Differential Loss and Retention of Cytoglobin, Myoglobin, and Globin-E during the Radiation of Vertebrates

Federico G. Hoffmann1,2, Juan C. Opazo3, and Jay F. Storz*,1

1School of Biological Sciences, University of Nebraska, Lincoln
2Department of Biochemistry and Molecular Biology, Mississippi State University
3Instituto de Ecología y Evolución, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile

*Corresponding author: E-mail: jstorz2@unl.edu.

Accepted: 30 May 2011

Abstract

If rates of postduplication gene retention are positively correlated with levels of functional constraint, then gene duplicates that have been retained in a restricted number of taxonomic lineages would be expected to exhibit relatively low levels of sequence conservation. Paradoxical patterns are presented by gene duplicates that have been retained in a small number of taxa but which are nonetheless subject to strong purifying selection relative to paralogous members of the same multigene family. This pattern suggests that such genes may have been co-opted for novel, lineage-specific functions. One possible example involves the enigmatic globin-E gene \( \text{GbE} \), which appears to be exclusively restricted to birds. Available data indicate that this gene is expressed exclusively in the avian eye, but its physiological function remains a mystery. In contrast to the highly restricted phyletic distribution of \( \text{GbE} \), the overwhelming majority of jawed vertebrates (gnathostomes) possess copies of the related cytoglobin \( \text{Cygb} \) and myoglobin \( \text{Mb} \) genes. The purpose of the present study was 1) to assess the phyletic distribution of \( \text{GbE} \), \( \text{Cygb} \), and \( \text{Mb} \) genes among vertebrates, 2) to elucidate the duplicative origins and evolutionary histories of these three genes, and 3) to evaluate the relative levels of functional constraint of these genes based on comparative sequence analysis. To accomplish these objectives, we conducted a combined phylogenetic and comparative genomic analysis involving taxa that represent each of the major lineages of gnathostome vertebrates. Results of synteny comparisons and phylogenetic topology tests revealed that \( \text{GbE} \) is clearly not the product of a recent, bird-specific duplication event. Instead, \( \text{GbE} \) originated via duplication of a proto-\( \text{Mb} \) gene in the stem lineage of gnathostomes. Unlike the \( \text{Mb} \) gene, which has been retained in all major gnathostome lineages other than amphibians, the \( \text{GbE} \) gene has been retained only in the lineage leading to modern birds and has been independently lost in at least four major lineages: teleost fish, amphibians, mammals, and nonavian reptiles. Despite the restricted phyletic distribution of this gene, our results indicate that \( \text{GbE} \) is one of the most highly conserved globins in the avian genome.

Key words: bird genome, gene duplication, gene family evolution, genome evolution, globins.

Introduction

Duplicated genes are most likely to be retained in the genome when the two descendant copies partition ancestral subfunctions of the single-copy progenitor or when one or both copies evolve new specializations of function (Lynch and Force 2000; Lynch et al. 2001). Because different gene functions are not all equally essential in the economy of the organism, it is not surprising that paralogous members of the same multigene family often exhibit substantial variation in levels of functional constraint. Comparative analysis of genomic sequence data can be used to evaluate variation in functional constraint among members of the same gene family by measuring rates of gene loss and retention among different taxonomic lineages and by measuring rates of amino acid substitution in orthologous comparisons. Highly essential genes are expected to be retained in phylogenetically diverse taxa and are also expected to exhibit especially slow rates of amino acid substitution relative to paralogous members of the same gene family. At the other end of the spectrum, dispensable genes may be inactivated or deleted in multiple taxa, and such genes are also expected to exhibit elevated rates of amino acid substitution due to a relative
relaxation of purifying selection. Seemingly paradoxical patterns are presented by genes that have been retained in a small number of taxa but which nonetheless exhibit high levels of sequence conservation relative to other paralogous gene duplicates. This pattern suggests that such genes may have been co-opted for a novel, lineage-specific function.

Comparative genomic studies of the globin gene superfamily have revealed a number of striking discrepancies in the phyletic distribution of paralogous gene lineages. For example, all jawed vertebrates (gnathostomes) examined to date possess copies of neuroglobin (Ngb) and cytoglobin (Cygb; Burmester et al. 2000, 2002, 2004; Awenius et al. 2001; 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; Hoffmann, Opazo, and Storz 2010). These genes are also characterized by the highest levels of sequence conservation among all members of the gene family (Burmester et al. 2004). In gnathostomes, hemoglobin (Hb) and myoglobin (Mb) also represent examples of seemingly indispensable globin proteins that play critical roles in the maintenance of cellular oxygen supply in support of aerobic metabolism. Antarctic ice fish in the family Channichthyidae represent one notable exception to what appears to be an otherwise universal rule, as these fish do not express Hb and many species in the family do not express Mb either (Hemmingsen 1991; Sidell and O’Brien 2006). Mb has also been secondarily lost in amphibians (Maeda and Fitch 1982; Xi et al. 2007).

