebrates are the two major groups of jawed vertebrates (Gnathostomata). Gnathostomes are sister to cyclostomes, the vertebrate group that includes jawless lampreys and hagfish and represent the most ancient lineage of extant vertebrates. The crucial phylogenetic position of cartilaginous fishes and cyclostomes makes them valuable in understanding the origin and early evolution of vertebrate gene families. In this study, we took advantage of the recently published genomes of the elephant shark (*Callorhinchus milii,* Venkatesh et al. 2014), a member of cartilaginous fishes, and two species of lampreys, the sea lamprey (*Petromyzon marinus*, Smith et al. 2013) and the Japanese lamprey (*Lethenteron japonicum*, Mehta et al. 2013) to gain insights into the origin and evolution of the *Myb* genes in vertebrates. We characterized the *Myb* repertoire from genomes representing all major lineages of vertebrates and also determined the expression profile of the different *Myb* paralogs in elephant shark, spotted gar, tilapia, coelacanth, frog, chicken, opossum, and human using RNA-seq–data and in the Japanese lamprey by quantitative polymerase chain reaction (qPCR). We then integrated phylogenetic reconstructions with synteny comparisons and analyses of protein structure and gene expression in a comparative framework that explicitly considers the organismal tree. Our results suggest that the *Myb* repertoires of lampreys and gnathostomes have a common evolutionary origin but differ in their functions. Finally, by tracking the evolution of expression of the *A-Myb* paralog, we provide testable hypotheses regarding the emergence of pachytene piRNAs, a class of piRNAs only found in vertebrates (Lau et al. 2006; Aravin et al. 2007; Houwing et al. 2007; Li et al. 2013).

Materials and Methods

Sequence Data

We used bioinformatic techniques to collect the full set of Myb-like genes in representatives of the two deepest divergences in the tree of extant vertebrates: cyclostomes,

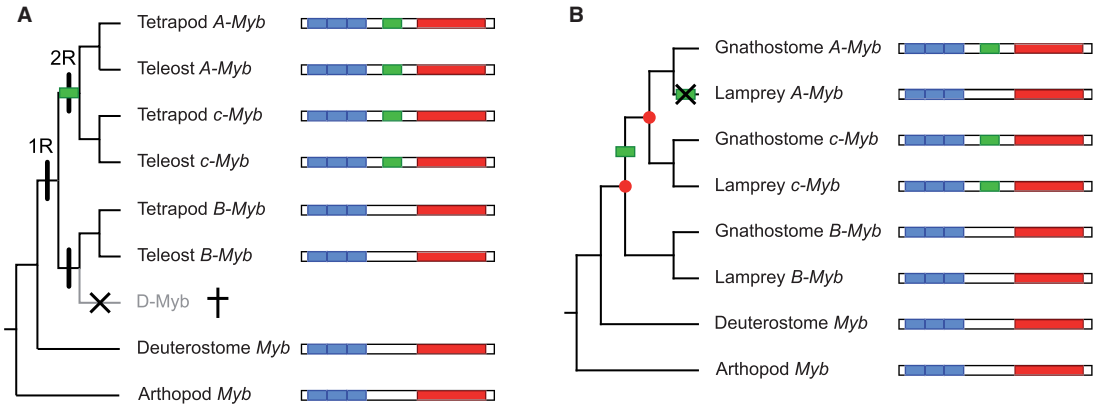

Fig. 1.—(*A*) Depiction of the evolutionary model proposed by Davidson et al. (2005, 2013), with the two rounds of WGD early in vertebrate evolution, 1R and 2R, giving rise to the vertebrate *Mybs*. WGDs indicated by black lines, the acquisition of a CTAD indicated by a green rectangle, the three Myb repeats of the DNA-binding domain in the N-terminal portion as blue rectangles, and the C-terminal regulatory domain as red rectangles. Adapted from Davidson et al. (fig. 1, 2013). (*B*) Depiction of the evolutionary model of vertebrate Mybs derived from our study. Duplications are indicated by red circles. Because of uncertainties regarding the placement of the 1R and 2R WGD on the vertebrate tree, we have remained agnostic on the matter.

represented by Japanese lamprey and sea lamprey; and cartilaginous fish, class Chondrichthyes, represented by elephant shark (*C. milii*). For comparative purposes, our analyses included 12 additional bony vertebrates, four nonvertebrate deuterostomes and three protostomes (see supplementary table S1, Supplementary Material online, for the complete list of sequences included). Our sample of bony vertebrates included four mammals (human, *Homo sapiens*; mouse, *Mus musculus*; gray short-tailed opossum, *Monodelphis domestica*; and platypus, *Ornithorhynchus anatinus*), four sauropsids (chicken, *Gallus gallus*; Chinese softshell turtle, *Pelodiscus sinensis*; American alligator, *Alligator mississippiensis*; and green anole lizard, *Anolis carolinensis*), one amphibian (western clawed frog, *Xenopus tropicalis*), one lobe-finned fish (West Indian Ocean coelacanth, *Latimeria chalumnae*), one holostean fish (spotted gar, *Lepisosteus oculatus*), and one teleost fish (zebrafish, *Danio rerio*). Our sample of outgroup sequences included both deuterostomes, represented by three nonvertebrate chordates, one urochordate (sea squirt, *Ciona intestinalis*), one cephalochordate (amphioxus, *Branchiostoma floridae*), plus a hemichordate (acorn worm, *Saccoglossus kowalevskii*), an echinoderm (purple sea urchin, *Strongylocentrotus purpuratus*), and protostomes, represented by three arthropods (fruit fly, *Drosophila melanogaster*; honey bee, *Apis mellifera*; and silkmoth, *Bombyx mori*).

In most cases, BLASTX (Altschul et al. 1990) searches were sufficient to identify the putative *Myb*-like sequences, with the exception of Japanese lamprey. In the latter case, we compared gnathostome- and sea lamprey-predicted *Mybs* to transcripts from multiple tissues and genomic sequence using a combination of tools to identify putative *Mybs*. After annotation, we confirmed the identity of the *Myb*-like gene models by comparing their predicted amino acid sequence to the NCBI database using BLASTX. In the case of elephant shark and Japanese lamprey, we searched genome assemblies of the elephant shark (http://esharkgenome.imcb.a-star.edu.sg/, last accessed October 27, 2015) and the Japanese lamprey (http://jlampreygenome.imcb.a-star.edu.sg/, last accessed October 27, 2015) by TBLASTN using human and zebrafish Myb protein sequences as queries. The regions that showed similarity to Myb proteins were extracted and searched against the NR protein database at NCBI using BLASTX. If this search confirmed the presence of a *Myb* gene in this region, the exon–intron boundaries of the gene were predicted based on BLASTX alignments. The predictions were further refined by manual inspection.

