Ancient Duplications and Expression Divergence in the Globin Gene Superfamily of Vertebrates: Insights from the Elephant Shark Genome and Transcriptome

Juan C. Opazo,*†,1 Alison P. Lee,‡,2 Federico G. Hoffmann,†3,4 Jessica Toloza-Villalobos,1 Thorsten Burmester,5 Byrappa Venkatesh,*,2,6 and Jay F. Storz2,*

1Instituto de Ciencias Ambientales y Evolutivas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile
2Institute of Molecular and Cell Biology, Comparative and Medical Genomics Laboratory, Agency for Science, Technology, and Research Biopolis, Singapore
3Department of Biochemistry, Molecular Biology, Entomology, and Plant Pathology, Mississippi State University
4Institute for Genomics, Biocomputing, and Biotecnology, Mississippi State University
5Institute of Zoology, University of Hamburg, Hamburg, Germany
6Departments of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
7School of Biological Sciences, University of Nebraska, Lincoln

These authors contributed equally to this work.

*Corresponding author: E-mail: jopazo@gmail.com; mcbbv@imcb.a-star.edu.sg; jstorz2@unl.edu.

Abstract

Comparative analyses of vertebrate genomes continue to uncover a surprising diversity of genes in the globin gene superfamily, some of which have very restricted phyletic distributions despite their antiquity. Genomic analysis of the globin gene repertoire of cartilaginous fish (Chondrichthyes) should be especially informative about the duplicative origins and ancestral functions of vertebrate globins, as divergence between Chondrichthyes and bony vertebrates represents the most basal split within the jawed vertebrates. Here, we report a comparative genomic analysis of the vertebrate globin gene family that includes the complete globin gene repertoire of the elephant shark (Callorhinchus milii). Using genomic sequence data from representatives of all major vertebrate classes, integrated analyses of conserved synteny and phylogenetic relationships revealed that the last common ancestor of vertebrates possessed a repertoire of at least seven globin genes: single copies of androglobin and neuroglobin, four paralogous copies of globin X, and the single-copy progenitor of the entire set of vertebrate-specific globins. Combined with expression data, the genomic inventory of elephant shark globins yielded four especially surprising findings: 1) there is no trace of the neuroglobin gene (a highly conserved gene that is present in all other jawed vertebrates that have been examined to date), 2) myoglobin is highly expressed in heart, but not in skeletal muscle (reflecting a possible ancestral condition in vertebrates with single-circuit circulatory systems), 3) elephant shark possesses two highly divergent globin X paralogs, one of which is preferentially expressed in gonads, and 4) elephant shark possesses two structurally distinct α-globin paralogs, one of which is preferentially expressed in the brain. Expression profiles of elephant shark globin genes reveal distinct specializations of function relative to orthologs in bony vertebrates and suggest hypotheses about ancestral functions of vertebrate globins.

Key words: Chondrichthyes, gene duplication, family evolution, globin, myoglobin, neuroglobin.

Introduction

Analyses of vertebrate genomes have led to the discovery of many new members of the globin gene superfamily (Hankeln et al. 2005; Hankeln and Burmester 2008; Burmester and Hankeln 2009, 2014; Storz et al. 2011, 2013). The canonical globin gene repertoire of jawed vertebrates (gnathostomes) comprises eight main elements, each of which may be represented by multiple, structurally distinct paralogs with distinct expression domains that are specific to different tissues or cell types: neuroglobin (Ngb), cytoglobin (Cygb), androglobin (Adgb), globin E (GbE), globin X (GbX), globin Y (GbY), myoglobin (Mb), and the α- and β-type subunits of hemoglobin (Hb). All vertebrate globins (Euteleostomi, which includes osteichthyans + tetrapods) that have been examined to date possess copies of Ngb, Cygb, and Adgb (Burmester et al. 2000, 2002, 2004; Awenuis et al. 2001; Pesce et al. 2002; Trent and Hargrove 2002; Fuchs et al. 2004, 2005; Kugelstadt et al. 2004; Wystub et al. 2004; Hankeln et al. 2005; Roesner et al. 2005; Hankeln and Burmester 2008; Burmester and Hankeln 2009, 2014; Hoffmann, Opazo, et al. 2010; Kakar et al. 2010; Hoogewijs et al. 2012; Schwarze and Burmester 2013). In bony vertebrates, the monomeric Ngb protein is expressed in the neurons of the central and peripheral nervous system, and in some endocrine tissues, whereas the homodimeric Cygb protein is expressed in fibroblasts and related cell types and in distinct nerve cells in the central
and peripheral nervous systems (reviewed by Hankeln and Burmester 2008; Burmester and Hankeln 2009, 2014). The more recently discovered Adgb gene is especially enigmatic, as it is a chimeric fusion gene; the encoded protein has an N-terminal calpain-like domain, an internal globin domain that has undergone an internal shuffling of α-helical subdomains, and an IQ calmodulin-binding motif (Hoogewijs et al. 2012). In mammals, the Adgb gene is preferentially expressed in testis (Hoogewijs et al. 2012). The heme coordination chemistries and other structural features of Ngb, Cygb, and Adgb suggest that these globins may perform redox-regulated signaling functions or oxygen-sensing functions that mediate oxygen-dependent protein activities, but the primary physiological functions of these globin proteins are still mostly shrouded in mystery (Burmester and Hankeln 2014).