In contrast to the highly conserved Ngb, Cygb, Hb, and Mb proteins that have been retained in all or nearly all vertebrate lineages, there are also a number of paralogous globin proteins that have much more restricted phyletic distributions. The globin X gene is the product of a duplicative event, which predates the divergence between proto- and deuterostomes, and among vertebrates, it has only been found in the genomes of teleost fish and Xenopus (Roesner et al. 2005). The globin Y (GbY) gene appears to be restricted to gnathostome vertebrates and has thus far only been found in the genomes of Xenopus, anole lizard, and platypus (Fuchs et al. 2006; Patel et al. 2008; Hoffmann, Opazo, and Storz 2010). The GbY gene and the proto-Hb gene represent paralogous products of an ancient tandem gene duplication in the stem lineage of vertebrates (Storz et al. 2011). The ancestral linkage arrangement of these two genes is still retained in the genomes of Xenopus, anole lizard, and platypus, as GbY is located at the 3’ end of the α-globin gene cluster in each of these taxa (Fuchs et al. 2006; Patel et al. 2008; Hoffmann, Storz, et al. 2010). Finally, the globin-E (GbE) gene has thus far only been found in the genome of birds (Kugelstadt et al. 2004; Hoffmann, Opazo, and Storz 2010). Available data suggest that GbE is expressed exclusively in the eye, and although structural considerations suggest that the protein could have an oxygen-binding respiratory function, its physiological role remains a mystery (Kugelstadt et al. 2004). The level of functional constraint of this gene has not been assessed previously because it has been characterized only in chicken (Kugelstadt et al. 2004). Thus, it has not been possible to make the same kinds of taxonomic cross-comparisons that have been made with other globin genes that have a broader phyletic distribution. However, with the availability of genomic sequence data from other bird species besides chicken (turkey, mallard duck, and zebra finch), it is now possible to compare levels of among-species sequence conservation across the full complement of globin genes in the avian genome.

There have been conflicting accounts concerning the phylogenetic placement of GbE relative to other vertebrate-specific globin genes. Based on results of a phylogenetic analysis, which indicated that GbE is most closely related to Cygb, Kugelstadt et al. (2004) suggested two alternative hypotheses: 1) GbE and Cygb trace their duplicative origins to the common ancestor of tetrapods and teleost fish, in which case the GbE gene would have been secondarily lost from all descendant lineages other than birds or 2) GbE is the product of a bird-specific duplication event. The phylogenetic reconstructions reported by Hoffmann, Opazo, and Storz (2010) are not consistent with either of these hypotheses. Instead, these more recent results indicate that GbE originated via duplication of a proto-Mb gene in the stem lineage of gnathostomes. The latter scenario also posits that the GbE gene was lost secondarily from all descendant lineages other than birds.

The purpose of the present study was 1) to assess the phyletic distribution of the Cygb, Mb, and GbE genes among vertebrates, 2) to elucidate the duplicative origins and evolutionary histories of these three genes, and 3) to evaluate the relative levels of functional constraint of these genes based on comparative sequence analysis. To accomplish these objectives, we conducted a combined phylogenetic and comparative genomic analysis involving taxa that represent each of the major lineages of gnathostome vertebrates. Results of synteny comparisons and phylogenetic topology tests revealed that the Mb and GbE genes represent products of a tandem gene duplication in the stem lineage of gnathostomes and that the proto-Mb/GbE gene and the Cygb gene represent paralogous products of a whole-genome duplication event in the stem lineage of vertebrates. Thus, GbE is clearly not the product of a recent, bird-specific duplication event. Rather, available data suggest that this ancient gene has been retained only in birds (and possibly in other archosaur lineages) and has been independently lost in all other gnathostome lineages. Despite the restricted phyletic distribution of this gene, our results also indicate that, in birds, GbE is one of the most highly conserved globin proteins and is characterized by a level of functional constraint similar to that of both Mb and Hb.
Materials and Methods

Data Collection

We had previously identified the full complement of globin genes in the genomes of 12 vertebrate species (Hoffmann, Opazo, and Storz 2010): five teleost fish (fugu, Takifugu rubripes; medaka, Orzias latipes; pufferfish, Tetraodon nigroviridis; three-spined stickleback, Gasterosteus aculeatus; and zebrafish, Danio rerio), one amphibian (western clawed frog, Xenopus tropicalis), one squamate reptile (green anole lizard, Anolis carolinensis), two birds (chicken, Gallus gallus and zebra finch, Taeniopygia guttata), and three mammals (human, Homo sapiens; gray short-tailed opossum, Monodelphis domestica; and platypus, Ornithorhynchus anatinus). For the present study, we added sequences corresponding to the Cygb, GBE, and Mb genes from the prerelease version of the mallard duck (Anas platyrhynchos) and turkey (Meleagris gallopavo) genomes available from the Ensembl-Pre database. For the purpose of investigating patterns of conserved synteny in mammals, we also annotated the genes found upstream and downstream of the Mb gene in cow (Bos taurus), dog (Canis familiaris), rabbit (Oryctolagus cuniculus), guinea pig (Cavia porcellus), mouse (Mus musculus), and rat (Rattus norvegicus). The additional globin sequences were annotated by comparing known coding sequences to genomic contigs using basic alignment search tool (Blast) (Altschul et al. 1990).