Phylogenetic Analyses

We inferred phylogenetic relationships among *Myb* genes using the full set of genes from the 22 species listed above. We aligned amino acid sequences using Kalign (Lassmann and Sonnhammer 2006), the E-INS-i, L-INS-i and G-INS-i strategies from MAFFT (Katoh et al. 2009; Katoh and Standley 2013), MUSCLE (Edgar 2004), and T-coffee (Notredame et al. 2000). We then compared the resulting alignments using MUMSA (Lassmann and Sonnhammer 2005, 2006), which compares alignment blocks from different alignment strategies to assess the difficulty of an alignment case and ranks each alignment based on a Multiple Overlap Score. Subsequently, we used the best-scoring alignment for all downstream analyses. Phylogenetic relationships were estimated using Bayesian (BA) and maximum likelihood (ML) approaches. Bayesian analyses were carried out in MrBayes v. 3.1.2 (Ronquist and Huelsenbeck 2003) under a mixed model of amino acid substitution. We set two independent runs of six simultaneous chains for 20,000,000 generations, sampling every 2,500 generations, and using default priors. Once convergence was verified, node support was derived from a majority rule

consensus of the last 8,000 trees. In the case of ML, analyses were done in Treefinder v. March 2011 (Jobb et al. 2004), using the "propose model" subroutine to estimate the best-fitting model of amino acid substitution. Tree searches were conducted under the selected model, and node support was evaluated from 1,000 bootstrap pseudoreplicates. All trees were rooted with arthropod Mybs. In addition, we used constrained searches to compare tree topologies corresponding to alternative evolutionary scenarios and compared the resulting trees using the approximately unbiased (AU, Shimodaira 2002) topology tests.

Synteny Comparisons

Patterns of conserved synteny were inferred using the data available on the Ensembl database (Cunningham et al. 2015) and using Genomicus (database, v. 78.01; Louis et al. 2015), except for the Japanese lamprey and elephant shark, where we inspected gene models visually and resolved orthology with reciprocal searches using BLAST (Altschul et al. 1990). We constructed synteny maps for the genomic neighborhoods surrounding the *A-*, *B-,* and *c-Myb* genes in a mammal (human), a bird (chicken), a squamate (green anole lizard), two ray-finned fish (spotted gar and medaka), a cartilaginous fish (elephant shark), and two cyclostomes (Japanese lamprey and sea lamprey). When BLAST comparisons where ambiguous, orthology and paralogy were resolved by additional ML phylogenetic analyses done in MEGA v. 6.06 (Tamura et al. 2013), under the best-fitting model of amino acid substitution.

Structural Analyses

We used the SMART tool (Letunic et al. 2015; available in: http://smart.embl-heidelberg.de, last accessed October 27, 2015) for structural analyses. We first verified the presence of the three DNA-binding Myb domains in all species analyzed and predicted the corresponding positions. To check for the presence of a CTAD, we aligned the amino acid sequence of the A-, B-, and c-Mybs of human, elephant shark, spotted gar, sea lamprey, and Japanese lamprey and visually inspected the homologous region in all sequences and verified our observation using SMART. Finally, we verified the presence of the regulatory C-terminal domains in all sequences as well.

Gene Expression Analyses

RNA-seq data from a diverse set of tissues (brain, heart, kidney, liver, muscle, ovary, spleen, and testis) from the elephant shark, spotted gar, coelacanth, western clawed frog, chicken, gray short-tailed opossum, and human, using tilapia (*Oreochromis niloticus*) as a teleost representative, were collected from NCBI SRA (see supplementary table S2, Supplementary Material online). For each species except for elephant shark, the entire set cDNA sequences were collected from Ensembl. Elephant shark cDNA sequences were collected from http://esharkgenome.imcb.a-star.edu.sg/. Gene expression levels were estimated by RSEM v. 1.2.3 (Li and Dewey 2011), which uses Bowtie v. 0.12.9 (Langmead et al. 2009) to map reads to the proper set of coding sequences. Transcript abundances were measured in transcripts per million.

Quantitative PCR for Lamprey Myb Genes

Total RNA was extracted from brain, gills, heart, intestine, kidney, liver, muscle, notochord, ovary, skin, and testis of adult Japanese lamprey using Trizol reagent (Life Technologies, Carlsbad, CA) according to manufacturer's protocol. One milligram of total RNA was reverse transcribed into 5′RACE-ready single strand cDNA by using the SMART RACE cDNA Amplification kit (Clontech, Palo Alto, CA). The RACE-ready single-strand cDNA was used as a template in qPCR reactions with SYBR Select Master Mix (Life Technologies). Sequences of primers used in qPCR are given in supplementary table S3, Supplementary Material online. All primer pairs were designed to span at least one intron to distinguish cDNA from genomic DNA products. The qPCR conditions comprised 50 °C for 2 min (for uracil DNA glycosylase activation), 95 °C for 3 min, followed by 40 cycles of 95 °C for 3 s, and 65 °C for 30 s. Each analysis was performed in triplicate. Quantification of gene expression levels was performed using the comparative CT method. Expression levels of the Japanese lamprey *Myb* genes were normalized using the β-actin gene as internal reference. The relative expression levels of each *Myb* gene in different tissues were estimated in relation to a reference tissue that showed the lowest level of expression among the tissues analyzed.

Analyses of 5′ Cis Regulatory Regions

For the human *A-*, *B-,* and *c-Myb* genes, we extracted the 5′-untranslated region (UTR) plus 2,000 bases upstream of the UTR and conducted a pairwise BLASTN among sequences.

Analyses of Coduplicated Genes

Davidson et al. (2013) described five gene families that had paralogs within regional proximity to *Myb* paralogs. These families were the EYA, SGK, PLAG, PDE7, and SRC gene families. To study the duplication history of these gene families with regard to *Myb* duplication, we constructed synteny maps and phylogenetic trees of the three families that were generally closest to *Mybs*: EYA, SGK, and PLAG. Given the large genetic distances among genes, we restrained synteny maps to species with the best constructed genome drafts, human, mouse, chicken, anole, and clawed frog.

