Gene Turnover and Diversification of the $\alpha$- and $\beta$-Globin Gene Families in Sauropsid Vertebrates

Federico G. Hoffmann, Michael W. Vandewege, Jay F. Storz, and Juan C. Opazo

Abstract

The genes that encode the $\alpha$- and $\beta$-chain subunits of vertebrate hemoglobin have served as a model system for elucidating general principles of gene family evolution, but little is known about patterns of evolution in amniotes other than mammals and birds. Here, we report a comparative genomic analysis of the $\alpha$- and $\beta$-globin gene clusters in sauropsids (archosaurs and nonavian reptiles). The objectives were to characterize changes in the size and membership composition of the $\alpha$- and $\beta$-globin gene families within and among the major sauropsid lineages, to reconstruct the evolutionary history of the sauropsid $\alpha$- and $\beta$-globin genes, to resolve orthologous relationships, and to reconstruct evolutionary changes in the developmental regulation of gene expression. Our comparisons revealed contrasting patterns of evolution in the unlinked $\alpha$- and $\beta$-globin gene clusters. In the $\alpha$-globin gene cluster, which has remained in the ancestral chromosomal location, evolutionary changes in gene content are attributable to the differential retention of paralogous gene copies that were present in the common ancestor of tetrapods. In the $\beta$-globin gene cluster, which was translocated to a new chromosomal location, evolutionary changes in gene content are attributable to differential gene gains (via lineage-specific duplication events) and gene losses (via lineage-specific deletions and inactivations). Consequently, all major groups of amniotes possess unique repertoires of embryonic and postnatally expressed $\beta$-type globin genes that diversified independently in each lineage. These independently derived $\beta$-type globins descend from a pair of tandemly linked paralogs in the most recent common ancestor of sauropsids.

Key words: gene family evolution, genome evolution, gene expression, hemoglobin, gene duplication, amniotes.

Introduction

Repeated rounds of gene duplication and divergence can lead to the functional and regulatory diversification of multigene families, where different members acquire distinct biochemical functions and/or patterns of expression. During the evolution of deuterostomes, the duplication and functional divergence of members of the globin gene superfamily has been an important source of physiological innovation (Hoffmann, Opazo, et al. 2010; Hoffmann, Storz, et al. 2010; Blank et al. 2011; Storz, Opazo, et al. 2011; Hoffmann, Opazo, and Storz 2012; Hoffmann, Opazo, Hoogewijs, et al. 2012; Hoogewijs et al. 2012; Schwarze and Burnesteiner 2013; Storz et al. 2013; Burnesteiner and Hanks 2014; Schwarze et al. 2014, 2015; Opazo, Lee, et al. 2015). Within this diverse and ancient superfamily, vertebrate hemoglobin (Hb) genes comprise one of the most intensively studied gene families from a functional and evolutionary perspective (Hardison 2012; Storz 2016), providing an outstanding opportunity to assess the phenotypic consequences of changes in gene content.

The $\alpha$- and $\beta$-globin genes of jawed vertebrates (gnathostomes) encode subunits of tetrameric ($\alpha_2\beta_2$) Hb, the red blood cell protein that is responsible for the circulatory transport of oxygen and carbon dioxide. The ancestral $\alpha$- and $\beta$-globin genes derive from the tandem duplication of a proto-Hb gene in the common ancestor of gnathostomes, and the resulting linked arrangement of $\alpha$- and $\beta$-globin genes has been retained in cartilaginous fishes, ray-finned fishes, and...
amphibians (Jeffreys 1982; Hosbach et al. 1983; Gillemans et al. 2003; Fuchs et al. 2006; Opazo et al. 2013; Opazo, Hoffmann, et al. 2015; Opazo, Lee, et al. 2015). However, this ancestral arrangement has been lost in amniotes due to the translocation of the $\beta$-globin locus to a new chromosomal location, so that the $\alpha$- and $\beta$-globin gene clusters are located on different chromosomes in this group (Hardison 2008; Patel et al. 2008; Hoffmann, Opazo, and Storz 2012).

For historical reasons, comparative studies of the $\alpha$- and $\beta$-globin gene clusters of amniotes have mainly focused on mammals because of the greater availability of whole-genome assemblies. These studies have revealed very rapid rates of gene turnover, which result in high levels of variation in gene content among species (Hoffmann et al. 2008a, 2008b; Opazo et al. 2008a, 2008b, 2009; Storz et al. 2008; Runck et al. 2009; Gaudry et al. 2014). Our understanding of globin gene clusters in other amniotes is much more limited. The release of multiple avian genomes provided expanded opportunities for comparative studies of the $\alpha$- and $\beta$-globin clusters among vertebrates and revealed far lower rates of gene turnover in birds than in mammals (Zhang et al. 2014; Opazo, Hoffmann, et al. 2015). Until recently, a comprehensive examination of variation in saurupsid, the sister groups of mammals, was not possible due to a lack of comparative data from crocodilians, turtles, squamates (lizards and snakes), and rhyynchocephalians (represented by tuatara). This dearth of genomic data has limited our ability to decipher orthologous and paralogous relationships among the globin genes of different taxa. With the recent release of multiple nonavian saurupsid genomes (Alfo¨lchi et al. 2011; Castoe et al. 2013; Wan et al. 2013; Wang et al. 2013; Green et al. 2014; Liu et al. 2015), we can now extend these studies to include all major groups of amniotes.

