J Mol Evol (2004) 59:177–189
DOI: 10.1007/s00239-004-2612-0

JOURNAL OF MOLECULAR EVOLUTION



# Gene Conversion and Functional Divergence in the β-Globin Gene Family

**Gabriela Aguileta, Joseph P. Bielawski, Ziheng Yang**

Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, England

Received: 17 July 2003 / Accepted: 16 February 2004 [*Reviewing Editor*: Martin Kreitman]

**Abstract.** Different models of gene family evolution have been proposed to explain the mechanism whereby gene copies created by gene duplications are maintained and diverge in function. Ohta proposed a model which predicts a burst of nonsynonymous substitutions following gene duplication and the preservation of duplicates through positive selection. An alternative model, the duplication–degeneration–complementation (DDC) model, does not explicitly require the action of positive Darwinian selection for the maintenance of duplicated gene copies, although purifying selection is assumed to continue to act on both copies. A potential outcome of the DDC model is heterogeneity in purifying selection among the gene copies, due to partitioning of subfunctions which complement each other. By using the $d_N/d_S$ ($\omega$) rate ratio to measure selection pressure, we can distinguish between these two very different evolutionary scenarios. In this study we investigated these scenarios in the β-globin family of genes, a textbook example of evolution by gene duplication. We assembled a comprehensive dataset of 72 vertebrate β-globin sequences. The estimated phylogeny suggested multiple gene duplication and gene conversion events. By using different programs to detect recombination, we confirmed several cases of gene conversion and detected two new cases. We tested evolutionary scenarios derived from Ohta's model and the DDC model by examining selective pressures along lineages in a phylogeny of β-globin genes in eutherian mammals. We did not find significant evidence for an increase in the $\omega$ ratio following major duplication events in this family. However, one exception to this pattern was the duplication of γ-globin in simian primates, after which a few sites were identified to be under positive selection. Overall, our results suggest that following gene duplications, paralogous copies of β-globin genes evolved under a nonepisodic process of functional divergence.

**Key words:** β-Globin gene family — Gene duplication — Gene conversion — Positive selection — Codon substitution models

## Introduction

Gene duplication is one of the most important mechanisms in the evolution of gene diversity, presumably because it is easier to achieve new functions by modifying preexisting genetic systems than by generating them *de novo* (Ohno 1970; Go 1981; Gilbert 1978; Hughes 1994). After gene duplication, gene copies can explore three possible routes: (1) one paralog can lose the original function by the accumulation of deleterious mutations (nonfunctionalization); (2) one paralog may gain a new function under positive selection for advantageous mutations (neofunctionalization); or (3) original functions are partitioned among the two paralogs (subfunctionalization) (Force et al. 1999, Lynch and Conery 2000). Since the vast majority of mutations are deleterious,

*Correspondence to:* Joseph P. Bielawski, Department of Biology, Dalhousie University, Halifax, Nova Scotia B3H 4J1, Canada; *email:* j.bielawski@dal.ca