In the majority of bony vertebrate taxa, tetrameric Hb and monomeric Mb play key roles in the maintenance of cellular oxygen supply to fuel aerobic metabolism. The proto-Hb gene and the single-copy ancestor of the Mb/GbE gene pair were paralogous products of a whole-genome duplication event that occurred in the stem lineage of vertebrates (Hoffmann, Opazo, Storz, et al. 2012). A subsequent tandem duplication of the proto-Hb gene gave rise to the progenitors of the α- and β-globin gene subfamilies. This tandem gene duplication event occurred in the stem lineage of gnathostomes prior to the split between cartilaginous fish and bony vertebrates (Hoffmann, Opazo, Storz, et al. 2012) which is estimated to have occurred approximately 450–500 Ma in the early Paleozoic (Hedges 2009; Inoue et al. 2010). The α2β2 Mb tetramer is most highly expressed in erythrocytes and is responsible for transporting oxygen in arterial blood from the respiratory surfaces (lungs, gills, or skin surface) to the cells of respiring tissues, and for transporting carbon dioxide in the venous blood from the tissues back to the gas exchange surfaces. In addition to the familiar respiratory function of Hb in blood-gas transport, recent reports of Hb expression in non-erythroid cell types suggest possible “moonlighting” functions that may involve the scavenging of reactive oxygen species (Nishi et al. 2008) or regulation of nitric oxide signaling (Straub et al. 2012). Mb is primarily expressed in myocytes of cardiac and skeletal muscle, where it regulates cellular oxygen tension and the bioavailability of nitric oxide (Wittenberg and Wittenberg 2003; Helbo et al. 2013). Similar to the case with Hb, recent discoveries of Mb expression in vascular smooth muscle and endothelial cells suggest that this protein may play a more versatile role than previously thought (Qiu et al. 1998; Wittenberg and Wittenberg 2003; Cossins et al. 2009; Helbo et al. 2013).

In addition to the Ngb, Cygb, Adgb, Hb, and Mb genes that have been retained in all or nearly all bony vertebrate lineages, a number of paralogous globins have been discovered that have far more restricted phyletic distributions. For example, one or more copies of GbX have been documented in lampreys and most examined bony vertebrates other than mammals and archosaurs (Roesner et al. 2005; Fuchs et al. 2006; Dröge and Makalowski 2011; Blank and Burmester 2012; Schwarze and Burmester 2013; Schwarze et al. 2014). Bony vertebrate GbX is a monomeric, membrane-bound globin that possesses N-terminal acylation sites, which suggests possible roles in cellular signaling and/or protection against the oxidation of membrane lipids (Blank, Wollberg et al. 2011; Blank and Burmester 2012). The GbY gene has been documented in the genomes of lobe-finned fishes (coelacanth, Latimeria chalumnae) and several tetrapod taxa such as the anole lizard (Anolis carolinensis) and platypus (Ornithorhynchus anatinus), but the gene does not appear to be present in lampreys, teleost fishes, birds, or therian mammals (marsupials and eutherians) (Hoffmann, Opazo, et al. 2010; Hoffmann, Storz, et al. 2010; Hoffmann et al. 2011; Schwarze and Burmester 2013; Schwarze et al. 2014). In adult Xenopus, GbY is expressed in a broad range of tissue types (Fuchs et al. 2006), but its physiological function has yet to be elucidated. The GbE gene was initially found only in birds (Kugelstadt et al. 2004; Hoffmann, Opazo, et al. 2010; Blank, Kiger, et al. 2011; Hoffmann et al. 2011). The monomeric GbE protein appears to perform a Mb-like function in regulating oxygen supply to photoreceptor cells in the avascular avian retina (Blank, Kiger, et al. 2011), although a role in regulating cellular redox homeostasis is also possible. Although initial genomic surveys suggested that GbE represented a bird-specific globin, phylogenetic analyses combined with assessments of conserved synteny revealed that GbE and Mb represent the paralogous products of a tandem gene duplication that occurred in the stem lineage of gnathostomes (Hoffmann et al. 2011; Hoffmann, Opazo, Storz, et al. 2012). The antiquity of the gene suggested the possibility that orthologs of GbE might be found in nonavian vertebrate taxa that had yet to be surveyed. Sure enough, a recent comparative genomic study identified orthologs of the hitherto “bird-specific” globin in the coelacanth (Schwarze and Burmester 2013).