Comparative studies suggest that most amniotes have retained copies of three tandemly linked $\alpha$-type globin genes that were present in the last common ancestor of tetrapods: $\alpha^A$-$\alpha^D$-$\alpha^A$-$3^\prime$ (Hoffmann and Storz 2007; Hoffmann, Storz, et al. 2010; Grispo et al. 2012). With the exception of mammals, where most species have multiple copies of $\alpha^A$- and $\alpha^D$-globin, most variation in gene family size among tetrapods is attributable to lineage-specific losses of one of the three ancestral paralogs. By contrast, the $\beta$-globin gene clusters have had a much more dynamic duplicative history, with multiple lineage-specific expansions of gene family size. As a result, the major lineages of amniotes share a common set of orthologous $\alpha$-globin genes, but they have unique sets of $\beta$-globin genes that diversified via independent rounds of lineage-specific duplication and divergence (Opazo et al. 2008a; Hoffmann, Storz, et al. 2010; Storz, Hoffmann, et al. 2011; Storz 2016). In general, the linkage order of the globin genes reflects their temporal order of expression during development (Cirotto et al. 1987; Ikehara et al. 1997; Alev et al. 2008, 2009; Storz, Hoffmann, et al. 2011). This is the case in the mammalian and avian $\beta$-globin gene clusters, even though the lineage-specific gene repertoires diversified independently. In both cases, the $\beta$-type globin genes located at the 5′ end of the cluster are expressed during the earliest stages of embryogenesis whereas the genes at the 3′ end are expressed during adulthood.

In the current study, we take advantage of newly released whole genome sequences to investigate patterns of diversification of saurupsid $\alpha$- and $\beta$-globin genes. Specifically, the objectives of the present study were 1) to characterize changes in the size and membership composition of the $\alpha$- and $\beta$-globin gene families within and among the major sauropсид lineages, 2) to reconstruct the evolutionary history of the saurupsid $\alpha$- and $\beta$-globin genes and to resolve orthologous relationships among them, and 3) to reconstruct evolutionary changes in the developmental regulation of gene expression. Our comparisons revealed contrasting patterns in the evolution of the $\alpha$- and $\beta$-globin clusters. Evolutionary changes in the $\alpha$-globin gene cluster are attributable to the differential retention of ancestral duplicates among lineages, whereas changes in the $\beta$-globin gene cluster are attributable to differential gene gains and losses. In addition, we identified differences in $\alpha$- and $\beta$-globin expression between sauropids and the group that includes testudines, crocodilians, and birds. In the case of $\beta$-globin, squamates and rhyynchosphenians express adult Hb isoforms that incorporate the products of different $\beta$-type globin genes, whereas a single but independently derived gene encodes the $\beta$-type subunits of adult Hb in mammals, archosaurs, and turtles. Taken together, our results suggest that the regulatory architectures of the $\beta$-globin gene clusters evolved independently in mammals and in the group comprising archosaurs plus turtles.

Materials and Methods

Bioinformatic Searches

We implemented bioinformatic searches for $\alpha$- and $\beta$-globin sequences in saurupsid genomes in two stages: we first focused on genomic data and we then incorporated auxiliary sequence data from transcriptomes. To characterize the saurupsid $\alpha$- and $\beta$-globin gene clusters we searched for traces of $\alpha$- and $\beta$-globin genes in the genomic sequence records from representative saurupids in the Ensembl or NCBI (refseq genomic, hts, and vgs) databases using BLAST (Altschul et al. 1990). In all cases we used low stringency settings, searching with the BLASTn algorithm, with match/mismatch scores of 1 and −1, and gap existence and extension costs of 2 and 1, respectively. The full set of species is listed in supplementary table S1, Supplementary Material online, and includes ten squamates, four testudines, four crocodilians, and two birds. Even though there are more bird genome sequences available, variation in the $\alpha$- and $\beta$-globin gene clusters of birds, which is the most studied group of
sauropsids, has been covered recently (Opazo, Hoffmann, et al. 2015). Thus, we have included two species in the current study, zebra finch and chicken, which have high-quality genomes and provide good representation of avian diversity for comparative purposes. Once we located all genomic fragments containing traces of $\alpha$- and $\beta$-like globin genes, we extracted the corresponding fragments and verified preliminary annotations by comparing identified sequences against the well-annotated $\alpha$- and $\beta$-globin genes of human and chicken using BLAST2 (Tatusova and Madden 1999), as in Hoffmann, Storz, et al. (2010). In a second stage, we interrogated sauropsid transcriptomes available in NCBI and the Reptilian Transcriptomics database (www.reptilian-transcriptomes.org; last accessed June 2017; Tzika et al. 2015) for $\alpha$- and $\beta$-type globin sequences in order to expand our taxonomic coverage. Information regarding the sources of all sequences in our analyses is provided in supplementary tables S1–S3, Supplementary Material online. For the purpose of our analyses, sequences were considered putatively functional if the predicted lengths of the exons matched expected lengths for amniote $\alpha$- and $\beta$-globins, if there were no premature stop codons, and if start and stop codons were found at expected positions. Conversely, $\alpha$- and $\beta$-like globin sequences were considered as pseudogenes if they include premature stop codons.