For each gene family, we extracted all Ensembl release 78 annotated paralogs from human, mouse, opossum, platypus, chicken, softshell turtle, anole, frog, coelacanth, and zebrafish. To find potential paralogs in nonvertebrate species, we conducted a BLASTP search of GenBank using the human proteins as a query and restricted results to acorn worm, sea

squirt, sea urchin, fruit fly, honey bee, and silkmoth. Protein sequences that were likely to have incorrectly annotated exons or exon sequences that were unique to one species and could potentially skew phylogenetic inference were discarded. In each case, we aligned amino acid sequences using the L-INS-i strategy from MAFFT and estimated phylogenetic trees from the resulting alignment following the same methods described for the Myb family.

Results and Discussion

We combined phylogenetic, synteny, structural, and gene expression analyses of the *Myb* paralogs from elephant shark and lampreys with data from other representative vertebrates and outgroups to reconstruct the early stages of evolution of the vertebrate Myb gene family. By incorporating data from elephant shark and two lampreys, we were able to trace changes in this gene family to the last common ancestor of all extant vertebrates and gain insights into the early stages of evolution of this gene family in vertebrates. Because of the GC-bias, codon-bias, and unusual amino-acid composition observed in cyclostome genomes (Qiu et al. 2011; Smith et al. 2013), resolving orthology between gnathostome and cyclostome genes using phylogenies has been challenging (Qiu et al. 2011; Schwarze et al. 2014; Opazo et al. 2015). To alleviate these problems, we first focused on comparing the *Myb* repertoire of the elephant shark, a cartilaginous fish, to the rest of the gnathostomes in our study, all of which are bony vertebrates, and incorporated lamprey genes in a second stage.

Elephant Shark Has Paralogs of A-, B-, and c-Myb

Our bioinformatic searches revealed the presence of three *Myb* paralogs in the elephant shark genome, located on three separate scaffolds. Phylogenetic, structural, and synteny comparisons described below indicate that they correspond to orthologs of the *A-*, *B-*, and *c-Myb* genes from bony vertebrates (figs. 2*A* and 3). Our phylogenies place gnathostome *Mybs* in a monophyletic group and confidently resolve orthology for all elephant shark *Mybs* (fig. 2*A*, supplementary fig. S1*A*, Supplementary Material online). The phylogenetic arrangement within the three gnathostome *Myb* clades matched the expected position given organismal relationships: Elephant shark paralogs were sister to all other gnathostome paralogs, and ray-finned fish, sauropsid, and mammal paralogs fell in monophyletic clades.

Synteny and structural comparisons support the results of phylogenetic analyses, with stronger synteny conservation for the gnathostome *A-* and *c-Mybs*. In the case of *A-Myb*, the elephant shark ortholog is flanked by *SGK3* and *VCPIP1* on one side and *C8Orf46* and *ADHFE1* on the other hand, as in most gnathostomes (fig. 3). *c-Myb* is flanked by *AHI1* and *PDE7B* on one side and by *HBS1L* and *ALDH8A1* in most gnathostomes, including elephant shark, and the elephant shark and bony vertebrate *B-Mybs* are flanked by *ITF52* (fig. 3). From a structural standpoint, the three elephant shark *Myb* paralogs were similar to the corresponding genes in bony vertebrates. All *Mybs* encode proteins that have an N-terminal DNA-binding domain with three repeats, and a C-terminal regulatory domain. In addition, the A- and c-Mybs of elephant shark also possess a CTAD (fig. 4, supplementary table S4, Supplementary Material online). Interestingly, the *Myb* paralogs of most gnathostomes are linked to paralogs of the SGK gene family, *SGK1*, *SGK2*, and *SGK3*, with *c-*, *B-*, and *A-Myb,* respectively (fig. 3, supplementary fig. S2*A*, Supplementary Material online).

From a numerical standpoint, all gnathostomes examined have three *Myb* genes, with the exception of platypus, which apparently possesses two *B-Myb* paralogs (fig. 2*A*). Structural analyses indicate both of them correspond to portions of *B-Myb*: the ENSOANG00000011081 gene model contains the N-terminal portion of the protein and is flanked by *SGK2*, like the *B-Myb* of most gnathostomes, whereas the ENSOANG00000029236 gene model contains the C-terminal portion and is found in a short contig that includes no additional genes. Thus, it would seem these two platypus gene models actually correspond to the single-copy *B-Myb* paralog, which is split among different contigs in the current assembly of this genome.

In terms of relationships among the paralogs, our phylogenies place gnathostome *A-* and *B-Mybs* as sister groups and *c-Myb* as the most divergent, an arrangement that does not match predictions from Davidson's model (supplementary fig. S1*A*, Supplementary Material online). However, support for the node joining *A-* and *B-Myb* was low, and a tree constraining gnathostome *A-Myb* to be sister to *c-Myb* was not significantly different from the best tree ($P \approx 0.29$ in AU topology test). The constrained tree will be preferred because gnathostome *A-* and *c-Myb* are the most derived from a functional standpoint and implies a single origin for the CTAD. Thus, our data indicate that the common ancestor of gnathostomes possessed at least three different *Myb* paralogs found in extant gnathostomes, an inference that is independent of the tree selected.

Lampreys Have Three Myb Paralogs

We then moved to include lamprey *Mybs* in our analyses. Our bioinformatic searches revealed the presence of three separate *Myb* paralogs in the genomes of the two lampreys, similar to gnathostome genomes. These lamprey paralogs were also arranged in three groups which were tentatively labeled as lamprey *A-*, *B-,* and *c-Mybs* based on traces of shared synteny between cyclostome and gnathostome *Mybs* discussed below (fig. 3). The Japanese lamprey genome includes full length copies of *A-*, *B-*, and *c-Myb*, whereas the sea lamprey genome includes full length copies of the *A-* and *c-Myb* paralogs, plus fragments similar to the Japanese lamprey *B-Myb* gene exons in scaffold GL480571, which does not include any