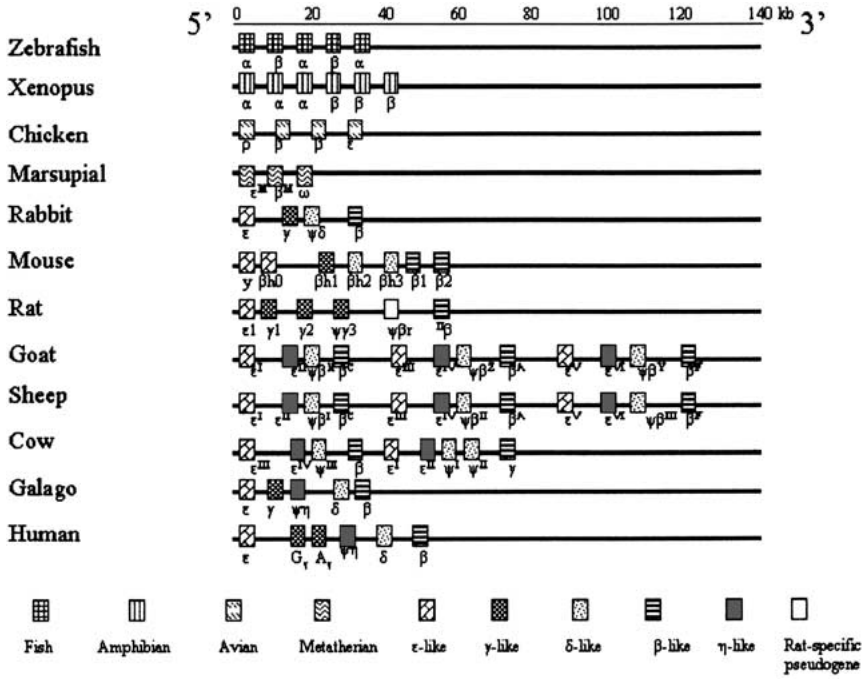


**Fig. 1.** β-Globin gene linkage in different vertebrates (Cooper et al. 1996; Garner and Lingrel 1989; Konkel et al. 1979; Lacy et al. 1979; Kretschmer et al. 1981; Lingrel et al. 1983; Satoh et al. 1999; Schon et al. 1981; Shapiro et al. 1983; Townes et al. 1984; Schimenti and Duncan 1985b). Orientation is variable in fish and amphibian clusters (Gillemans et al. 2002; Hosbach et al. 1983).

the usual fate for one duplicate is nonfunctionalization (Ohno 1970).

Until recently, the traditional view was that positive selection is the usual mechanism whereby a duplicated gene avoids nonfunctionalization. Positive selection leads to fixation of mutations conferring a new or modified function in one of the copies shortly after the duplication event (Ohta 1988). Interestingly, there is often an acceleration of the nonsynonymous substitution rate following gene duplication (Li 1997; Lynch and Conery 2000). After a new function has evolved, however, amino acid evolution is expected to be dominated by purifying selection and the rate of nonsynonymous substitution should decrease (Ohta 1993). Under an alternative model, the duplication–degeneration–complementation (DDC) model, duplicates are maintained because the functions are partitioned between paralogs which complement each other. This occurs as a result of degenerative mutations, which accumulate differentially in functional domains or in regulatory regions of genes (Force et al. 1999). It is the partitioning of subfunctions, rather than the acquisition of new functions, that preserves the duplicates; hence, this model does not explicitly require a role for positive Darwinian selection, although differential purifying selection may be at work. Recent empirical studies suggest that this model is applicable to some gene families (e.g., Gerhardt and Kirchner 1997; DiLeone et al. 1998; Force et al. 1999). However, the relative importance of these models remains a matter of controversy (Mazet and Shimeld 2002).

The eutherian mammal β-globin family comprises five functional genes (β-, δ-, ε-, G$\gamma$-, and A$\gamma$- globin) and one pseudogene ($\psi\beta$) typically arranged in a specific linkage order (Fig. 1). All the functional β-globin genes encode the β chain of hemoglobin, a tetramer composed of two α and two β chains, which binds oxygen noncovalently (Perutz 1983). The β-globin family constitutes a classic example of molecular evolution by gene duplication. Globin paralogs have explored diverse evolutionary pathways, with some functional genes retaining their original function (i.e., encode the β chain of adult hemoglobins) (Bunn 1981), others having become nonfunctional (Lacy and Maniatis 1980; Cleary et al. 1981; Li et al. 1981; Martin et al. 1983; Goodman et al. 1984), and yet others having changed their function and time of expression (Farace et al. 1984; Hutchinson et al. 1984; Fitch et al. 1991; Meireles et al. 1995; Johnson et al. 1996). Expression is partitioned among developmental stages, with β- and δ-globins expressed entirely in adults, ε-globin expressed solely in the embryo, and γ-globin expressed in the embryo in some placental mammals and in the fetus in simian primates (Hardison et al. 1997). Hence in the evolution of this gene family, both the partitioning of expression and the divergence of the proteins are important factors.