Phylogenetic Analyses

To verify the phylogenetic position of the additional avian globin genes included in this study, we added sequences from the Cygb, GBE, and Mb genes of turkey and mallard duck to an alignment of vertebrate globins that included representative amino acid sequences of Cygb (including cyclostome Hbs; Hoffmann, Opazo, and Storz 2010), GBE, GbY, α- and β-chain Hbs, and Mb. As outgroup sequences, we included three globin sequences from amphioxus (Branchiostoma floridae, Cephalochordata), which are sister to the complete set of vertebrate-specific globins (Ebner et al. 2010; Storz et al. 2011), plus Ngb sequences from eight vertebrate taxa (supplementary table 1, Supplementary Material online). We aligned amino acid sequences using the L-INS-i strategy from Maftf version 6.8 (Kato in and Toh 2008; Kato in et al. 2009), and we estimated phylogenetic relationships using maximum likelihood and Bayesian approaches (the resulting alignment is provided as supplementary fig. 1, Supplementary Material online). Maximum likelihood searches were carried out using Treefinder version October 2008 (Jobb et al. 2004) using the LG + I + G model (Le and Gascuel 2008), which was identified as the best fitting model of amino acid substitution with the “propose model” subroutine from Treefinder. Support for the nodes was evaluated with 1,000 bootstrap pseudoreplicates. Bayesian analyses were conducted in MrBayes v.3.1.2 (Ronquist and Huelsenbeck 2003) running six simultaneous chains for 10,000,000 generations, sampling every 1,000 generations, under a mixed model of amino acid substitution and using default priors. Support for the nodes and parameter estimates were derived from a majority rule consensus of the last 2,500 trees. The average standard deviation of split frequencies remained less than 0.01 after the burn-in threshold. MrBayes analyses were run using the Cipres portal (Miller et al. 2009).

Synteny Analyses

To assess synteny conservation among chromosomal regions that harbor the Cygb, Mb, and GBE genes, we annotated the genes found upstream and downstream of the paralogous globin genes in each of the 14 vertebrate genome sequences examined. Initial ortholog predictions were derived from EnsemblComparer (Vilella et al. 2009). We also used the program Genscan (Burge and Karlin 1997) to identify additional unannotated genes lying upstream and downstream of the annotated globin genes. The unannotated genes were compared with the nonredundant protein database using Blast (Altschul et al. 1990). Finally, partial sequences for genes of interest (representing either pseudogenes or incomplete coverage) were identified and annotated with Blast. Additional synteny comparisons were conducted using Genomicus (Muffato et al. 2010).

Hypothesis Testing

To test alternative phylogenetic hypotheses about pathways of globin gene family evolution, we used parametric bootstrapping tests (Swofford et al. 1996; Goldman et al. 2000) as well as the approximately unbiased (AU) test proposed by Shimodaira (2002). Results of Kugelstadt et al. (2004) suggested that the GBE and Cygb genes are either products of a relatively ancient duplication event in the gnathostome lineage that occurred prior to the split between tetrapods and teleost fish or they represent the products of a much more recent bird-specific duplication. By contrast, the phylogenetic reconstruction reported by Hoffmann, Opazo, and Storz (2010) suggests that GBE originated via duplication of a proto-Mb gene in the stem lineage of gnathostomes. To test these alternative hypotheses about the duplicative origins of the Cygb, Mb, and GBE genes, we compared the likelihood scores associated with the following phylogenies: 1) a constrained tree in which GBE was sister to the gnathostome Cygb clade, 2) a constrained tree in which GBE was sister to the avian Cygb clade, 3) a constrained tree in which GBE was sister to the avian Mb clade, and 4) an unconstrained tree in which GBE was sister to the clade that included the Mb genes of gnathostomes. All tests were carried out in Treefinder version October 2008 (Jobb et al. 2004).
Molecular Evolution Analyses

To assess relative levels of functional constraint among the different globin paralogs and to evaluate the effects of substitution rate heterogeneity on the phylogenetic reconstructions, we compared rates of amino acid substitution among avian orthologs of Cygb, GbE, and Mb using Tajima’s (1993) relative test, as implemented in MEGA version 5 (Kumar et al. 2008).