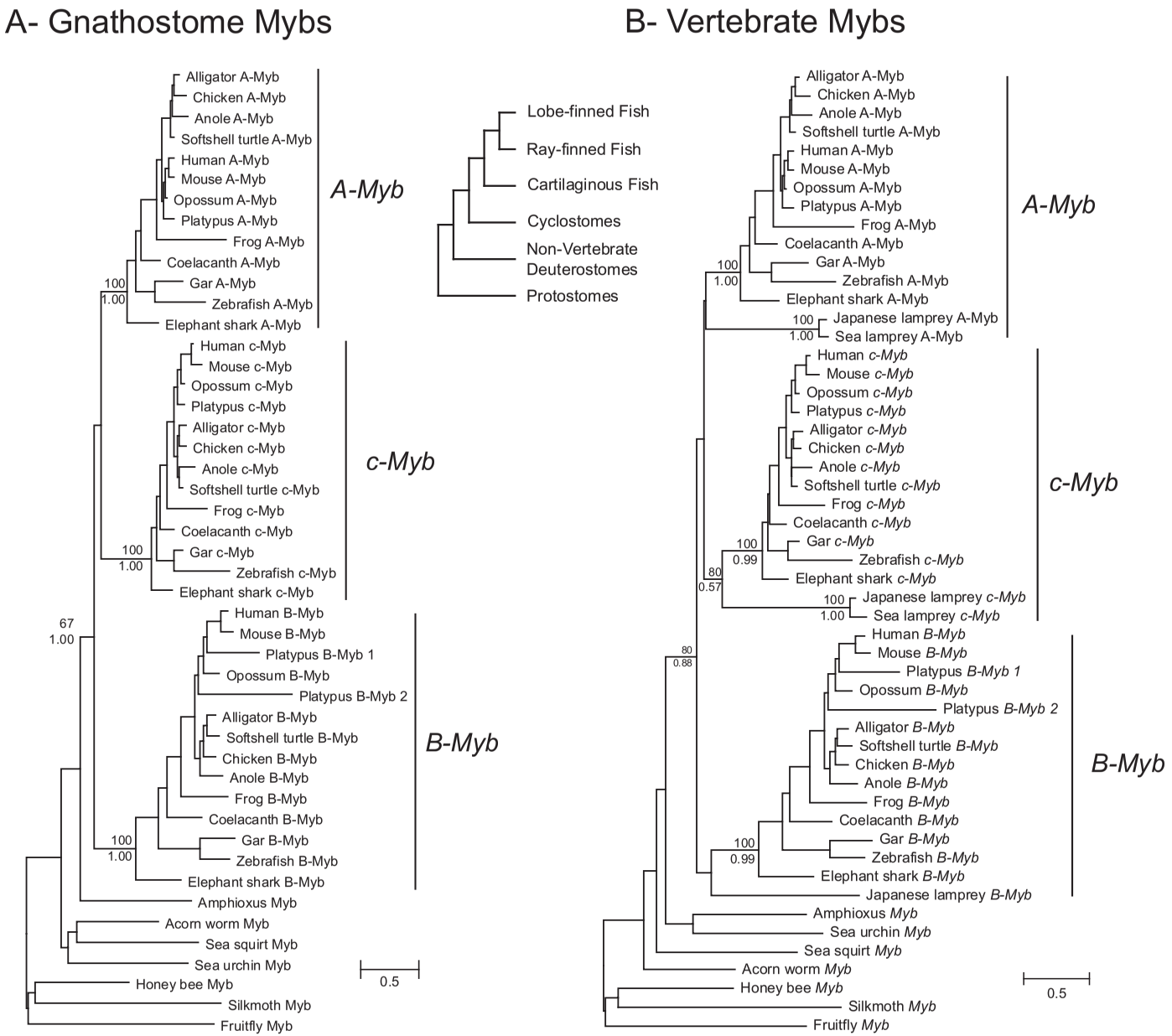

FIG. 2.—Phylogenetic trees for gnathostome (*A*) and vertebrate (*B*) *Mybs*, where syntenic *Mybs* are constrained to be monophyletic, and *A*- and *c-Myb* are constrained to be sister clades. Numbers next to the nodes correspond to ML bootstrap support (above) and Bayesian posterior probabilities (below).