In this paper, we investigate the pattern and process of evolution by gene duplication in the β-globin family. We assembled a dataset of 72 DNA sequences that include mammals, amphibians, fish, and birds. We inferred a phylogeny for the β-globin family and identified duplication events and gene conversions, some of which have not been reported previously. We were interested in testing some predictions of the neofunctionalization and subfunctio-

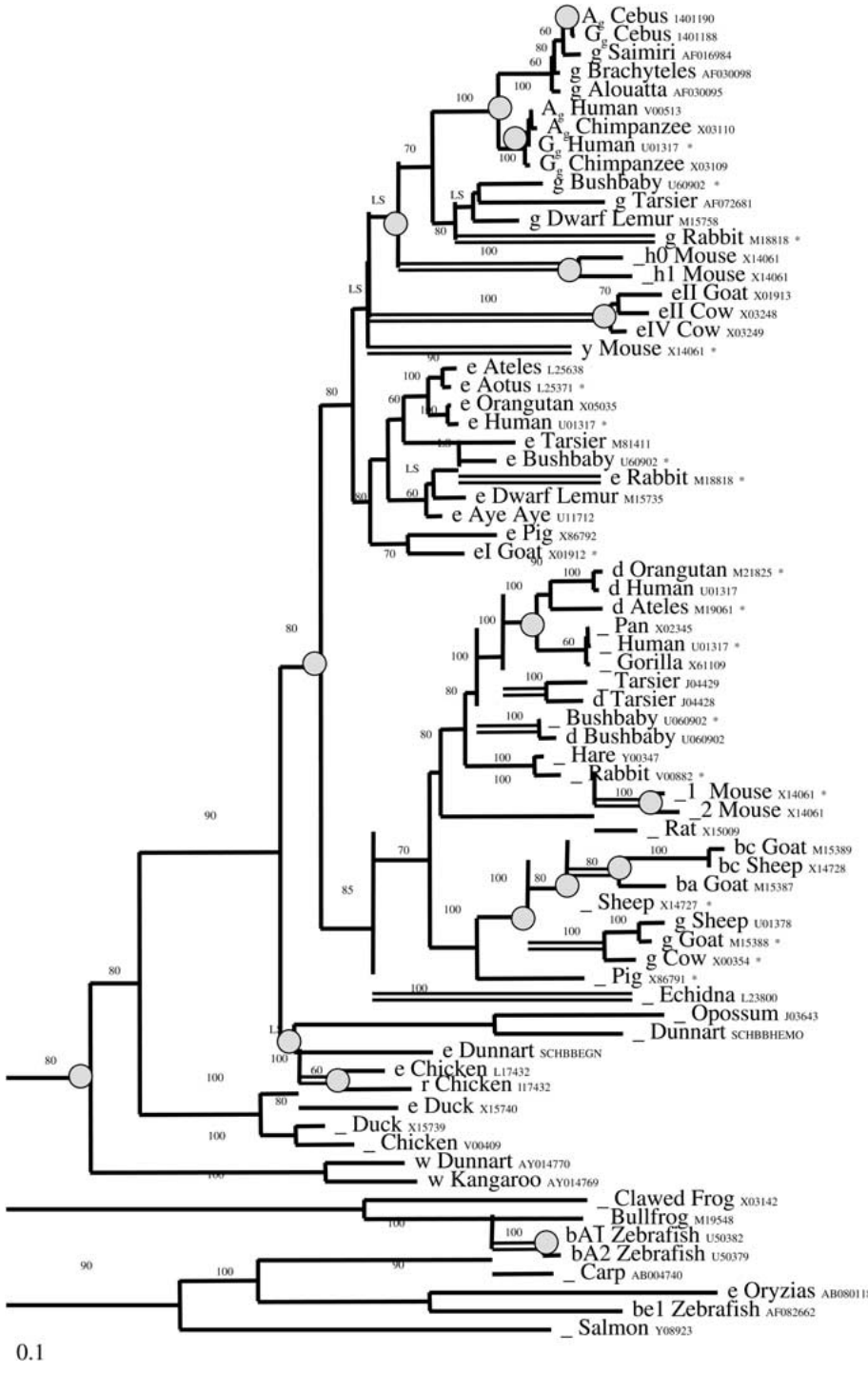


**Fig. 2.** Maximum likelihood tree of the β-globin gene family. GenBank accession numbers are provided after species names. Support values for nodes are bootstrap proportions. Duplication events are marked by circles. Asterisks show species used in the eutherian mammal dataset (see Fig. 3). Double-line branches indicate taxa misplacement relative to the expected species topology. LS—Low support ( < 40%).

nalization models in eutherian mammals, as their phylogenetic relationships are well studied and their sequences are not overly divergent. We wanted to contrast Ohta's model with the DDC model, as they represent extremes in the debate on gene family evolution, although other models have been proposed (Ohno 1970, 1984; Patthy 1985; Gilbert 1978; Krakauer and Nowak 1999; Clark 1999; Kondrashov et al. 2002). Specifically we tested for (i) a significant increase in the rate of nonsynonymous substitution following gene duplication events, a consequence of neofunctionalization predicted by Ohta (1988); and (ii) significant differences in selective constraints among paralogs. Even though the DDC model is concerned with the evolution of regulatory regions we hypothesize that, if subfunctionalization occurs in the protein coding sequences, as well as in the regulatory sequences, selective pressure should differ between paralogs. We measured selective pressure by using the nonsynonymous/synonymous substitution rate ratio, as implemented in codon models of sequence evolution (Nielsen and Yang 1998; Yang et al. 2000). An