Data Curation

In the current assembly of the anole lizard genome, the $\alpha^A$ globin gene lies on a separate contig from the $x^D$ and Gby$^\gamma$ genes, whereas the king cobra assembly is missing the $\alpha^A$ gene altogether. However, we suspect that these are issues related to the current assemblies of anole lizard and king cobra genomes because the gekko assembly places $\alpha^A$ in the canonical genomic location, flanked by $x^D$ and Gby$^\gamma$. In contrast to mammals, pseudogenes are rare in the $\alpha$-globin clusters of sauropsids; the sole exception was a clearly recognizable $\alpha^A$ pseudogene found in the python. In addition, our bioinformatic searches identified two different $\alpha^D$ sequences in the common viper genome, two different $\alpha^D$ sequences in the garter snake genome, and two identical $\alpha^D$ paralogs in the same fragment of the Chinese softshell turtle genome. Given the nature of the current assemblies, we assume that this apparent sequence heterogeneity is attributable to assembly artifacts, although we cannot rule out the possibility that the sequences represent products of recent duplication events within each of the species.

The current assemblies of most of the sauropsid $\beta$-globin gene clusters are fragmented and appear to be missing genes in some cases. For example, the most current assembly of the anole lizard genome, AnoCar2.0, includes a single $\beta$-globin gene (Ensembl: ENSACAG00000012173, GenEd: 100552694), but analyses of transcriptomic and proteomic data indicate the presence of at least one additional $\beta$-type globin gene that is clearly distinct at the sequence level (Storz, Hoffmann, et al. 2011), and was present in AnoCar1.0 (ENSACAG00000010799). In the case of the Chinese alligator genome, the assembly includes an $\alpha$-globin gene that is clearly distinct at the sequence level (Hoffmann et al. 2011; Gaudry et al. 2014), to err on the side of caution, we provisionally consider these transcripts as bioinformatic artifacts.

Because some squamate genomes have incomplete coverage of the genes of interest, we completed the set of genomic sequences with additional transcriptome records whenever possible. For example, in the case of the corn snake and the python, the current assembly includes portions of two distinct genes, which correspond to the two $\beta$-globin gene sequences derived from their transcriptomes (available at the Reptilian Transcriptomics 2.0 database; Tzika et al. 2015). In the case of the two rattlesnake species for which genome sequence data were available, the current assemblies lack complete coverage of the $\alpha$-globin gene cluster, so we also included the globin repertoire of the closely related Eastern diamond rattlesnake, derived from a venom-gland transcriptome (Rokyta et al. 2012). Similarly, in the case of the garter snake, the $\alpha^A$-globin gene is not present in the current assembly, so we included a copy of this gene from a transcriptome of the same species. In a second stage, we extended our analyses to include sequences derived from transcriptomes, with a view to increasing taxonomic coverage of squamates and testudines. Finally, in the case of the tuatara, because it is the sole representative of the order Rhynchocephalia, we included three transcriptome-derived sequences in the analyses of genome-derived sequences (two $\alpha$-type globins and one $\beta$-type globin). We also compared amino acid sequences derived from direct peptide sequencing to conceptual translations of the coding sequences. The list of these additional amino acid sequences is provided in supplementary table S4, Supplementary Material online.

In many squamate transcriptomes we identified $\beta$-globin paralogs that appear to be assembly artifacts, as initial searches identified the presence of three or four putative paralogs, but careful examination revealed that these sequences represent different chimeric combinations of two closely related sequences. By default, we assumed that the paralogs more similar to genome-derived sequences were the “true” paralogs. Although chimeric fusion genes are relatively common among mammalian $\beta$-globins (Hoffmann et al. 2008b; Opazo et al. 2009; Gaudry et al. 2014), to err on the side of caution, we provisionally consider these transcripts as bioinformatic artifacts.
Phylogenetic Analyses

Phylogenetic relationships among the α- and β-type globin sequences were estimated using maximum likelihood (ML) and Bayesian analyses (BA). For these analyses, nucleotide and amino acid identity of vertebrate globin genes were aligned using the program MAFFT version 7.304 (Katoh and Standley 2013), as implemented in the following server: http://mafft.cbrc.jp/alignment/server/, last accessed on January 2017. ML analyses were run using IQ-Tree ver 1.5.5 (Nguyen et al. 2015) in the implementation of IQ-Tree available from the IQ-Tree web server (Trifinopoulos et al. 2016) last accessed on June 2017, and support for the nodes was evaluated with 1, 000 pseudoreplicates of the ultrafast bootstrap procedure (Minh et al. 2013). Bayesian Analyses were performed in MrBayes version 3.2 (Ronquist et al. 2012), running four simultaneous chains for 2 × 10^7 generations, sampling trees every 1,000 generations, and using default priors. We assessed convergence by measuring the SD of the split frequency among parallel chains. Chains were considered to have converged once the average split frequency was <0.01. We discarded trees collected before the chains reached convergence, and we summarized results with a majority-rule consensus of trees collected after convergence was reached. Alternative topologies were compared using the approximately unbiased test proposed by Shimodaira and Goldman (2002), as implemented in IQ-TREE.