To make more refined inferences about levels of sequence conservation and functional constraint among avian globin genes, we quantified sequence conservation among all members of the globin gene superfamily shared by boreoeutherian mammals: Cygb, Mb, GbE, Ngb, αF-globin, α2-globin, αA-globin, p-globin, ε-globin, βA-globin, and βE-globin (supplementary table 2, Supplementary Material online). Of the four bird species included in this analysis (all of which are representatives of the supraordinal clade Neognathae), the deepest split is between Neoeoa (represented by zebra finch [order Passeriformes]) and Galloanserae (represented by chicken and turkey [order Galliformes] and mallard duck [order Anseriformes]). The Galloanserae–Neoeoa divergence is thought to have occurred ~100–130 Ma in the Cretaceous (Ericson et al. 2006; Brown et al. 2008). For comparative purposes, we conducted a similar analysis of sequence conservation among mammalian globin genes. To do this, we chose a phylogenetic assemblage of mammalian taxa that are thought to have diversified over roughly the same timescale as the four bird species that we examined. Specifically, we analyzed the evolution of globin genes in representative species from the supraordinal clade Boreoeuthera, which contains Lausariatheria (the superorder that includes bats, carnivores, cetaceans, artiodactyls, and perissodactyls) and Euarchontoglires (the superorder that includes primates, rodents, and lagomorphs). This clade provides a good basis for comparison with the neognath birds, as the split between Lausariatheria and Euarchontoglires is thought to have occurred roughly 100 Ma in the mid-Cretaceous (Bininda-Emonds et al. 2007; Murphy et al. 2007). We compared relative levels of sequence conservation among all members of the globin gene superfamily shared by boreoeutherian mammals: Cygb, Mb, Ngb, αF-globin (=ζ-globin), α2-globin, αA-globin, ε-globin, and β-globin. This analysis included the full complement of globin genes found in the genomes of mega- and microchiropteran bats, cow, dog, guinea pig, mouse, pika, rabbit, human, macaque, and marmoset (supplementary table 2, Supplementary Material online).

Nucleotide sequences were retrieved from the GenBank or Ensembl databases, and conceptual translations of the nucleotide sequences were then aligned using the L-INS-i strategy in Mafft version 6.8 (Katoh and Toh 2008; Katoh et al. 2009). The amino acid alignments were then back translated to using the Mobyle web server, and the resulting nucleotide alignments were used in all subsequent analyses of molecular evolution. The alignment of globin sequences for neognath birds is available as supplementary figure 2 (Supplementary Material online) and the alignment of globin sequences for boreoeutherian mammals is available as supplementary figure 3 (Supplementary Material online). In separate analyses of the avian and mammalian globin genes, we calculated the average pairwise amino acid differences using MEGA version 5 (Kumar et al. 2008), and we compared estimates of ω (=dN/dS), the ratio of the rate of nonsynonymous substitution per nonsynonymous site (dN) to the rate of synonymous substitution per synonymous site (dS) using the codeml program from the PAML package version 4.4c (Yang 2007).

Results and Discussion

Phyletic Distribution

To unravel the evolutionary history of the Cygb, Mb, and GbE gene lineages, we first assessed the phyletic distribution of the three paralogs among representatives of each of the major vertebrate classes for which complete genomic sequence was available: teleost fish, amphibians, mammals, birds, and nonavian reptiles. All tetrapod species examined in this study possess a single copy of the Cygb gene, and all teleost fish possess either two duplicated copies or, in the case of zebra fish, three copies (fig. 1). Analysis of conserved synteny revealed that the two Cygb paralogs possessed by all teleost fish, Cygb-1 and Cygb-2, are embedded in chromosomal segments that share multiple paralogous gene duplicates in the upstream and downstream flanking regions. These homologous chromosomal segments (or “paralogons”) appear to represent products of a large-scale segmental duplication or possibly the whole-genome duplication event that occurred in the stem lineage of teleost fish (Meyer and Schartl 1999; Taylor et al. 2001, 2003). In fact, we were able to confirm that the Cygb-1 and Cygb-2 paralogs derive from the fish-specific whole-genome duplication because both Cygb-defined paralogons descend from chromosome “e” in the reconstructed (unduplicated) protokaryotype of the teleost common ancestor (Nakatani et al. 2007). This supports the conjecture of Fuchs et al. (2005) that the Cygb-1 and Cygb-2 paralogs are products of whole-genome duplication.

Almost all species examined possess a single copy of the Mb gene. Consistent with previous reports (e.g., Fuchs et al. 2006; Xi et al. 2007), we found no trace of an Mb gene in the Xenopus genome. The absence of Mb in amphibians in the order Anura, which includes frogs and toads, and the order Caudata, which includes salamanders and newts, was previously noted in a survey of protein sequence data in the American bullfrog (Rana catesbeiana; Maeda and Fitch 1982), a survey of genomic sequence data in the western long-clawed frog (X. tropicalis; Fuchs et al. 2006), and...
southern hybridization and reverse transcriptase-polymerase chain reaction experiments in Japanese fire-bellied newts (Cynops pyrrhogaster; Xi et al. 2007). Surprisingly, we found suggestive evidence that the $\text{Mb}$ gene may have been wholly or partially deleted in two other vertebrate species, the three-spined stickleback and the gray short-tailed opossum. We found no trace of the $\text{Mb}$ gene in the genome assembly of the three-spined stickleback, and we found only a partial coding sequence (exon 1) in the genome assembly of the opossum. We also searched the available stickleback and opossum expressed sequence tag (EST) databases and trace archives and found no $\text{Mb}$ sequence matches.