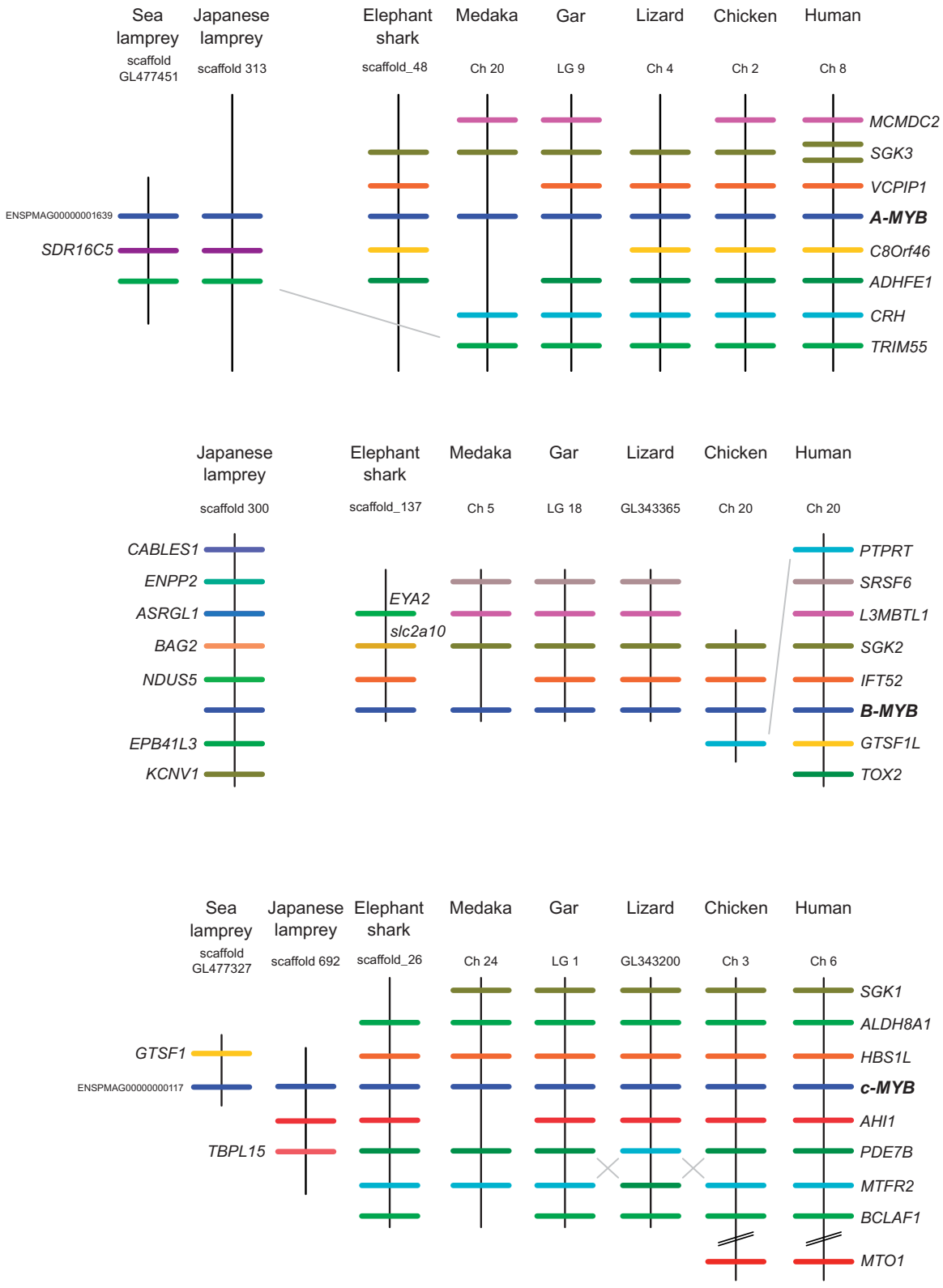

Fig. 3.—Patterns of conserved synteny in genomic regions that harbor paralogous *Myb* genes in representative vertebrate taxa.

FIG. 4.—Alignment of the elephant shark, gar, human, and lamprey Mybs highlighting the three Myb repeats of the DNA-binding domain in the N-terminal portion (in blue), the central transcription activation domain in the center (in green), and the C-terminal regulatory domain (in red).

additional gene models. Phylogenetic analyses place the *A-* and *c-Mybs* of lampreys in strongly supported clades (fig. 2*B*). Thus, we infer that the common ancestor of these two species of lampreys had at least three different *Myb* paralogs, similar to the common ancestor of gnathostomes, and that a future assembly of the sea lamprey genome will probably include a full copy of the lamprey *B-Myb* gene. Like in gnathostomes, there is also some level of conserved synteny among the lamprey *Mybs*. The lamprey *A-Myb* gene is flanked by copies *TRIM55* in the two lampreys (fig. 4). From a structural standpoint, lamprey *Mybs* encode proteins that have the three repeats that correspond to the DNA-binding domain on the N-terminal region and the two C-terminal regulatory domains, and the lamprey *c-Myb* also encodes for a CTAD (fig. 4, supplementary table S4, Supplementary Material online).

Synteny Resolves Orthology among Lamprey and Gnathostome Mybs

Our combined phylogenies of lamprey and gnathostome Mybs place 1) gnathostome *A-* and *B-Mybs* as sister clades in a monophyletic group, 2) lamprey *c-Mybs* as sister to gnathostome *c-Mybs* in a clade labeled as vertebrate *c-Myb* that received moderate support in BA and ML, 3) lamprey *A-* and *B-Mybs* in a clade supported in BA, and 4) vertebrate *c-Mybs* as sister to the clade joining lamprey *A-* and *B-Mybs* with moderate support (supplementary fig. S1*B*, Supplementary Material online). Genomic comparisons reveal patterns of conserved synteny between the *Myb* genes of gnathostomes and lampreys: There is a *TRIM55* gene linked to the *A-Myb* paralogs of cyclostomes and gnathostomes, and there is an *AHI1* paralog linked to the *c-Myb* paralogs of most gnathostomes and the sea lamprey. Thus, orthology for the *c-Myb* paralogs of gnathostomes and lampreys is resolved by phylogenetic, synteny, and structural analyses, but this is not the case for the *A-* and *B-Myb* genes of lampreys and gnathostomes. Topology tests reject a tree where *A-Mybs* of vertebrates are forced together as sister to the *c-Myb* clade ($P \approx 0.047$ in AU topology test, fig. 2*B*).

We speculate that the observed discrepancy between synteny and phylogeny is probably due to the extreme GC-, codon, and amino acid composition bias observed in lampreys (Smith et al. 2013). Notably, phylogenetic and synteny discrepancies have also been observed in the case of the *KCNA* and *Globin X* paralogs of gnathostomes and cyclostomes (Qiu et al. 2011; Schwarze et al. 2014; Opazo et al. 2015). Because the biases listed above are not likely to affect synteny, we infer that the *A-* and *c-Myb* paralogs of lampreys and gnathostomes are orthologous, where the cyclostome *A-Myb* secondarily lost its CTAD. In addition, given that the single-copy *Myb* gene of invertebrates and the *B-Myb* of gnathostomes appear to be functionally equivalent (Davidson et al. 2005) and that the gnathostome *B-Myb* paralog appears to be essential (Tanaka et al. 1999), we hypothesize that the remaining lamprey paralog corresponds to this gene. The corresponding evolutionary scenario is shown on figure 1*B*.

Myb Expression Varies along the Vertebrate tree

Previous studies suggest that the three *Myb* genes show distinct temporal and spatial patterns of expression, which are associated with their different biological functions. In mammals, *c-Myb* is primarily expressed in the immature hematopoietic cells (Gonda et al. 1982; Westin et al. 1982; Duprey and Boettiger 1985), whereas *A-Myb* is predominantly expressed in the male germ cells and breast epithelial cells of pregnant mice (Mettus et al. 1994; Toscani et al. 1997). Thus, *c-Myb* and *A-Myb* are tissue restricted in their expression. On the other hand, *B-Myb* expression has been recorded in mitotically active cells of all tissues (Nomura et al. 1988; Mettus et al. 1994; Trauth et al. 1994; Sitzmann et al. 1996).

To gain an evolutionary perspective on patterns of gene expression of vertebrate *Mybs*, we first focused on comparing the abundance of reads corresponding to each of the three *Mybs* in RNA-seq data from gnathostomes. Specifically, we compared the abundance of reads mapping to *A-*, *B-,* and *c-Myb* genes in brain, heart, kidney, liver, muscle, ovary, spleen, and testis from elephant shark, spotted gar, tilapia, coelacanth, western clawed frog, chicken, gray short-tailed opossum, and human. Our RNA-seq comparisons between the three different *Myb* paralogs within species revealed that, as expected, the *B-Myb* paralog was the most abundantly expressed in most gnathostomes (fig. 5). Comparisons among the different tissue samples indicate that with few exceptions, the three gnathostome *Mybs* were most highly expressed in the gonads and that patterns of gene expression were variable among the different species. *A-Myb* was preferentially expressed in the testis, but in almost all species, *B-* and *c-Myb* were highly expressed in the testis as well, with the exception of elephant shark and spotted gar. Low levels of *A-Myb* expression were previously reported in mouse ovaries, brain, and spleen (Mettus et al. 1994), and we found high expression of *A-Myb* in ovaries and testis for elephant shark and in the brain and heart of spotted gar. The strong expression of gnathostome *A-Myb* in testes is consistent with experimental data from mice that show its involvement in spermatogenesis and piRNA biogenesis. Similarly, the high expression of elephant shark *c-Myb* in spleen is in agreement with its function in hematopoiesis. Comparisons of flanking sequences suggest that the expression pattern similarities between *A-* and *c-Myb* are not driven by the shared presence of putative regulatory elements. We failed to find conserved noncoding elements among the different human *Myb* paralogs, as there were no sequence similarities among any of the 5′-UTRs; however, there were ~300 bp shared between the *A-* and *B-Myb*

flanking sequences, corresponding to insertions of an *Alu* repetitive element.