The Ngb, Adgb, and GbX genes originated prior to the split between protostomes and deuterostomes, and were inherited by both descendant lineages (Burmester et al. 2000; Blank and Burmester 2012; Hoffmann, Opazo, Hoogewijs, et al. 2012; Hoogewijs et al. 2012; Burmester and Hankeln et al. 2013. Vertebrate Ngb shows phylogenetic affinities with a diversity of nerve globins in annelid worms, echinoderms, and cephalochordates (amphioxus), and vertebrate GbX shows affinities with an equally diverse set of putatively N-acylated globins in protostomes and nonvertebrate deuterostomes (Blank and Burmester 2012; Hoffmann, Opazo, Hoogewijs, et al. 2012). The remaining members of the vertebrate globin gene repertoire are products of vertebrate-specific gene duplications or whole-genome duplications (Hoffmann, Storz, et al. 2010; Hoffmann et al. 2011, 2013; Hoffmann, Opazo, Hoogewijs, et al. 2012; Hoffmann, Opazo, Storz, et al. 2012). Cyclostomes (jawless fishes represented by lampreys and hagfish), the sister group of gnathostomes, possess a surprisingly diverse repertoire of globin genes, but available data indicate that Cygb is the only vertebrate-specific globin shared between cyclostomes and gnathostomes (Schwarze et al. 2014). Analyses of available globin sequences from sharks and rays have demonstrated that cartilaginous fish and bony vertebrates share orthologous α- and β-type Hb genes (Hoffmann, Opazo,
Hoogewijs, et al. 2012; Hoffmann, Opazo, Storz, et al. 2012; Opazo et al. 2013), but it remains possible that cartilaginous fish and cyclostomes share orthologs of other globin types that were secondarily lost in the ancestor of bony vertebrates. It is possible that additional globin genes still await discovery as full genome sequences become available for more vertebrate taxa. Among vertebrates, it should be especially enlightening to characterize the complete globin gene repertoire for a representative of Chondrichthyes (cartilaginous fish), as divergence between Chondrichthyes and bony vertebrates represents the most basal split within the gnathostomes. Comparative expression analyses of orthologous genes shared between cartilaginous fish and bony vertebrates could potentially shed light on ancestral functions and could reveal whether certain physiological functions of modern-day globins represent derived, “repurposed” modifications of ancestral functions. Here, we report a comparative genomic analysis that includes the globin gene repertoire of the elephant shark (Callorhinus milius), the first cartilaginous fish to have its full genome sequenced (Venkatesh et al. 2007, 2014). The main objectives of the study were 1) to unravel the duplicative history of globin genes during the early evolution of vertebrates, 2) to reconstruct the globin gene repertoire of the vertebrate common ancestor, and 3) to test whether orthologous globins shared between cartilaginous fish and bony vertebrates have conserved patterns of tissue-specific expression.

**Results and Discussion**

Using gene and genome sequences from representatives of all major vertebrate classes (Materials and Methods), we integrated analyses of conserved synteny and phylogenetic relationships to unravel the evolutionary history of vertebrate globin genes. Combined with expression data, the inventory of globin genes in the elephant shark genome yielded four especially surprising findings: 1) there is no trace of the Ngb gene (a highly conserved gene that is possessed by all other gnathostome taxa that have been examined to date), 2) Mb is highly expressed in heart, but not in skeletal muscle, 3) elephant shark possesses two highly divergent GbX paralogs, one of which is preferentially expressed in gonads (ovaries and testis), and 4) elephant shark possesses two structurally distinct α-globin paralogs, one of which is preferentially expressed in the brain. The divergent expression domains of the two GbX paralogs and the two α-globin paralogs have not been previously documented in any vertebrate taxon.