Moreover, this alignment reveals that functionally important residues, such as the CD1 phenylalanine, the proximal (F8) histidine, and the distal (E7) histidine are conserved in all GbE sequences (fig. 2). Bioinformatic searches for previously unidentified orthologs of the avian GbE gene in vertebrate sequence databases failed to detect any significant matches in the genomes of nonavian taxa. These searches were based on both nucleotide and amino acid sequences and included EST and trace records from cyclostomes, cartilaginous fish, teleost fish, and tetrapods. It remains to be seen whether GbE is present in the genomes of other extant archosauromorphs (such as crocodilians or tuatara) or whether it is exclusively restricted to modern birds.

**Phylogenetic Analyses**

Results of maximum likelihood and Bayesian phylogenetic searches yielded a tree that is consistent with previous phylogenetic reconstructions of vertebrate-specific globin genes (Hoffmann, Opazo, and Storz 2010; Storz et al. 2011). The different vertebrate-specific globin paralogs included in this study formed four distinct and strongly supported clades: 1) $\text{Cygb} + \text{cyclostome Hb}$, 2) $\text{Mb} + \text{GbE}$, 3) $\text{GbY}$, and 4) the $\alpha$- and $\beta$-chain Hbs of gnathostomes (fig. 3). Support for the monophyly of each of these four clades was strong in Bayesian analyses but moderate in the case of the maximum likelihood bootstrap analysis. In all analyses, we confirmed a sister relationship between the ($\text{Cygb} + \text{GbE}$) complex and the the $\alpha$- and $\beta$-globins $\text{Hb}$ and $\text{GbE}$, respectively, but the deeper relationships among these clades are only weakly supported.
cyclosteone Hb) and the (Mb + GbE) clades, whereas the deepest split among the four lineages of vertebrate-specific globins corresponds to the split between the \( \alpha \) and \( \beta \)-Hbs of gnathostomes and the common ancestor of the remaining three clades of paralogous globins. Again, only Bayesian analyses provided relatively high levels of support for these different clades. As shown in figure 3, the avian Cygb, Mb, and GbE sequences each form monophyletic groups, and in the case of Cygb and Mb, they are nested within the expected set of orthologous genes from other tetrapods.

The phylogenetic reconstruction shown in figure 3 is not consistent with that reported by Kugelstadt et al. (2004) (see also Burmester et al. 2004; Burmester and Hankeln 2009), probably due to the more dense sampling of vertebrate globin diversity made possible by the increased availability of genomic data. This previous study suggested that GbE was more closely related to Cytg than to other vertebrate-specific globins. Accordingly, Kugelstadt et al. (2004) suggested two alternative hypotheses: 1) GbE and Cygb genes are products of an ancient gene duplication that occurred prior to the split between teleost fish and tetrapods or 2) the GbE gene arose via duplication of the Cygb gene in the common ancestor of birds. With respect to the latter hypothesis, Kugelstadt et al. (2004) suggested that the observed placement of the avian GbE gene outside the clade of tetrapod Cygb sequences might be an artifactual result that is attributable to an accelerated rate of amino acid substitution.

We conducted topology tests to evaluate alternative phylogenetic hypotheses regarding the duplicative origins of GbE (fig. 4). Topology tests implemented with constrained trees strongly rejected the bird-specific duplication hypotheses in favor of the tree that placed GbE sister to the Mb clade (table 1). Moreover, the parametric bootstrap tests strongly rejected the (GbE, Cygb, Mb) topology in favor of the (GbE, Mb, Cygb) topology. The AU test also favored this latter topology although the difference in likelihood scores was only marginally significant (\( P < 0.08 \)). Because our sample of Mb sequences included cartilaginous fish, teleost fish, and tetrapods, our results indicate that the tandem gene duplication that gave rise to the GbE and Mb genes occurred prior to the radiation of gnathostome vertebrates. These results also indicate that orthologs of the GbE gene have been lost independently in all gnathostome lineages other than the lineage leading to modern birds. The differential retention of Mb and GbE mirrors a similar pattern of gene turnover in the \( \beta \) globin gene clusters of tetrapod vertebrates (Hoffmann et al. 2008a,b; Opazo et al. 2008a,b).

**Synteny Analyses**

Because of the additional round of whole-genome duplication in teleost fish and the ensuing differential loss of paralogous gene duplicates, we initially conducted separate analyses on genomic data from tetrapods and teleost fish. In the case of the Cygb gene region, the analysis revealed a high degree of conserved synteny within tetrapods and within teleost fish (with the exception of zebra fish; fig. 5). Furthermore, the genes found downstream of the Cygb-1 paralog in medaka, pufferfish, and stickleback are nearly identical to the genes found downstream of the Cygb gene in all tetrapods, although the linear order has been inverted (fig. 5). Similarly, the SPHK1 and PRSAP1 genes, which bracket the Cygb-2 paralog in teleost fish, are also found in close proximity to the single-copy Cygb gene of tetrapods.