We then used real-time qPCR to estimate transcript abundance of the *Myb* paralogs in the Japanese lamprey, to provide a first glimpse of expression patterns of this gene family in lampreys and compare it to gnathostomes. Our analyses indicate that the three lamprey *Mybs* are most heavily expressed in the notochord, with lower levels of expression in kidney and ovary. The Japanese lamprey *B-Myb* gene was expressed in testes but at relatively lower levels, and similarly, *A-Myb* was detected at low levels in the liver. Thus, our results suggest that the Japanese lamprey *Mybs* have patterns of expression (fig. 5) that are markedly different from the gnathostome paralogs.

Evolution and Functional Differentiation of Vertebrate Mybs

Integrating phylogenetic, synteny, structural, and expression data from elephant shark, lampreys, plus a representative sample of bony vertebrates, we were able to shed light on the early stages of evolution of the vertebrate *Myb* paralogs. In our unconstrained trees, we found that 1) all vertebrate *Mybs* were monophyletic relative to invertebrate *Mybs*, 2) both lampreys and gnathostomes have three different *Myb* paralogs in their genomes, 3) *c-Mybs* share some synteny, the presence of a central domain and were placed in a monophyletic clade, 4) lamprey and gnathostome *A-Mybs* are linked to copies of *TRIM55*, but lamprey *A-Myb* apparently lacks a CTAD, and 5) gnathostome and lamprey *Mybs* have distinct patterns of gene expression that are not shared between the groups.

If we restrict our analyses to gnathostome *Mybs*, results are straightforward: All three *Mybs* fall in strongly supported monophyletic clades, which are also supported by synteny and structural analyses, with some conservation in gene expression patterns. Integrating lamprey *Mybs* required the use of synteny to resolve orthology for *A-Mybs*, and the assumption that *B-Myb* is essential to gnathostomes to suggest the third *Myb* paralog in lamprey is orthologous to *B-Myb*. Under this assumption, our analyses would indicate that 1) the *Myb* repertoire of vertebrates was established early in their evolutionary history, prior to divergence between lampreys and gnathostomes; 2) the genes have conserved their syntenic position in the corresponding genomes, 3) the functional differentiation of these genes occurred separately in gnathostomes and lampreys, and 4) the A-Myb of lampreys probably lost the CTAD secondarily (fig. 1*B*).

The model outlined by Davidson et al. (2005, 2013, fig. 1*A*) postulates that the *Myb* paralogs derive from the two rounds (2R) of WGD early in the evolution of vertebrates. In agreement with this prediction, there is extensive shared synteny between the gnathostome *Myb* paralogs, and the three human *Mybs* are located in regions of the genome that can be traced back to linkage group four in the inferred karyotype of the common ancestor of amphioxus and human (Putnam et al. 2008). The 2R model predicts the presence of four separate vertebrate paralogs per invertebrate gene. However, because of extensive gene loss after WGD, this is seldom the case (Dehal and Boore 2005). In the case of *Mybs* and the flanking coduplicated gene families, duplications map to the deepest branch of the vertebrate tree we could identify (supplementary fig. S2*B*, Supplementary Material online). The EYA gene family conforms to the 4:1 prediction from the 2R of WGD, but the Myb, PLAG, and SKG gene families show a 3:1 ratio of vertebrate to invertebrate genes which would require the secondary loss of one of the resulting paralogs to reconcile our data with the 2R model. Taken together, our analyses support the role of the 2R of WGD as the source of the vertebrate *Myb* paralogs. In addition, our model extends and refines the one proposed by Davidson et al. (2005, 2013) to include cartilaginous fish and lampreys in addition to bony vertebrates and provides a more complex picture of the evolution and functional differentiation among the paralogs (fig. 1*B*), where distinct expression patterns evolved independently in lampreys and gnathostomes, and even among gnathostomes, we found lineage-specific differences in this regard.

From a physiological standpoint, our analyses suggest that patterns of expression of the *Myb* paralogs are not conserved between lampreys and gnathostomes and probably indicate that different functional roles evolved independently in these two groups. All the *Myb* paralogs in gnathostomes are most heavily expressed in gonad tissues and expression varies among species (fig. 5), whereas the *Myb* paralogs of the Japanese lamprey are most abundantly expressed in the notochord. The clearest evidence of a difference in functional role between lamprey and gnathostome Mybs comes from the *A-Myb* paralog, which is involved with several testis-specific functions in gnathostomes but not expressed in the testes of Japanese lamprey. These results would suggest that the involvement of the *A-Myb* paralog with spermatogenesis and the pachytene piRNA pathway is likely to be a gnathostome-specific innovation.

In vertebrates, piRNAs are broken into two groups, prepachytene piRNAs and pachytene piRNAs. Prepachytene piRNAs are expressed in premeiotic and early prophase 1 germ cells and play a role in TE expression regulation (Aravin et al. 2007), similar to the role piRNAs play in *Drosophila* (Brennecke et al. 2007). However, pachytene piRNAs become expressed in the pachytene stage of prophase 1 through maturation and are mostly derived from intergenic transcripts regulated by *A-Myb* (Li et al. 2013). The latter piRNAs appear to have a role regulating and eliminating gene transcripts from the cytoplasm, in a manner resembling that of the miRNA pathway (Gou et al. 2014). Given that *A-Myb* is a driver of the production of pachytene piRNAs in mature testes, we would predict this class of small RNAs would be absent from lampreys. This difference might be