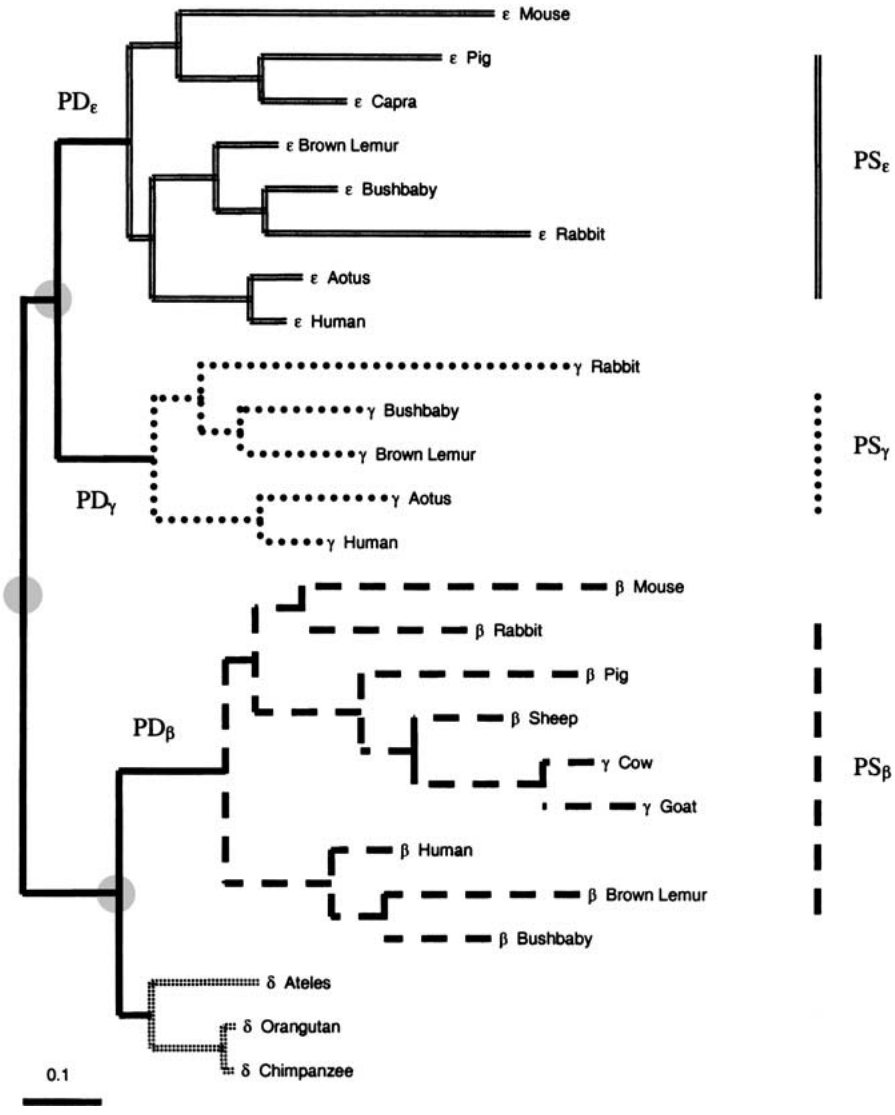


**Fig. 3.** Maximum likelihood tree of the β-, δ-, ε-, and γ-globin genes from eutherian mammals. Gray circles indicate gene duplication events. Branches in the tree are partitioned into postduplication (PD; those that immediately postdate the gray circles) and postspeciation (PS; those that postdate species-divergences) branches. The tree is rooted at the proto-ε- and proto-β-globin duplication event to help interpret classification of branches. All analyses were conducted using unrooted trees. Labels indicate the classification of PD (—) and PS branches (= = =, - - -, = = =, and ···). This is a general representation; different tests focused on different PD and PS branches according to each case described in the text.

$\omega < 1$ indicates purifying selection, $\omega = 1$ is consistent with neutral evolution, and $\omega > 1$ indicates positive Darwinian selection (Yang and Bielawski 2000).

## Materials and Methods

### *Sequence Data*

For the phylogenetic study of the β-globin gene family, 72 sequences from various vertebrates including fish, amphibians, bird, and mammals were obtained from GenBank. The nomenclature of β-globin genes is rather chaotic. To avoid confusion, we have included species names and GenBank accession numbers next to each sequence in Fig. 2. We used the bony fish clade as outgroup. Sequences were aligned using Clustal X (Thompson et al. 1997), followed by manual adjustments. Alignment gaps were removed.

### *Phylogenetic Analysis*

We constructed phylogenies for the β-globin genes to understand the relative order of duplication and speciation events and to identify gene conversions. Trees were estimated from the nucleotide sequences using maximum parsimony, maximum likelihood, and Bayesian analysis. Relative support for internal branches was measured using bootstrap analysis with PAUP* (Swofford 1998). We performed the SH (Shimodaira and Hasegawa 1999), KH (Kishino and Hasegawa 1989), and RELL (Kishino and Hasegawa 1989) tests to compare the inferred gene tree with an alternative topology derived from the expected species relationships. We compared two trees each time, the tree in Fig. 2 and a tree modified by relocating the misplaced taxa according to the species phylogeny. We used the programs PLATO (Crassly and Holmes 1997), Pist (Worobey 2001), GENECONV (Sawyer 1999), Reticulate (Jakobsen and Eastel 1996), and Partimatrix (Jakobsen et al. 1997) and estimated the Homoplasy Index (Maynard Smith and Smith 1998) to test for nonreciprocal recombination between paralogous genes, i.e., gene conversion. Conflict between the estimated gene tree and the species tree determined which sequences were tested for gene conversion.