The pattern of conserved synteny spanning the GbE and Mb genes is more complex than that observed for Cygb. The genes found adjacent to Mb are well conserved within tetrapods (fig. 6a) and within teleost fish (fig. 6b). Among teleost fish, the zebra fish genome is again characterized by the lowest level of conserved synteny (fig. 6b). Synteny conservation...
FIG. 3.—Maximum likelihood phylogram describing phylogenetic relationships among vertebrate-specific globins from representative taxa obtained under the LG + I + Γ model of amino acid substitution. As outgroup sequences, we included three globin sequences from amphioxus that are sister to the complete set of vertebrate-specific globins, plus Ngb sequences from eight representative vertebrate taxa. Numbers above the nodes correspond to maximum likelihood bootstrap support values, and those below the nodes correspond to Bayesian posterior probabilities.
is stronger in the case of amniotes, as represented by human, opossum, platypus, anole lizard, chicken, and zebra finch. If we ignore variation in the number of tandemly linked APO paralogs, the set of genes found upstream and downstream of the \textit{Mb} gene are almost identical in this set of taxa (fig. 6A).

Despite the fact that the stickleback and opossum appear to be missing orthologs of the \textit{Mb} gene, synteny is conserved for the genes flanking the ancestral \textit{Mb} locus in both species (fig. 6A and B). Contrary to the highly conserved pattern shown in figure 6, Xi et al. (2007) reported a low degree of conserved synteny in the \textit{Mb} gene region of vertebrates. The discrepancy between our results and those of Xi et al. (2007) is attributable to the fact that these authors included mouse as a representative mammal in their synteny comparisons. Due to a chromosomal translocation that occurred in the common ancestor of rat and mouse, the \textit{RASD2}, \textit{MCM5}, \textit{HMOX1}, \textit{TOM1}, and \textit{ISX} genes of these two rodent species are not found immediately adjacent to the \textit{Mb} gene as they are in most other amniote taxa (supplementary fig. 4, Supplementary Material online).

In each of the examined avian genomes, the \textit{GbE} gene is located between the \textit{Luc7L2} and the \textit{POLDIP3} genes (fig. 6A). In the anole lizard genome assembly, the \textit{Luc7L2} and \textit{POLDIP3} genes are located 5 kb apart on the same scaffold, but there is no trace of \textit{GbE} in the intervening chromosomal region. In placental mammals, the \textit{Luc7L2} and \textit{POLDIP3} genes are located on separate chromosomes, whereas in opossum, they are separated by over 17 Mb on chromosome 8. Although the avian \textit{Mb} and \textit{GbE} genes are separated by physical distances of over 3 Mb in each of the bird species examined here, \textit{Mb}- and \textit{GbE}-associated genes that are separated by vast physical distances on chromosome 1 of galliform birds (and chromosome 1A of zebra finch) are found in close physical proximity in the homologous chromosomal segments of teleost fish and amphibians (fig. 6A and B; supplementary fig. 5, Supplementary Material online). This indicates that the \textit{Mb} and \textit{GbE} genes were closely linked in the genome of the gnathostome common ancestor.

Patterns of Molecular Evolution

In chicken and zebra finch, we did not detect significant differences in relative rates of amino acid substitution in the comparison between \textit{Mb} and \textit{GbE}. In turkey and mallard duck, by contrast, test results indicated that the \textit{GbE} gene is evolving at a significantly slower rate than the \textit{Mb} gene (table 2). The discrepancy in substitution rates between the
Mb and the GbE paralogs is also reflected in the reconstructed phylogeny, as the branch leading to the GbE clade is clearly shorter than the branch leading to the Mb clade (fig. 3). All gnathostome vertebrates examined to date possess one or more copies of the Cygb gene, and the overwhelming majority possess a single Mb gene. By contrast, the GbE gene is restricted to birds and possibly other archosaurs. If rates of gene retention are positively correlated with the degree of functional constraint, then we would expect to observe the following rank order of sequence conservation: Cygb

![Figure 5](image_url)

**Fig. 5.**—Patterns of conserved synten in the chromosomal region that harbors the Cygb gene in gnathostome vertebrates. Horizontal lines denote orthologous relationships.

![Figure 6](image_url)

**Fig. 6.**—Patterns of conserved synten in the chromosomal region that harbors the GbE and Mb genes in gnathostome vertebrates. (A) Synten comparisons of the GbE and Mb gene regions among representative amniote taxa. (B) Synten comparisons of the Mb gene region among representative teleost fish. Horizontal lines denote orthologous relationships.
Evolution of Cytoglobin, Myoglobin, and Globin-E

Table 2
Results from Tajima’s (1993) Relative Rate Tests of Amino Acid (aa) Substitution

<table>
<thead>
<tr>
<th></th>
<th>Unique aa Changes in Mb</th>
<th>Unique aa Changes in GbE</th>
<th>Unique aa Changes in Cygb</th>
<th>Identical Sites</th>
<th>Divergent Sites*</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chicken</td>
<td>27</td>
<td>15</td>
<td>21</td>
<td>27</td>
<td>57</td>
<td>NS</td>
</tr>
<tr>
<td>Duck</td>
<td>27</td>
<td>13</td>
<td>20</td>
<td>27</td>
<td>60</td>
<td>P &lt; 0.03</td>
</tr>
<tr>
<td>Turkey</td>
<td>28</td>
<td>15</td>
<td>20</td>
<td>27</td>
<td>57</td>
<td>P &lt; 0.05</td>
</tr>
<tr>
<td>Zebra finch</td>
<td>24</td>
<td>16</td>
<td>19</td>
<td>26</td>
<td>62</td>
<td>NS</td>
</tr>
</tbody>
</table>

Note.—NS, not significant.