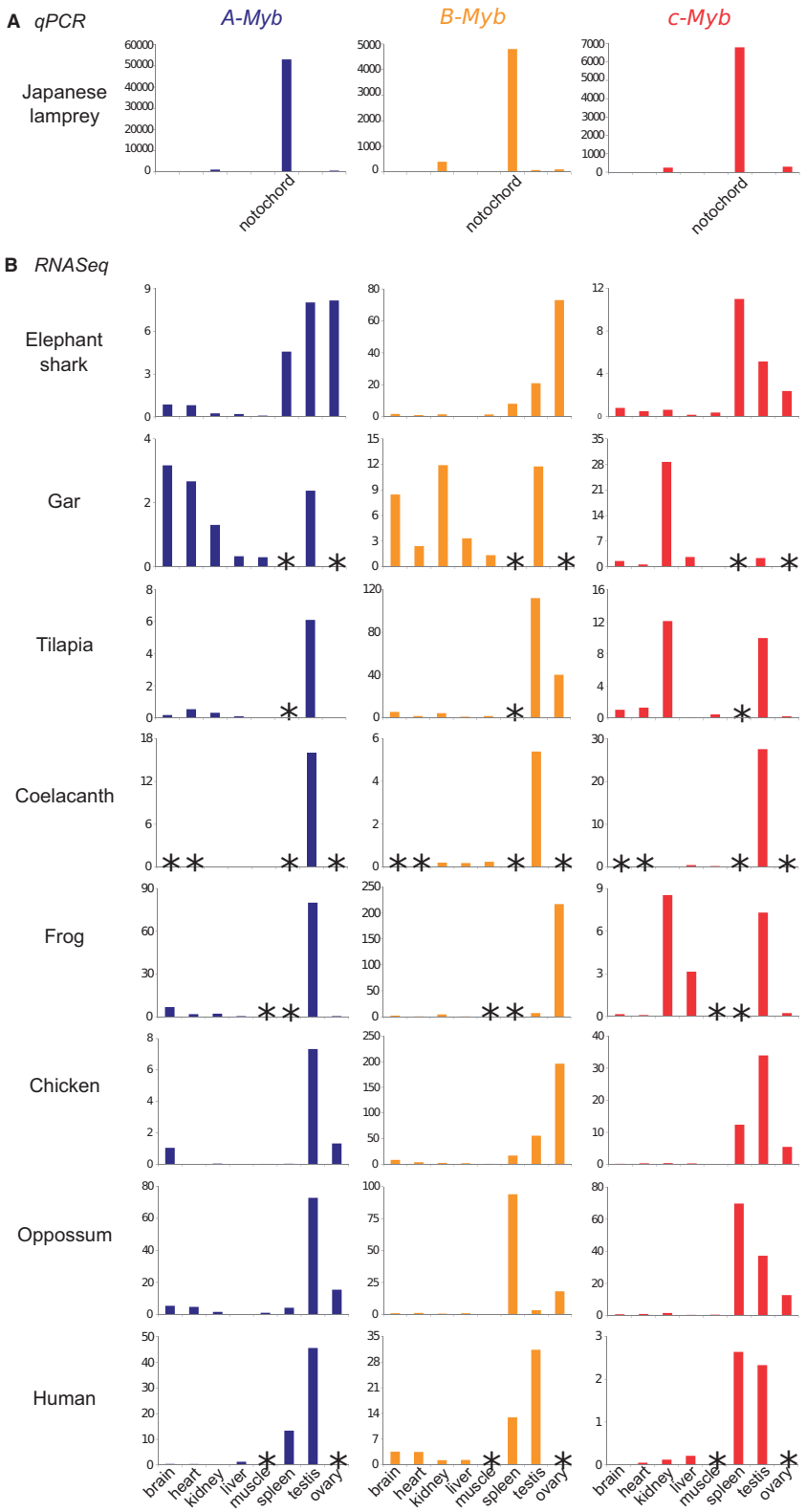

Fig. 5.—Comparative expression profiles of vertebrate *Mybs* across multiple tissues. (*A*) Comparative expression of Japanese lamprey *Mybs* estimated via qPCR. Mean standard deviations were less than 1.04 and not displayed. (*B*) Comparative expression of gnathostome *Mybs*, where gene-specific mRNA levels were quantified using RNA-Seq. Transcript abundances are measured in transcripts per million (TPM). Asterisks indicate tissues for which data were not available.

related to the secondary loss of the CTAD by the A-Myb protein of cyclostomes. Thus, our analyses would suggest that gnathostome and cyclostome *Mybs* have evolved different functions, despite a common evolutionary origin. Further, it remains to be checked whether the *B-Myb* paralog of lampreys is also functionally equivalent to the single copy *Myb* gene of invertebrates, as is the *B-Myb* gene of gnathostomes, or whether the two lamprey paralogs without a CTAD, *A-* and *B-Myb*, have partitioned that functional role in a different manner. Unlike most gnathostomes, cyclostomes retain the notochord as adults, and the high expression of the cyclostome *Myb* paralogs in the notochord might be related to this retention.

Hidden paralogy or the differential evolution of the gene complements of cyclostomes and gnathostomes from a common ancestor represents an important challenge when reconstructing the phenotype of their common ancestor (Kuraku 2013). Our results illustrate an additional facet of this challenge, as a conserved set of paralogs have evolved specialized functional roles independently in cyclostomes and gnathostomes. As a further note of caution, it is worth pointing out that lampreys might have undergone an additional round of WGD after divergence from gnathostomes (Mehta et al. 2013), which might present an additional layer of difficulty in reconciling our phylogenetic and synteny analyses. A more detailed analysis of the evolution of cyclostome *Mybs*, including hagfish genomes, might result in alternative explanations for the observed patterns of relationships between lamprey and gnathostome *Mybs*.

Supplementary Material

Supplementary figures S1–S2 and tables S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the National Science Foundation (EPS-0903787, DBI-1262901 DEB-1354147) to F.G.H., the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (99999.004252/2014-04) and the Fundação de Amparo à Pesquisa do Estado de São Paulo (2012/17160-8) to E.B.C., and a grant from the Biomedical Research Council at the Agency for Science, Technology and Research to B.V. The authors thank Juan C. Opazo and two anonymous reviewers for helpful suggestions.

Literature Cited

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

Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. 2007. Developmentally regulated piRNA clusters implicate MILI in transposon control. Science 316:744–747.

Bolcun-Filas E, et al. 2011. A-MYB (MYBL1) transcription factor is a master regulator of male meiosis. Development 138:3319–3330.

Braasch I, Volff J, Schartl M. 2009. The endothelin system: evolution of vertebrate-specific ligand-receptor interactions by three rounds of genome duplication. Mol Biol Evol. 26:783–799.

Brennecke J, et al. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128:1–15.

Cunningham F, et al. 2015. Ensembl 2015. Nucleic Acids Res. 43:D662–D669.