Organismal Phylogeny and Divergence Dates

In all cases, we assumed that relationships among the different lineages studied: birds, crocodilians, rhynchocephalians, squamates, testudines, plus amphibians and mammals which were included as outgroups follow the arrangement reported by Crawford et al. (2012), based on the analyses of ultraconserved genetic elements. Estimates of divergence times among the lineages were obtained from the TimeTree server (Kumar, Stecher, et al. 2015). Sequences corresponding to the α^1 and α^0 genes were identified using BLASTX and a database of previously annotated α^1 and α^0 amino acid sequences. Raw paired-end Illumina reads (SRA062458 and SRP071466) were cleaned and trimmed using Trimomatic (Bolger et al. 2014) and were used to estimate expression of α^1 and α^0-globin in each species with RSEM (Li and Dewey 2011). Transcripts per million (TPM) was used as the expression metric.

Gene Expression Analyses

We collected publically available squamate, crocodilian, and turtle liver transcriptomes from McGaugh et al. (McGaugh et al. 2015). Sequences corresponding to the α^0 and α^1 genes were identified using BLASTX and a database of previously annotated α^1 and α^0 amino acid sequences. Raw paired-end Illumina reads (SRA062458 and SRP071466) were cleaned and trimmed using Trimomatic (Bolger et al. 2014) and were used to estimate expression of α^1 and α^0-globin in each species with RSEM (Li and Dewey 2011). Transcripts per million (TPM) was used as the expression metric.

Results

We integrated synteny, phylogenetic, and expression analyses to reconstruct the evolution of the α- and β-globin gene clusters of sauropsids. Briefly, we surveyed genome databases to locate the sauropsid α- and β-globin gene clusters. We then annotated the α- and β-type globin genes in conjunction with the genes flanking the 5’ and 3’ ends of the globin clusters, and we used phylogeny reconstructions to resolve orthology and paralogy. We then compared the resulting trees with the organismal phylogeny to reconstruct ancestral repertoires and we incorporated gene and/or protein expression data to infer developmental patterns of expression. These surveys spanned all orders of sauropsids other than rhynchocephalians and included ten species of squamates (both lizards and snakes), four testudines, four crocodilians, plus zebra finch, and chicken as representative birds (see full list in supplementary table S1, Supplementary Material online).

Synteny

Comparative genomic analyses revealed that the α- and β-globin gene clusters of sauropsids both exhibit a high level of
conserved synteny. The 5' end of the α-globin cluster is flanked by orthologs of NPRL3, which spans multiple important cis-regulatory elements that govern the expression of α-type globins (fig. 1). Available data suggest that the conserved synteny of the genes flanking the sauropod α-globin gene cluster extends to RBHDF1 and MPG (fig. 1). On the 3' end of the α-globin gene cluster, sauropods share a derived, small-scale inversion, so that the cluster is flanked by TMEM8A, rather than by Luc7L, as in mammals and other vertebrate groups (Opazo, Lee, et al. 2015). In addition, squamates, crocodilians, and testudines, but not birds, have retained a copy of Globin-Y (GbY) at the 3' end of the α-globin gene cluster, an ancient vertebrate globin with an unknown functional role that also flanks the α-globin gene cluster of platypus, and the α- and β-globin gene clusters of frog (Xenopus), spotted gar, and elephant shark (fig. 1).

Patterns of conserved synteny in the β-globin gene cluster are less clear due to the fragmentary nature of this chromosomal region in the current genome assemblies of most sauropods. Among the species studied, flanking gene information for both sides of the β-globin gene clusters is available for chicken, the two alligators, and painted turtle, and in all these cases the β-globin gene cluster is flanked by olfactory receptors on both sides, as is the case with mammals (fig. 1). This also seems to be the case for king cobra and common viper as well, but the current assemblies are too

---

**Fig. 1.**—Synteny of the α- and β-globin gene clusters of representative vertebrates (arranged from left to right in 5' to 3' order) with a focus on sauropods. Genes inferred to be orthologs are linked by vertical lines, and adult-expressed genes are identified by an asterisk. White boxes correspond to pseudogenes. In the case of the to squamates, the lizard gene clusters combine data from Japanese gekko (α-globin) and the AnoCar1.0 release of the Green anole (β-globin), and the snake gene clusters combine data from king cobra (α-globin) and Burmese python (β-globin). Note that we lack a genome assembly for the tuatara, the sole representative of the order Rhynchocephalia. The secondary loss of GbY in birds and humans is denoted by an “x.”
fragmentary to draw definitive conclusions. Because of the high rate of gene gain and loss among the olfactory receptors of the different sauropsid lineages (Vandewege et al. 2016), establishing 1:1 orthologous relationships among olfactory receptors of these different lineages was not possible.