**Phylogenetic Relationships**

In broad outline, our reconstruction of phylogenetic relationships among vertebrate globins was consistent with previous studies (Hoffmann, Opazo, et al. 2010; Blank and Burmester 2012; Hoffmann, Opazo, Hoogewijs, et al. 2012; Hoffmann, Opazo, Storz, et al. 2012; Schwarze and Burmester 2013; Schwarze et al. 2014). Vertebrate globins were arranged into three separate clades, two representing the ancient GbX and Ngb gene lineages and one representing the vertebrate-specific globins, including gnathostome α- and β-globin, GbE, GbY, and Mb, cyclostome Hbs and Mbs, and the Cygb genes of gnathostomes and cyclostomes (fig. 1). Support for monophyly of the three clades of vertebrate globins, expressed as Bayesian posterior probabilities, was higher than 0.99, and support for monophyly of the different vertebrate globin types was also high, with posterior probabilities higher than 0.99 in all cases except gnathostome α-globins (fig. 1). Compared with a recent phylogenetic analysis of vertebrate globins presented in Schwarze et al. (2014), the tree shown in figure 1 shows relatively strong support for the node joining cyclostome and gnathostome Cygb (posterior probability = 0.91). Vertebrate-specific globins are arranged into three strongly supported clades containing 1) cyclostome Hbs, 2) cyclostome and gnathostome Cygb, and 3) gnathostome α- and β-globin, GbE, GbY, and Mb.

Two of the nine elephant shark globins were placed in distinct GbX clades, and the remainder fell in well-supported clades of vertebrate-specific globins (fig. 1). None of the elephant shark globins was placed in the Ngb clade. For the majority of genes, phylogenetic placement of the elephant shark globins provided a clear means of inferring orthology; we identified two α-globin paralogs, two GbX paralogs, and single copies of β-globin, GbY, and Mb. Within the clade of vertebrate-specific globins, the position of the elephant shark globins matched expectations of the organismal phylogeny: the α/β-Hb and Mb genes of the elephant shark were grouped with corresponding sequences from other cartilaginous fish, and elephant shark GbY was placed sister to GbY sequences from bony vertebrates. Assigning orthologous relationships of the GbX sequences was more complicated, as one of the two elephant shark GbX paralogs was placed sister to cyclostome GbX genes, and the second was placed sister to a group that included teleost GbXs plus the coelacanth GbX2 paralog (fig. 1).

A comparison of the organismal phylogeny with the phylogeny of GbX genes in figure 1 suggests the presence of multiple GbX paralogs in the common ancestor of vertebrates. To better address this issue, we performed a second phylogenetic analysis restricted to GbX paralogs, with increased taxonomic sampling of GbX sequences relative to the tree in figure 1, and using a group of relatively closely related acorn worm globins as outgroup. In this second analysis, GbX paralogs fell into four main clades: 1) GbX sequences from western clawed frog, teleost fish, plus one of the two GbX paralogs present in the gar, coelacanth, and elephant shark genomes, 2) GbX sequences from squamates and testudines, plus the GbX1 paralogs of gar and coelacanth, 3) a group that includes all GbX paralogs of cyclostomes, and 4) the second elephant shark globin paralog, labeled as GbX2 (fig. 2). Relationships among these lineages are well supported, and a comparison of the gene tree with the organismal phylogeny strongly suggests that the common ancestor of vertebrates possessed at least four GbX paralogs.

**Patterns of Conserved Synteny**

Previous studies of globin phylogeny and conserved synteny have shown that the diversification of vertebrate-specific
Fig. 1. Bayesian phylogram depicting relationships among globin genes from representative deuterostome taxa. Numbers next to the nodes are Bayesian posterior probabilities. Terminal branches leading to elephant shark globins are shown in red. The tree was rooted by using plant globins as outgroup sequences (see Materials and Methods). The inset on top shows the organismal phylogeny for the representative deuterostome taxa included in the analysis.
globins is largely attributable to two rounds of whole-genome duplication that occurred in the stem lineage of vertebrates (Hoffmann, Opazo, Storz, et al. 2012). The hypothesis that the vertebrate ancestor possessed a total of four GbX paralogs is consistent with the expected outcome of two successive rounds of whole-genome duplication (Meyer and Schartl 1999). If the four GbX paralogs represent quadruplicated products of the “2 R” genome duplications in the vertebrate ancestor, then this should be reflected in a 4-fold pattern of intragenomic synteny. Specifically, each of the four GbX paralogs should be embedded in unlinked chromosomal segments that share similar, interdigitated arrangements of paralogous genes (Storz et al. 2013). To test this, we examined the physical linkage arrangements of the GbX paralogs and we characterized patterns of conserved synteny among vertebrates. We excluded teleost fish from this analysis because this group has undergone an additional round of whole-genome duplication, which greatly complicates synteny comparisons.