### *Analysis of Selective Pressure*

To examine the selective pressure acting on genes from the β-globin family (i.e., β-, δ-, ε-, and γ-globins), we used sequences from eutherian mammals only. We analyzed a dataset comprising β-, δ-, ε-, and γ-globin genes from eutherian mammals. The

sequences are identified in the tree in Fig. 3. The dataset included the 20 sequences marked with an asterisk in Fig. 2 plus the following: brown lemur β-globin gene (M15734), ε-globin gene (M15735), γ-globin gene (M155757), chimpanzee δ-globin gene (AF339363), and *Aotus* γ-globin gene (AF016985). Primates possess two copies of γ-globin; we chose to sample the $G_\gamma$ copy because it is less likely to be affected by gene conversion, as gene conversion is almost exclusively unidirectional with $G_\gamma$ converting $A_\gamma$ (Fitch et al. 1990). β-globin genes converted δ-globin genes in some lineages (Koop et al. 1989); therefore we excluded the converted δ copies. Also excluded were some of the internally duplicated genes in the ruminant β-globin cluster (goat εIII, εIV, εV, and εVI and cow εI and εIII) (Fig. 1). These sequences are very divergent due to inserted sequences (Saban and King 1994). From the mouse β-globin cluster (Fig. 1), we sampled one of the three copies of fetal globin (βh1) and one of the two adult globin genes (β1). Separate datasets also were constructed for β-, ε-, and γ-globin genes. There were too few sequences available for a separate analysis of the δ-globin gene.

## Site-Based Analyses

A statistical approach was taken to study the selective pressure on the β-globin gene family in eutherian mammals. We used several codon models of molecular evolution that allow for heterogeneous $d_N/d_S$ ratios at sites (Nielsen and Yang 1998; Yang et al. 2000). In the simplest model (M0 or one-ratio model), the $\omega$ ratio is an average over all the sites. The "neutral" model (M1) allows for conserved sites where $\omega = 0$ and completely neutral sites where $\omega = 1$. The "selection" model (M2) adds a third class to M1 at which $\omega$ can take values $> 1$. The discrete model (M3) uses an unconstrained discrete distribution with different $\omega$ ratios for $K$ different classes of sites. Model M7 (beta) assumes a beta distribution of $\omega$ over sites. Model M8 (beta and $\omega$) adds an extra class of sites to M7, thereby allowing $\omega$ values $> 1$. Likelihood ratio tests (LRTs) were conducted to test M0 (one-ratio) against M3, M1 (neutral) against M2 (selection), and M7 (beta) against M8 (beta and $\omega$). All analyses were based on the unrooted gene-tree topologies and used the codeml program in the PAML package (Yang 1997).

## Branch-Based Analyses

To study changes in selective pressure in the context of gene duplication we implemented several models that allow for variable $\omega$ ratios among branches in the tree (Yang 1998; Bielawski and Yang 2003). The null model assumed the same $\omega$ for all lineages in the tree. The "PD–PS" model assigns different $\omega$ ratios for postspeciation and postduplication branches in the tree (e.g., Fig. 3). This is based on the hypothesis that duplicated genes avoid nonfunctionalization because positive Darwinian selection promoted fixation of amino acid mutations that led to a new or modified gene function (Ohta 1988). The hypothesis predicts a burst of amino acid replacements in the branches postdating duplication events (Ohta 1983). After a new function evolves, however, amino acid evolution is expected to be dominated by purifying selection and the rate of nonsynonymous substitution should decrease (Ohta 1993). Hence there should be a higher rate of amino acid substitution along branches that immediately postdate duplication events (PD branches) compared with those branches that immediately postdate speciation events (PS branches). An LRT can be conducted to compare the one-ratio model ($\omega_{PD} = \omega_{PS}$) with the two-ratio model PD-PS ($\omega_{PD} \neq \omega_{PS}$).

Another alternative model was based on the hypothesis that duplicated genes avoid nonfunctionalization because expression patterns and/or functions are partitioned among paralogs following gene duplication (Force et al. 1999). If subfunctionalization had indeed occurred in the protein-coding sequences, sites associated with such partitioning are expected to exhibit long term differences in selection pressure. If the difference between paralogs is large, we might be able to detect paralog-specific differences in average selective constraint. We formalized this in a model called "Paralog," where an independent $\omega$ ratio is specified for each paralogous clade (e.g., $\omega_\beta \neq \omega_\gamma \neq \omega_\varepsilon$). To test for a significant difference in selective pressure among paralogs we conducted an LRT comparing the one-ratio model (e.g., $\omega_\beta = \omega_\gamma = \omega_\varepsilon$) with the three-ratio Paralog model.