The α- and β-Globin Gene Clusters of Sauropsids

We found limited variation in the number of genes in the α- and β-globin gene clusters of most sauropsid taxa relative to mammals (fig. 1). Birds and turtles have retained the ancestral complement of three α-type globin genes (αA, αD, and αE), whereas the other sauropsid taxa possessed alternative combinations of two genes: αD and αA in squamates, and αE and αA in crocodilians. We verified the absence of the αE-globin gene in squamates by comparing crocodilian and testudine αE-globin sequences against squamate sequences in GenBank and the Reptilian transcriptomics databases using low stringency searches. We were able to retrieve the αD and αA paralogs that were already present in our analyses, and the more distant β-paralogs as well, but we found no matches that corresponded to an αE-like globin sequence. We also verified the absence of αD-globin in crocodiles by comparing avian and testudine αE-globin sequences against crocodilian sequences in GenBank and the Reptilian transcriptomics databases. In this case, we were able to retrieve the αE and αA sequences that were already present in our analyses, and we also recovered an intronless retroprocessed αD-globin pseudogene in the genomes of all crocodilians, located on a separate scaffold from the α-globin gene cluster, and flanked by the craniofacial development protein 1 (CFDP1) and the breast cancer antiestrogen resistance protein 1 (BCAR1) in each of the four crocodilian species surveyed. The fact that our protocols can identify β-type globin sequences when seeded with an α-like sequence, and can also recover divergent pseudogenes such as the αE-globin pseudogene of crocodilians, suggests that our searches were comprehensive. In addition, average pairwise distances among the αA, αD, and αE paralogs range from 38% to 40% at the nucleotide level, and from 48% to 50% at the amino acid level (supplementary table S5 and supplementary files S1 and S2, Supplementary Material online), indicating that these three α-globin paralogs are clearly distinct from each other. We therefore conclude that the apparent absence of αD in crocodilians and the absence of αE-globin in squamates are not artifacts of our search protocol.

The β-globin gene cluster of sauropsids is more variable than the α-globin cluster. In this case, we could not resolve orthology relationships using bioinformatic searches. Thus, we initially labeled the genes using numbers, which in the case of well-resolved clusters use the prefix “T,” and the numbers denote the order of the gene in the tandem array, so that the T1 is the first gene on the 5′ end, followed by T2 and so forth. We found two intact β-globin genes in most squamates, but could not determine their order in any of the species examined. On the other hand, we were able to resolve the tandem arrays of β-globin genes for birds, crocodiles, and testudines. There were two or three genes in turtles, labeled T1–T3, four in birds, labeled T1–T4 (which correspond to ρ-, β1-, β2-, and ε-globin), and four genes with three pseudogenes in crocodilians, labeled T1–T7 (fig. 1). The Chinese alligator was the only exception, as this species appears to have an en bloc duplication of the β-globin gene cluster, which includes five putatively functional genes and seven pseudogenes labeled T1–T12, where the T7–T10 block of genes is a duplication of the T3–T7 paralogs (supplementary fig. S1, Supplementary Material online). Average pairwise distances for the different sets of β-globins paralogs were lower than the distances among α-globin paralogs, ranging from 8% to 33%, (supplementary table S6 and supplementary files S3 and S4, Supplementary Material online).
Fig. 2.—Phylogeny of sauropsid α-type globin genes derived from genomes and transcriptomes. Support for the relevant nodes are indicated as bootstrap percentages from IQ-Tree above the nodes, and as Bayesian posterior probabilities from MrBayes below.
**Fig. 3.**—Phylogeny of sauropsid β-type globin genes derived from genomes and transcriptomes. Genes or groups of genes that are expressed in adulthood are underlined. Support values for relevant nodes are indicated as bootstrap percentages (above) and as Bayesian posterior probabilities (below).
containing β-globin sequences of birds, crocodilians, and testudines. In general, the arrangements do not deviate significantly from the expected organismal relationships (Crawford et al. 2012). The one exception was the presence of one tuatara beta globin gene, SPU_ENSACAG00000012173, in the clade that includes birds, turtles, and crocodiles, which suggests that this gene lineage traces back to the ancestor of sauropsids and that it was secondarily lost in the ancestor of lizards and snakes. Forcing the β-globins of this second clade to follow the expected organismal arrangement, with tuatara as the deepest node, followed by testudines, with crocodilian β-globins as sister to avian β-globins did not result in a significant loss in likelihood score in an Approximately Unbiased topology test.

Within the squamate clade, there are two well-defined groups of snake β-globins, corresponding to the βA and βB clades defined by Storz et al. (2015). Lizard β-globins are paraplethetic relative to the snake β-globin clade (fig. 3) and relationships among the lizard sequences do not deviate from known organismal relationships (Pyron et al. 2013). Orthology among lizard β-globins can only be resolved for a small subset of the genes because when multiple β-globin paralogs for a lizard species are present, they are usually very similar (fig. 3), a pattern that may reflect a history of lineage-specific gene duplications, interparalog gene conversion, or a combination of these two processes. For example, intraspecific distances among the β-globin paralogs of lizards range from 1% to 15%, lower than the comparisons between the snake βA- and βB-globins, which range from 22% to 27% (supplementary file S3, Supplementary Material online).