Our comparative genomic analyses revealed conserved synteny for two of the four GbX groups defined in the phylogeny (figs. 2 and 3). One of these GbX paralogs is flanked by PLEKHG3 and SRP14 in most species examined, whereas a second one is flanked by TIMM50 and SUPTSH. In coelacanth and spotted gar, this latter GbX paralog is also flanked by PLEKHG2, a paralog of PLEKHG3. Thus, the linked PLEKHG and GbX genes appear to have coduplicated, possibly during one or two rounds of whole-genome duplication prior to the divergence of cyclostomes and gnathostomes. With the notable exception of the GbX1 sequences from lampreys, shared synteny was reflected in the phylogenetic reconstructions; the GbX paralogs that are flanked by PLEKHG3 and SRP14 on either side were phylogenetically placed with the GbX paralogs from teleost fish, whereas the GbX paralogs that are flanked by TIMM50 and SUPTSH2 were placed with the GbX paralogs from squamates and testudines. The GbX1 paralogs of lamprey deviate from this pattern, as they are flanked by PLEKHG3 and SRP14 on either side, but they were phylogenetically placed with another cyclostome paralog, GbX2. Although the inferred sister relationship between the cyclostome GbX1 and GbX2 paralogs has high statistical support (posterior probability = 0.99; fig. 2), such results should be interpreted cautiously due to the possible influence of base-composition bias (Qiu et al. 2011; Mehta et al. 2013; Smith et al. 2013) and other factors. Unfortunately, there is no synteny information for the cyclostome GbX2.
genes or the elephant shark GbX2 gene, as the cyclostome genes are found in isolated contigs and the elephant shark gene is not present in the genome assembly. Additional genomic data from cartilaginous fishes and cyclostomes will be required to assess which GbX paralogs might have originated via whole-genome duplication. Specifically, synteny information for the chromosomal region harboring orthologs of elephant shark GbX2 in other cartilaginous fish may reveal whether this gene represents a product of the 2 R whole-genome duplication that was secondarily lost in the ancestor of bony vertebrates.

Loss of Ngb in Chondrichthyes

All gnathostome vertebrates that have been investigated so far possess a Ngb gene, which is expressed in the nervous system. It was therefore surprising that Ngb was not detected in the genome or brain transcriptome of the elephant shark. The gene synten POMT2 - Ngb - TMEM63C - ZDHHC22 is conserved among vertebrates; in the elephant shark, the genes POMT2, TMEM63C, and ZDHHC22 were found on scaffold_319 in the expected order, but with Ngb missing. Further database searches did not reveal any evidence for a Ngb gene in the genome and transcriptome of the little skate (Leucoraja erinacea), or in the transcriptomes of the cloudy catshark (Scyllorhinus torazame), the spiny dogfish shark (Squalus acanthias), the nurse shark (Ginglymostoma cirratum), or the Pacific electric ray (Torpedo californica), suggesting that Ngb was lost in the stem lineage of Chondrichthyes.

The Globin Gene Repertoire of the Vertebrate Common Ancestor

Comparative analysis of complete globin gene repertoires from representatives of all vertebrate classes indicates that the last common ancestor of vertebrates possessed a repertoire of at least seven globin genes: single copies of Adgb and Ngb, four paralogous copies of GbX, and the single-copy progenitor of the entire set of vertebrate-specific globins (α1β-Hb, Mb, Gbe, GbY, Cygb, and the cyclostome Hbs and Mbs). During the course of vertebrate evolution, GbY, Mb, Gbe, GbX, and Ngb have each been lost in at least two different lineages (fig. 4). Given that GbX is present in the genomes of turtles and squamate reptiles (fig. 4), the principle of parsimony suggests that this gene must have been deleted independently in mammals and archosaurs (represented by birds and crocodilians). Cygb and Adgb are the only globins that have been retained in all vertebrate taxa examined to date.