“Sites that were different in all three sequences.

> Mb > GbE. Under the model in which ω (dS/dD) was allowed to vary among the three paralogous gene lineages, Cygb was characterized by the lowest ω value, whereas GbE and Mb were very similar (table 3). However, forcing all three paralogs Cygb, GbE, and Mb to have a single ω value did not result in a significant change in the likelihood score (supplementary fig. 6, Supplementary Material online). We therefore conclude that these three genes have been evolving under similar levels of purifying selection. The comparative analysis of neognathous birds and boreoeutherian mammals revealed similar patterns of sequence conservation, and the rank order of ω values for the subset of avian and mammalian globin genes that are 1:1 orthologs (Ngb, Cygb, Mb, aG, aD, and aE) was similar in both groups (table 3).

Table 3
Evolutionary Comparisons for Members of the Globin Gene Superfamily in the Genomes of Neo­gnathe Birds (Top) and Boreoeutherian Mammals (Bottom)

<table>
<thead>
<tr>
<th>Gene</th>
<th>ω*</th>
<th>P distance (aa)b</th>
<th>ω*</th>
<th>P distance (aa)b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neognathae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cygb</td>
<td>0.05</td>
<td>6.2</td>
<td>8.4</td>
<td>0.03</td>
</tr>
<tr>
<td>GbE</td>
<td>0.08</td>
<td>6.1</td>
<td>7.3</td>
<td>0.11</td>
</tr>
<tr>
<td>Ngb</td>
<td>0.12</td>
<td>8.5</td>
<td>11.3</td>
<td>0.15</td>
</tr>
<tr>
<td>Mb</td>
<td>0.19</td>
<td>19.1</td>
<td>25.5</td>
<td>0.15</td>
</tr>
<tr>
<td>Ngb</td>
<td>0.15</td>
<td>5.9</td>
<td>19.0</td>
<td>0.09</td>
</tr>
<tr>
<td>Mb</td>
<td>0.15</td>
<td>9.5</td>
<td>13.6</td>
<td>0.04</td>
</tr>
<tr>
<td>Mb</td>
<td>0.21</td>
<td>13.2</td>
<td>17.0</td>
<td>0.07</td>
</tr>
<tr>
<td>Mb</td>
<td>0.06</td>
<td>5.9</td>
<td>7.5</td>
<td>0.02</td>
</tr>
<tr>
<td>e-Globin</td>
<td>0.11</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Mb</td>
<td>0.08</td>
<td>5.6</td>
<td>9.1</td>
<td>0.08</td>
</tr>
<tr>
<td>Ngb</td>
<td>0.10</td>
<td>—</td>
<td>11.9</td>
<td>0.08</td>
</tr>
<tr>
<td>Boreoeutheria</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cygb</td>
<td>0.03</td>
<td>4.8</td>
<td>4.2</td>
<td>0.09</td>
</tr>
<tr>
<td>aG-Globin</td>
<td>0.15</td>
<td>20.8</td>
<td>17.0</td>
<td>0.12</td>
</tr>
<tr>
<td>aD-Globin</td>
<td>0.11</td>
<td>12.5</td>
<td>15.6</td>
<td>0.09</td>
</tr>
<tr>
<td>aE-Globin</td>
<td>0.17</td>
<td>17.7</td>
<td>12.0</td>
<td>0.08</td>
</tr>
<tr>
<td>e-Globin</td>
<td>0.17</td>
<td>13.7</td>
<td>12.4</td>
<td>0.11</td>
</tr>
<tr>
<td>p-Globin</td>
<td>0.27</td>
<td>19.1</td>
<td>13.6</td>
<td>0.16</td>
</tr>
<tr>
<td>Mb</td>
<td>0.10</td>
<td>16.0</td>
<td>15.6</td>
<td>0.12</td>
</tr>
<tr>
<td>Mb</td>
<td>0.05</td>
<td>5.9</td>
<td>4.6</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Notes.—aa, amino acid.

* ω = (dS/dD) estimates from codeml.

b Calculated in MEGA, excluding sites with alignment gaps in a pairwise manner.

c ω = (dS/dD) estimates from yN00.

Neognathae is the clade that includes chicken, duck, turkey, and zebra finch (Supplementary data file 2, Supplementary Material online).