In our maximum-likelihood analyses, avian, crocodilian, and testudine β-globins form a monophyletic clade (fig. 3). In all cases, we found that genes at the 5’ and 3’ ends of the cluster grouped together in a clade, and the genes in center of the cluster are grouped in a second clade. Thus, the avian T1 (α) and T4 (ω) paralogs are sister to the clade that includes the avian T2 (βA) and T3 (βB) paralogs, the crocodilian T1 and T7 paralogs are sister to the clade that includes the T2–T6 paralogs, and the testudine T1 and T3 paralogs are sister to the clade of testudine T2s (fig. 3). It is noteworthy that crocodilian pseudogenes are shared across all species, which indicates they were already inactive in their last common ancestor, which existed ~80 Ma.

We then estimated phylogenetic affinities based on amino acid sequences in order to include two tuatara β-globin amino acid sequences, P10060 and P10061, as well protein records from testudines, squamates, and crocodilians that have no associated nucleotide sequence (fig. 4). The tree based on amino acid sequences was largely congruent with the one based on nucleotide sequences, but resolution was relatively poor due to the large number of sequences included relative to the number of characters (108 vs. 147). One of the tuatara β-globin amino acid sequences, P10060, grouped with the squamate β-globins, whereas the other two tuatara β-globins, protein record P10061 and the one derived from the transcriptome, fell in the clade that includes the β-globins of crocodilians, birds, and the T1 and T3 paralogs of testudines. Integrating the nucleotide and amino acid based phylogenies with protein and transcript sequences indicates that the β-chain subunits of adult Hb are products of single genes: Hbb-T3 (βA) in birds, Hbb-T4 in crocodilians, and Hbb-T2 in testudines.

**Discussion**

Our analyses capture the contrasting evolutionary pattern of the α- and β-globin gene families of sauropsids. In the case of the α-globin gene family, differences are restricted to the differential loss of one of the three paralogs present in the common ancestor. As a result, testudines and birds have retained copies of αA, αD, and αE, whereas squamates have retained αD and αA and crocodilians have retained αE and αD (along with a vestige of αD in the form of processed pseudogene). By contrast, variation in gene copy number is more extensive in the β-globin gene cluster (fig. 1). Our previous study suggested that each of the major lineages of amniotes evolved distinct β-globin repertoires via repeated rounds of lineage-specific gene duplication (Hoffmann, Storz, et al. 2010). However, that study only included anole lizard as a representative of squamates and it did not include turtles or crocodilians. Interestingly, the phylogenies obtained with our increased taxonomic sampling suggest an even more extreme pattern of lineage-specific β-globin expansion (fig. 3), where the major groups of sauropsids (birds, crocodilians, testudines, and squamates) each possess repertoires that derive from lineage-specific duplications. In spite of the observed contrast in evolutionary dynamics between the α- and β-globin gene families, variation in gene content within the different sauropsid groups appears to be limited. This is in stark contrast to mammals, which exhibit high levels of variation in the size and membership composition of the α- and β-globin gene clusters (Hoffmann et al. 2008a, 2008b; Opazo et al. 2008a, 2008b; Gaudry et al. 2014).

**Evolution of the Sauropsid α- and β-Globin Gene Families**

Reconciling individual gene trees with the organismal phylogeny is straightforward in the case of the α-globin gene family, as variation in gene content among groups is strictly attributable to lineage-specific losses of single members of the three-gene set that was present in the common ancestor of sauropsids. In the case of the β-type globins, we base our inferences on the tree based on nucleotide data, as it is better resolved and includes the crocodilian pseudogenes (fig. 3). The monophyly of amniote β-globins suggests that these genes can be traced back to a single gene in the last common ancestor of amniotes, and implies that a single β-globin was translocated to the novel chromosomal location. In turn, the
**FIG. 4.**—Phylogeny of β-type globin genes based on amino acid sequences, including additional protein records. Genes or groups of genes that are expressed in adulthood are underlined. Support values for relevant nodes are indicated as bootstrap percentages (above) and as Bayesian posterior probabilities (below).
affinity of the one tuatara \(\beta\)-globin transcript with the \(\beta\)-globins of birds, crocodilians, and turtles implies that the last common ancestor of sauropsids had at least two different \(\beta\)-globins: one retained by tuatara and squamates, and one retained by tuatara, archosaurs, and turtles. This inference is confirmed by the tree based on amino acid sequences, which includes two additional tuatara \(\beta\)-globin sequences, P10060 and P10061. The tuatara P10060 \(\beta\)-globin sequence groups with the \(\beta\)-globins of squamates, and the other two tuatara \(\beta\)-globins, P10061 and SPU_ENSACAG00000012173, group with the \(\beta\)-globins of birds, crocodiles, and turtles, indicating the presence of at least two \(\beta\)-globin genes in the last common ancestor of sauropsids (fig. 5).

Differences in the \(\beta\)-globin gene repertoires among sauropsid groups are the result of the differential retention of two ancestral genes, followed by lineage-specific duplications, which in the case of some squamates have been obscured by a history of interparalog gene conversion (Hoffmann, Storz, et al. 2010; Opazo, Hoffmann, et al. 2015). Among the \(\beta\)-type globin genes of squamates, the phylogenetic analyses indicate that the snake \(\beta^\text{II}\)-globins represent the most early branching lineage (figs. 3 and 4), which suggests that the common ancestor of squamates had at least two \(\beta\)-globin genes, which have only been retained as a distinct pair in snakes. Most lizards analyzed have two \(\beta\)-globin genes that are most closely related to snake \(\beta^\text{I}\)-globins, but robust orthology inferences are not possible due to an apparent history of interparalog gene conversion, gene turnover, or both.