Divergent Expression Domains of Paralogous Globins

Some of the elephant shark globins displayed expression profiles that were consistent with experimental results from orthologous globins in bony vertebrates. For example, Adgb was preferentially expressed in testis (fig. 5A), consistent with experimental results for mammals (Hoogewijs et al. 2012). Cygb was expressed in a broad range of tissues, including brain (fig. 5B), although overall expression levels were low relative to other bony vertebrates (Nakatani et al. 2004; Schmidt et al. 2004; Motoyama et al. 2014). GbY was expressed in a similarly broad range of tissues (fig. 5C), consistent with data from Xenopus (Fuchs et al. 2006). In contrast, the elephant shark Mb, α-globin, and GbX genes exhibited highly unexpected expression patterns.

Although Mb is most highly expressed in the myocytes of cardiac and skeletal muscle in most bony vertebrates, elephant shark Mb is expressed at a high level only in heart, but not in skeletal muscle (fig. 5D). The Mb gene has been deleted in anuran amphibians (Fuchs et al. 2006; Hoffmann et al. 2011), and it has been hypothesized that Cygb performs a compensatory Mb-like O2-storage function in skeletal muscle (Xi et al. 2007). In elephant shark there is no detectable expression of Cygb in skeletal muscle (fig. 5B), so the compensatory mechanism hypothesized for anuran amphibians is clearly not operative in elephant shark. It may be that the ancestral function of Mb was to facilitate oxygen delivery to cardiac myocytes in early gnathostomes that had single-circuit circulatory systems. The prominent expression of Mb in the skeletal muscle of many gnathostomes may have been a later innovation that coincided with the evolution of higher rates of aerobic metabolism and the concomitant increase in tissue oxygen demands.

One of the elephant shark α-globin paralogs (α1) and the sole β-globin gene showed highest expression in heart and spleen (fig. 5E and G), as expected for subunit isoforms of a tetrameric α1β2 red blood cell Hb. Surprisingly, the other α-globin paralog (α2) was most highly expressed in the brain (fig. 5F). Expression of heterotetrameric Hb in brain neurons has been reported in mammals (Humphries et al. 1976; Ohyagi et al. 1994; Russo et al. 2013) and is associated with an enhanced neuronal oxygen supply (Schelshorn et al. 2009; Biagioli et al. 2009). The expression of a single Hb chain in
the shark brain provides evidence that this globin acts as a monomer, which excludes an Hb-like physiological function involving cooperative oxygen binding. It is intriguing that the brain-specific expression of this Hb chain is associated with deletion of the Ngb gene, and it suggests the hypothesis that α-globin 2 was co-opted to perform the brain-specific physiological functions that are performed by Ngb in bony vertebrates.

The two elephant shark GbX paralogs exhibited a surprising divergence in expression domains. Although GbX1 was expressed in a diverse range of tissues, GbX2 was most highly expressed in ovary and testis (fig. 5H and I). Studies of zebrafish and Xenopus documented predominant expression of GbX in the nervous system (Fuchs et al. 2006; Blank, Wollberg, et al. 2011). However, those studies did not examine gene expression in the gonads, which suggests that the main
site of \( GbX \) expression may have been overlooked. Alternatively, \( GbX \) genes may have evolved different expression domains in cartilaginous fish and bony vertebrates.