Boreoeutheria is the clade that unites Euarcho­nog­lires and Laurasiatheria. We included sequences from megafauna and micro­chiropteran bats, cow, dog, guinea pig, human, macaque, marmoset, mouse, pika, and rabbit (Supplementary data file 3, Supplementary Material online).

In cases where only two avian sequences were available for comparison, codeml analyses are not reported.

Evolution of the Vertebrate Cygb, GbE, and Mb Genes

We integrated results of the combined phylogenetic and synteny analysis to reconstruct the evolutionary history of the paralogous Cygb, Mb, and GbE genes during the radiation of gnathostome vertebrates (fig. 7). Patterns of conserved macro­synteny demonstrate that the proto-Cygb and proto-GbE/Mb genes represent paralogous products of a whole­genome duplication event in the stem lineage of vertebrates (Storz et al. 2011). Results of the analyses presented here confirm that the vertebrate GbE and Mb genes represent the paralogous products of a subsequent tandem gene duplication that occurred in the stem lineage of gnathostome vertebrates (fig. 7). We therefore conclude that the Cygb, GbE, and Mb genes were present in the genome of the gnathostome common ancestor. In the branch leading to teleost fish, GbE was probably deleted prior to the fish-specific genome duplication (fig. 7). Based on data from the five teleost fish examined here, it appears that only one paralogous Mb gene was retained following the fish-specific genome duplication.

The Avian GbE Gene: Paradoxical Patterns of Evolution and Enigmatic Function

Our results indicate that the GbE gene has been secondarily lost in at least four major gnathostome lineages: teleost fish, amphibians, mammals, and nonavian reptiles. We cannot rule out the possibility that GbE has been retained in the genomes of cartilaginous fish, and despite the apparent absence of GbE in a representative squamate reptile (anole lizard), we cannot rule out the possibility that the gene has been retained in other nonavian reptiles. It remains to be seen whether GbE is found only in birds or whether it is also present in the genomes of other extant archosauromorph taxa (e.g., crocodilians and tuatara). At face value,
the low rate of retention of GbE among gnathostome lineages would seem to suggest that the gene is not performing an essential physiological function. Paradoxically, however, relative rate tests revealed that the avian GbE gene is characterized by a higher average level of amino acid sequence conservation than the Mb gene (table 2), and in orthologous comparisons among all members of the globin gene superfamily in birds, the GbE gene was characterized by higher levels of sequence conservation than all genes other than Ngb, Cygb, and βA-globin (table 3). We therefore conclude that GbE is performing an important physiological function and that function appears to be unique to the avian eye (Kugelstadt et al. 2004).

Other globin proteins such as Ngb and Cygb are also expressed in the retina of birds and other gnathostome vertebrates (Burmester et al. 2002, 2004; Kugelstadt et al. 2004; Fuchs et al. 2005, 2006; Hankeln et al. 2005; Xi et al. 2007; Hankeln and Burmester 2008; Burmester and Hankeln 2009). It has been suggested that these globins may perform an oxygen supply function to sustain mitochondrial respiration in the inner segments of retinal photoreceptors and other aerobically metabolizing cell types (Schmidt et al. 2003, 2005; Bentmann et al. 2005; Hankeln and Burmester 2008; Burmester and Hankeln 2009). For example, globins such as Ngb, Cygb, and GbE that are expressed in the retina could be performing Mblike respiratory functions by buffering transient fluctuations in cellular oxygen tension or by facilitating intracellular oxygen diffusion to the mitochondria (Sun et al. 2001, 2003). Alternatively (but also in analogy with Mb function), these globins could play a role in the detoxification of reactive oxygen and nitrogen species, they could help maintain nitric oxide (NO) homeostasis, or they could supply oxygen to cytoplasmic enzymes that catalyze oxygen-requiring reactions, such as NO production by NO synthases (Fago et al. 2004; Schmidt et al. 2004; Weber and Fago 2004; Hankeln et al. 2005; Fordel et al. 2007; Hankeln and Burmester 2008; Burmester and Hankeln 2009). Experimental efforts to characterize the structure, reactivity, and cellular localization of GbE will likely provide additional clues to the physiological function of this protein. The observed levels of sequence conservation suggest that such experiments would be worthwhile because whatever GbE is doing, it appears to be something important.

**Supplementary Material**

Supplementary figures 1–6 and supplementary tables 1–3 are available at Genome Biology and Evolution online (http://gbe.oxfordjournals.org/).

**Acknowledgments**

We thank Z. Cheviron, J. Projecto-Garcia, and three anonymous reviewers for helpful comments and suggestions.
This work was funded by grants to J.F.S. from the NSF (IOS-0949931) and NIH/NHLBI (R01 HL087216 and HL087216-S1) and grants to J.C.O. from the Fondo Nacional de Desarrollo Científico y Tecnológico (11080181), the Programa Bicentenario en Ciencia y Tecnología (PSDB9), and the Concurso Estadía Jóvenes Investigadores en el Extranjero from the Universidad Austral de Chile.

**Literature Cited**


Associate editor: Ross Hardison