The \(\beta\)-globin gene repertoires of birds, crocodilians, and testudines diversified independently via repeated rounds of lineage-specific duplication and descend from only one of the two ancestral sauropsid \(\beta\)-globin paralogs. Interestingly, even though bird, crocodilian, and testudine \(\beta\)-globins fall in reciprocally monophyletic groups, reflecting independent duplicative histories, in all cases the genes at the two ends
of the cluster group together, and are sister to the clade of genes in the center of the cluster. Because there is evidence for gene conversion between the embryonically expressed T1 (μ) and T4 (ω) globin genes of birds, we suspect a similar process is partly responsible for this arrangement in testudines and crocodilians. (Hoffmann, Storz, et al. 2010; Opazo, Hoffmann, et al. 2015). Interestingly, most gene turnover seems to have occurred during the early evolution of sauropsids, as the repertoires are different among the major lineages in this group, but highly conserved within them.

Evolution of Globin Expression

In all gnathostome vertebrates studied to date, the expression of the α- and β-type globin genes is ontogenetically regulated so that structurally and functionally distinct Hb isoforms are expressed during different stages of development (Storz 2016). Thus, changes in the size and membership composition of the α- and β-globin gene families may produce changes in the developmental regulation of Hb synthesis which can then constrain or potentiate the functional differentiation between Hb isoforms (Opazo et al. 2008a, 2013; Runck et al. 2009; Hoffmann, Storz, et al. 2010; Storz, Hoffmann, et al. 2011; Storz, Opazo, et al. 2011; Grispo et al. 2012; Damsgaard et al. 2013; Storz et al. 2013; Gaudry et al. 2014). In fact, the α- and β-globin gene clusters of teleost fish, lobe-finned fish, amphibians, squamates, birds, and mammals have all diversified in a lineage-specific manner such that each of those groups has distinct repertoires that are differentially expressed during embryonic development and postnatal life (Opazo et al. 2008a, 2013; Hoffmann, Storz, et al. 2010; Storz, Hoffmann, et al. 2011).

Adult birds typically coexpress two structurally and functionally distinct Hb isoforms in definitive red blood cells: the major HbA isoform, which incorporates products of the αA-globin gene, and the minor HbD isoform, which incorporates products of the αD-globin gene (Grispo et al. 2012; Opazo, Hoffmann, et al. 2015). Turtles also coexpress different adult Hb isoforms that incorporate the αA- and αD-globin genes. In both turtles and birds, the Hb isoform that incorporates the product of the αA-globin gene is the major isoform in adult red blood cells, and in both taxa the minor “D” type isoform exhibits an appreciably higher O2-binding affinity (Grispo et al. 2012; Damsgaard et al. 2013; Projecto-Garcia et al. 2013; Cheviron et al. 2014; Galen et al. 2015; Natarajan et al. 2015, 2016; Opazo, Hoffmann, et al. 2015; Kumar, Natarajan, et al. 2017). Squamates exhibit a different pattern of expression. In fact, in rattlesnakes and anole lizards, Hb isoforms that incorporate products of the αD-globin gene are actually expressed at higher levels than those that incorporate products of the αA-globin gene (Storz, Hoffmann, et al. 2011; Storz et al. 2015). Consistent with these measurements of protein abundance, the squamate transcriptome data revealed a higher abundance of αD transcripts relative to αA transcripts in adult red blood cells. An assessment of RNA libraries from adult liver provides further support for this observation, as the ratio of reads mapping to αD and αA is generally larger in squamates than in turtles (supplementary table S5, Supplementary Material online). In snakes, the major adult D-type Hb isoform actually exhibits a lower O2-affinity than the minor A-type isoform (Storz et al. 2015). Thus, in birds, turtles, and squamates, it appears that the major adult Hb isoform always has a lower O2-affinity than the minor isoform, but the subunit composition of the major and minor isoforms varies among lineages.

Analyses based on protein data suggest that Hb isoform composition is qualitatively different in tuatara and squamates relative to other amniotes (Abbasi et al. 1988; Gorr et al. 1998; Storz, Hoffmann, et al. 2011; Lu et al. 2015; Storz et al. 2015). Both squamates and tuatara express multiple, structurally distinct β-type globins during adulthood, whereas birds, crocodilians, turtles, and mammals generally possess a single adult β-type globin gene, or they possess two or more copies that are identical or nearly so (Abbasi et al. 1988; Gorr et al. 1998; Opazo et al. 2009; Runck et al. 2009, 2010; Storz, Hoffmann, et al. 2011; Storz et al. 2012, 2015; Weber et al. 2013; Gaudry et al. 2014; Lu et al. 2015; Schwarze et al. 2015). It is worth noting the long branch leading to the adult crocodilian β-type globin gene, Hbb-T4, reflects a high rate of amino acid substitution that may be related to the evolution of uniquely derived functional properties of crocodilian Hb (Bauer et al. 1981; Perutz et al. 1981; Komiyama et al. 1995; Weber et al. 2013). Interestingly, even though tuatara and squamates both possess two β-type globins, these genes clearly do not descend from the same ancestral gene pair (i.e., the tuatara and squamate β-globins are not 1:1 orthologs; figs. 3–5).