**Evolution of Vertebrate Globins**

Historically, much of the uncertainty about the phylogenetic relationships among vertebrate globins was attributable to inadequate sampling of taxa and/or paralogous gene lineages. In recent years, the inclusion of a broader diversity of vertebrate globins in phylogenetic analyses has led to surprising findings regarding the antiquity of “taxon-specific” globins like \( GbE \), which often turn out to be older than originally suspected (Hoffmann et al. 2011; Schwarze and Burmester 2013), and the independent origins of functionally analogous globins in disparate taxa that were previously assumed to be orthologous. For example, it was previously assumed that \( Mb \) and the single-copy progenitor of the \( \alpha/\beta \)-Hbs originated via duplication of an ancestral, single-copy globin gene prior to the divergence between cyclostomes and gnathostomes, such that the ancestors of these two groups inherited orthologous copies of the same “proto-Hb” gene (Goodman et al. 1975, 1987). According to this scenario, the \( Hb \) genes of cyclostomes would be sister to the clade of gnathostome \( \alpha \)- and \( \beta \)-globin genes: (\( Mb \) [cyclostome \( Hb \), gnathostome \( Hb \)]). Only by including the full panoply of cyclostome-specific globins in phylogenetic analyses was it possible to infer that functionally analogous “Hbs” and “Mbss” evolved independently from different precursor globins that were present in the vertebrate ancestor (Hoffmann, Opazo, et al. 2010; Schwarze et al. 2014). Similarly, because of extensive lineage-specific gene losses, the single copy \( GbX \) genes of frog, most teleost fish, testudines, and squamates were thought to be 1:1 orthologs, and the possession of two copies in coelacanth was attributed to lineage-specific duplication (Schwarze and Burmester 2013). However, the inclusion of globin sequences from elephant shark, coelacanth, and lampreys in our analyses suggests a more complex historical scenario. The phylogenetic reconstructions and patterns of conserved synteny suggest that whole-genome duplication and/or smaller scale duplication events produced a set of four unlinked \( GbX \) paralogs in the common ancestor of vertebrates. During the ensuing approximately 500 My of vertebrate evolution, lineage-specific gene losses led to a gradual winnowing of \( GbX \) diversity, with different lineages retaining different \( GbX \) paralogs.

**Materials and Methods**

**DNA Sequence Data**

We characterized the complete globin gene repertoire of the elephant shark (\( C. milii \)) by assembling RNA-seq reads from ten tissues (brain, gills, heart, intestine, kidney, liver, muscle, ovary, spleen, and testis; Venkatesh et al. 2014). We then used the identified transcripts to search against databases of known vertebrate globins. We also used the elephant shark transcripts to search against the nonredundant (NR) protein database available at NCBI using BLASTX to confirm their identities as authentic globins. We then mapped all confirmed globin transcripts to the elephant shark genome assembly (Assembly Callorhinchus_miliii-6.1.3, available at http://esharkgenome.imcb.a-star.edu.sg/) using BLASTN, where full-length coding sequences were predicted by a combination of similarity of the transcripts to the genome sequence (BLASTN) and similarity of the genome sequence to known globin protein sequences from the NCBI database (BLASTX). The predicted protein sequences were further refined by manual inspection of exon–intron boundaries. One of the transcripts, which is full-length and encodes \( GbX2 \), had no match in the genome assembly. This gene is presumably present in the gaps of the genome assembly. Our searches in the NCBI database identified an additional globin gene, \( \alpha \)-globin 2, which was previously identified in the full-length cDNA data set (accession number AFK11062; Tan et al. 2012) but is not present in the current genome assembly.

Our comprehensive analyses also included the complete globin gene repertoire of 19 deuterostome taxa, including 17 chordates (14 vertebrates, two urochordates, and one cephalochordate), one hemichordate and one echinoderm. The vertebrate species included three mammals (human, \( Homo sapiens \); gray short-tailed opossum, \( Monodelphis domestica \); and platypus, \( O. anatinus \)), three birds (chicken, \( Gallus gallus \); turkey, \( Meleagris gallopavo \); and zebra finch, \( Taeniopygia guttata \)), one squamate reptile (green anole lizard, \( A. carolinensis \)), one amphibian (western clawed frog, \( Xenopus tropicalis \)), one lobe-finned fish (West Indian Ocean coelacanth, \( L. chalumnae \)), and five teleost fish (fugu, \( Takifugu rubripes \); medaka, \( Oryzias latipes \); green-spotted pufferfish, \( Tetraodon nigroviridis \); three-spined stickleback, \( Gasterosteus aculeatus \); and zebrafish, \( Danio rerio \)). The nonvertebrate species included two sea squirts (\( Ciona intestinalis \) and \( C. savignyi \), Urochordata), amphioxus (\( Branchiostoma floridae \), Cephalochordata), the acorn worm (\( Saccoglossus kowalevskii \), Hemichordata), and the purple sea urchin (\( Strongylocentrotus purpuratus \), Echinodermata). Because the sea urchin possesses a globin gene that encodes 16 putative globin domains (Bailly and Vinogradov 2008), we considered each of the domains as separately alignable globin sequences in our analyses. In addition to retrieving the complete globin gene repertoires of the species listed above, we also included individual globin sequences from several additional taxa. For echinoderms, we included sequences from two subunits of the coelomic red blood cell Hbs of the sea cucumber, \( Caudina arenicola \). For cartilaginous fish, we obtained \( \alpha \)- and \( \beta \)-Hb sequences from the red stingray (\( Dasyatis akajei \)) and gummy houndshark (\( Mustelus antarcticus \)), as well as \( Mb \) sequences from the gummy houndshark and the Port Jackson shark (\( Heterodontus portusjacksoni \)). For birds, we included the \( GbE \) sequence of the mallard duck (\( Anas platyrhynchos \)). For cyclostomes, we included the \( GbX1 \), \( GbX2 \), and \( Cygb \) genes of the Arctic and sea lampreys (\( Lethenteron japonicum \) and \( Petromyzon marinus \), respectively), the \( Cygb \) of the European brook lamprey (\( Lamproptera planeri \)), and the \( Hb \) of the arctic lamprey and Atlantic hagfish (\( Myxine glutinosa \); Schwarze et al. 2014). In order to include representatives from each of the main deuterostome lineages, we also included two globin sequences from the acel \( Symphys gistifera roscoffensis \),
three globin sequences from the nemertodermatid Nemertodera westbladi, plus one sequence from the xenoturbellid, Xenoturbella bocki. Nine plant globin sequences were included as outgroup. The full list of sequences and the corresponding accession numbers are presented in supplementary table S1, Supplementary Material online. The order of the helical domains of the globin domain of Androglobin has been rearranged relative to all other globins in this study (Hoogewijs et al. 2012). Because of the corresponding alignment challenges, this sequence was not included in the phylogenetic analyses.