Davidson CJ, Guthrie EE, Lipsick JS. 2013. Duplication and maintenance of the *Myb* genes of vertebrate animals. Biology Open 2:101–110.

Davidson CJ, Tirouvanziam R, Herzenberg LA, Lipsick JS. 2005. Functional evolution of the vertebrate Myb gene family: B-Myb, but neither A-Myb nor c-Myb, complements *Drosophila* Myb in hemocytes. Genetics 169:215–229.

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

Duprey SP, Boettiger D. 1985. Developmental regulation of *c-Myb* in normal myeloid progenitor cells. Proc Natl Acad Sci U S A. 82:6937–6941.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Gonda TJ, Shieness DK, Bishop JM. 1982. Transcripts from the cellular homologs of retroviral oncogenes: distribution among chicken tissues. Mol Cell Biol. 2:617–624.

Gou L-T, et al. 2014. Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. Cell Res. 24:680–700.

Hoffmann FG, Opazo JC, Storz JF. 2012. Whole-genome duplications spurred the functional diversification of the globin gene superfamily in vertebrates. Mol Biol Evol. 29:303–312.

Houwing S, et al. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in zebrafish. Cell 129:69–82.

Jobb G, von Haeseler A, Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol. 4:18.

Katoh K, Asimenos G, Toh H. 2009. Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol. 537:39–64.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772–780.

Kuraku S. 2013. Impact of asymmetric gene repertoire between cyclostomes and gnathostomes. Semin Cell Dev Biol. 24:119–127.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.

Larhammar D, Nordstrom K, Larsson TA. 2009. Evolution of vertebrate rod and cone phototransduction genes. Philos Trans R Soc. 364:2867–2880.

Lassmann T, Sonnhammer ELL. 2006. Kalign, Kalignvu and Mumsa: web servers for multiple sequence alignment. Nucleic Acids Res. 34:W596–W599.

Lau NC, et al. 2006. Characterization of the piRNA complex from rat testes. Science 313: 363–367.

Letunic I, Doerks T, Bork P. 2015. SMART: recent updates, new developments and status in 2015. Nucleic Acids Res. 43:D257–D260.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics 12:323.

Li XZ, et al. 2013. An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. Mol Cell. 50:67–81.

Lipsick JS, et al. 2001. Functional evolution of the *Myb* oncogene family. Blood Cells Mol Dis. 27:456–458.

Louis A, Nguyen NTT, Muffato M, Crollius HR. 2015. Genomicus update 2015: KaryoView and MatrixView provide a genome-wide

perspective to multispecies comparative genomics. Nucleic Acids Res. 43:D682–D689.
Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.
Mehta TK, et al. 2013. Evidence for at least six Hox clusters in the Japanese lamprey (*Lethenteron japonicum*). Proc Natl Acad Sci U S A. 110:16044–16049.
Mettus RV, et al. 1994. Murine *A-Myb*: evidence for differential splicing and tissue-specific expression. Oncogene 9:3077–3086.
Mucenski ML, et al. 1991. A functional *c-Myb* gene is required for normal murine fetal hepatic hematopoiesis. Cell 65:677–689.
Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 39:121–152.
Nomura N, et al. 1988. Isolation of human cDNA clones of *Myb*-related genes, *A-Myb* and *B-Myb*. Nucleic Acids Res. 16:11075–11089.
Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302:205–217.
Oh IH, Reddy EP. 1998. The C-terminal domain of B-Myb acts as a positive regulator of transcription and modulates its biological functions. Mol Cell Biol. 18:499–511.
Oh IH, Reddy EP. 1999. The *Myb* gene family in cell growth, differentiation and apoptosis. Oncogene 18:3017–3033.
Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.
Opazo JC, et al. 2015. Ancient duplications and expression divergence in the globin gene superfamily of vertebrates: insights from the elephant shark genome and transcriptome. Mol Biol Evol. 32:1684–1694.
Putnam NH, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064–1071.
Qiu H, Hildebrand F, Kuraku S, Meyer A. 2011. Unresolved orthology and peculiar coding sequence properties of lamprey genes: the *KCNA* gene family as test case. BMC Genomics 12:325.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.
Sakura H, et al. 1989. Delineation of three functional domains of the transcriptional activator encoded by the *c-Myb* protooncogene. Proc Natl Acad Sci U S A. 86:5758–5762.
Sala A. 2005. B-MYB, a transcription factor implicated in regulating cell cycle, apoptosis and cancer. Eur J Cancer. 41:2479–2484.
Schwarze K, et al. 2014. The globin gene repertoire of lampreys: convergent evolution of hemoglobin and myoglobin in jawed and jawless vertebrates. Mol Biol Evol. 31:2708–2721.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 51:492–508.
Sitzmann J, Noben-Trauth K, Kamano H, Klempnauer KH. 1996. Expression of *B-Myb* during mouse embryogenesis. Oncogene 12:1889–1894.
Smith JJ, et al. 2013. Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. Nat Genet. 45:415–421.
Soza-Ried C, Hess I, Netuschil N, Schorpp M, Boehm T. 2010. Essential role of c-Myb in definitive hematopoiesis is evolutionarily conserved. Proc Natl Acad Sci U S A. 107:17304–17308.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics Analysis version 6.0. Mol Biol Evol. 30:2725–2729.
Tanaka Y, Patestos NP, Maekawa T, Ishii S. 1999. B-Myb is required for inner cell mass formation at an early stage of development. J Biol Chem. 274:28067–28070.
Tarasov KV, et al. 2008. B-MYB is essential for normal cell cycle progression and chromosomal stability of embryonic stem cells. PLoS One 3:e2478.
Toscani A, et al. 1997. Arrest of spermatogenesis and defective breast development in mice lacking *A-Myb*. Nature 386:713–717.
Trauth K, et al. 1994. Mouse A-Myb encodes a trans-activator and is expressed in mitotically active cells of the developing central nervous system, adult testis and B lymphocytes. EMBO J. 13:5994–6005.
Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10:725–732.
Venkatesh B, et al. 2014. Elephant shark genome provides unique insights into gnathostome evolution. Nature 505:174–179.
Westin EH, et al. 1982. Differential expression of the *amv* gene in human hematopoietic cells. Proc Natl Acad Sci U S A. 79:2194–2198.
Zhang G, Cohn MJ. 2008. Genome duplication and the origin of the vertebrate skeleton. Curr Opin Genet Dev. 18:387–393.
Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18(6):292–298.

Associate editor: Naruya Saitou