Incorporating an explicit evolutionary framework into the comparative genomic analysis of these differentially expressed globins can reveal important clues as to when and how the genetic control of their expression emerged. This is particularly interesting in the case of the amniote β-globin gene cluster, because it translocated to a novel position in the genome, altering its original genomic context (Hardison 2012). In mammals and birds, the switch from embryonic to fetal and then adult Hb during development is a complex process that involves interactions between distal cis regulatory sequences (located 40- to 60-kb upstream of the β-globin cluster) and more proximal elements (Alev et al. 2009; Wilber et al. 2011; Hardison 2012; Ulianov et al. 2012). Moreover, position effects appear to be important, since the 5′ to 3′ gene order generally matches the order of their expression during development. This also seems to be the case in crocodilians and turtles, where the genes in the 5′ end of the cluster apparently do not encode β-chain subunits of adult Hb.
Evolution of the α- and β-Globin Gene Clusters in Amniotes

Our evolutionary reconstructions show a clear contrast in the tempo and mode of evolution of the α- and β-globin gene clusters in amniotes. In the case of the α-globin gene cluster, which has remained in the ancestral genomic location, variation in gene content is largely limited to differential losses of α-globin paralogs that were present in the common ancestor of the group. However, expression changes have evolved: in squamates and tuatara, the major adult Hb isoform incorporates products of the α2-globin gene, whereas in archosaurs and turtles, the major isoform incorporates products of the α2-globin gene. In contrast to the high degree of conserved synteny in the α-globin gene cluster, the β-globin cluster, which translocated to a new chromosomal location in the common ancestor of amniotes, has experienced a much higher rate of gene gain and loss. Consequently, all major groups of amniotes (i.e., mammals, birds, crocodilians, turtles, rhynchocephalians, and squamates) possess unique repertoires of β-type globin genes that diversified independently in each lineage. These independently derived β-globin gene clusters trace back to two genes in the last common ancestor of sauropsids, which in turn trace back to a single copy gene in the last common ancestor of amniotes. Tuatara is the only extant taxon that retains descendant copies of both ancestral sauropsid β-globins. The two β-globin paralogs of squamates descend from one of these ancestral genes, whereas the β-globin paralogs of archosaurs and turtles descend from the other ancestral gene copy (fig. 5). In adult red blood cells, tuatara and squamates coexpress structurally distinct Hb isoforms that incorporate products of both β-type globin paralogs. By contrast, the adult Hbs of archosaurs and turtles incorporate the products of a single β-type globin gene. Interestingly, however, the adult β-type globin genes of the archosaurs and turtles are not 1:1 orthologs, and in each case the adult-expressed paralogs are located downstream from early and late-expressed adult-expressed paralogs. By contrast, the adult Hbs of archosaurs and turtles trace back to two genes in the last common ancestor of sauropsids of the α2-globin gene cluster, whereas in archosaurs and turtles, the major isoform incorporates products of the α2-globin gene. In contrast to the high degree of conserved synteny in the α-globin gene cluster, the β-globin cluster, which translocated to a new chromosomal location in the common ancestor of amniotes, has experienced a much higher rate of gene gain and loss. Consequently, all major groups of amniotes (i.e., mammals, birds, crocodilians, turtles, rhynchocephalians, and squamates) possess unique repertoires of β-type globin genes that diversified independently in each lineage. These independently derived β-globin gene clusters trace back to two genes in the last common ancestor of sauropsids, which in turn trace back to a single copy gene in the last common ancestor of amniotes. Tuatara is the only extant taxon that retains descendant copies of both ancestral sauropsid β-globins. The two β-globin paralogs of squamates descend from one of these ancestral genes, whereas the β-globin paralogs of archosaurs and turtles descend from the other ancestral gene copy (fig. 5). In adult red blood cells, tuatara and squamates coexpress structurally distinct Hb isoforms that incorporate products of both β-type globin paralogs. By contrast, the adult Hbs of archosaurs and turtles incorporate the products of a single β-type globin gene. Interestingly, however, the adult β-type globin genes of the archosaurs and turtles are not 1:1 orthologs, and in each case the adult-expressed paralogs are located downstream from embryonically expressed paralogs. Thus, our results indicate that the stage-specific expression of early and late-expressed β-globins evolved at least twice in amniotes, once in mammals, and once in the lineage leading to birds, crocodilians and turtles.

Supplementary Material
Supplementary data are available at Genome Biology and Evolution online.

Acknowledgments
The authors thank Mary Clay Bailey and Amanda Black for editorial assistance. This work was supported by the National Institutes of Health/National Heart, Lung, and Blood Institute (HL087216 to J.F.S.), the National Science Foundation (MCB-1517636 and RII Track-2 FEC 1736206 and DEB-1354147 to F.G.H.), and the Fondo Nacional de Desarrollo Científico y Tecnológico from Chile (FONDECYT 1160627 to J.C.O.).

Literature Cited


Gene Turnover and Diversification in Sauropsid Vertebrates