For the analyses that focused on resolving orthologous relationships among vertebrate GbX paralogs, we also included additional GbX genes from other vertebrate groups to increase our taxonomic coverage. We added GbX paralogs from an holostean fish (spotted gar, Lepisosteus oculatus), three testudines (green sea turtle, Chelonia mydas; painted turtle, Chrysemys picta; and Chinese softshell turtle, Pelodiscus sinensis), two squamates (Burmese python, Python morulus; and anole lizard, A. carolinensis), and additional teleost fish (Atlantic cod, Gadus morhua; tilapia, Oreochromis niloticus; goldfish, Carassius auratus; platyfish, Xiphophorus maculatus; and cave fish, Astyanax mexicanus). For the sake of providing comprehensive coverage of amniotes, we also surveyed genome assemblies and sequence databases to identify globin genes in crocodilians (Green et al. 2014) and birds (Grispo et al. 2012; Zhang et al. 2014; Opazo et al. 2015) using the bioinformatic strategies outlined above.

Phylogenetic Analyses

We estimated phylogenetic relationships among globin genes of deuterostomes using a Bayesian approach, as implemented in the program MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003) under a mixed model of amino acid substitution. Sequence alignment was carried out using the G-INS-i strategy from MAFFT v6 (Kato et al. 2009). Two simultaneous independent runs were performed for 10 × 10⁶ iterations of a Markov Chain Monte Carlo algorithm, with six simultaneous chains sampling trees every 1,000 generations. Support for the nodes and parameter estimates were derived from a majority rule consensus of the last 5,000 trees sampled after convergence. The average standard deviation of split frequencies remained 0.01 after the burn-in threshold.

Gene Expression Analyses

Paired-end RNA-seq data for ten adult elephant shark tissues (brain, gills, heart, intestine, kidney, liver, muscle, ovary, spleen, and testis) were obtained from SRA accession SRA054255 (Venkatesh et al. 2014). Sequence reads were aligned against the elephant shark transcriptome (NR set of ~18,900 transcripts; Venkatesh et al. 2014) using Bowtie v. 0.12.9 (Langmead et al. 2009). Transcript abundances were estimated using RSEM v1.2.3 (Li and Dewey 2011). Default settings were used for both programs. Transcript abundances are measured in transcripts per million.

Supplementary Material

Supplementary table S1 is available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was funded by grants from the National Institutes of Health/National Heart, Lung, and Blood Institute to J.F.S. (R01 HL087216 and HL087216-S1), grants from the National Science Foundation to J.F.S. (IOS-0949931), and F.G.H. (EPS-0903787), a grant from Fondo Nacional de Desarrollo Cientifico y Tecnologico to J.C.O. (FONDECYT 112032), a grant be the Detusche Forschungsgemeinschaft to T.B. (BU 956/18-1), and a grant from the Biomedical Research Council of A*STAR to B.V. The authors thank the associate editor and two anonymous reviewers for helpful suggestions.

References