## Branch-Site Analysis

The above approaches might not detect a short episode of positive Darwinian selection, such as immediately following a gene duplication event, if it occurs at just a fraction of amino acid sites. The "branch-site" models (models A and B) recently developed allow the $\omega$ ratio to vary both among lineages and among sites and permits detection of lineage-specific changes in selective pressure at specific amino acid sites (Yang and Nielsen 2002). Branch-site models A and B have four $\omega$ site classes. The first two site classes, with $\omega_0$ and $\omega_1$, are uniform across the phylogeny, whereas the other two site classes are allowed to change from $\omega_0 \rightarrow \omega_2$ and from $\omega_1 \rightarrow \omega_2$ in a pre-specified branch of interest (the "foreground" branch). Note that $\omega_2$ can take values $> 1$, thus allowing for positive selection. In branch-site model A, $\omega_0$ is fixed to 0 and $\omega_1$ is fixed to 1; hence positive selection is permitted at only the foreground branch. Model A is compared with model M1 (neutral) with degrees of freedom (df) = 2. In model B, $\omega_0$ and $\omega_1$ are free parameters; therefore some sites can evolve under positive selection across all the branches in the phylogeny, whereas other sites are permitted to take $\omega$ values $> 1$ in the foreground branch. An LRT compares model B with model M3 (discrete) with $K = 2$ site classes and df = 2. We used branch-site models A and B to test for possible adaptive evolution along lineages following gene duplications.

# Results

## *Phylogenetic Analysis*

The 72 β-globin family genes in Fig. 2 were used for phylogenetic reconstruction. The ML tree is shown in Fig. 2. Both ML and Bayesian methods resulted in similar topologies, with support values for the internal nodes shown in Fig. 2. The only case of disagreement between the two methods was in the placement of marsupial and monotreme sequences. In the Bayesian tree the echidna β-globin gene was sister to a marsupial clade (opossum and dunnart β-globins), and in turn this clade was placed sister to the eutherian β-globin clade. In the ML tree (Fig. 2), the echidna β-globin gene was sister to the eutherian β-globin clade. Clearly, placement of the monotreme and marsupial β-globins is problematic and will probably require additional sampling to resolve. Interestingly, the marsupial $\omega$-globin genes were placed outside the mammalian β-globin clade, consistent with the earlier study of Wheeler et al. (2001).

Assuming no gene conversion, we expected (i) monophyly for each set of paralogs (i.e., β-, δ-, ε-, and γ-globins) and (ii) to recover the expected species tree within each paralogous clade (Rowe 1999; O'Brien et al. 2001; Springer et al. 2003). However,

**Table 1.** Parameter estimates and likelihood scores in separate analyses of the β-, γ-, and ε-globin genes under site-specific models

| Model | Parameter estimate(s) | $\ell$ |
|---|---|---|
| M0 (one-ratio) | | |
| β | $\omega = 0.27$ | −1676.08 |
| γ | $\omega = 0.26$ | −1609.76 |
| ε | $\omega = 0.17$ | −2137.83 |
| M1 (neutral) | | |
| β | $(\omega_0 = 0), f_0 = 0.60, (\omega_1 = 1), (f_1 = 0.40)$ | −1621.00 |
| γ | $(\omega_0 = 0), f_0 = 0.57, (\omega_1 = 1), (f_1 = 0.43)$ | −1598.16 |
| ε | $(\omega_0 = 0), f_0 = 0.54, (\omega_1 = 1), (f_1 = 0.46)$ | −2145.53 |
| M2 selection | | |
| β | $(\omega_0 = 0), f_0 = 0.60, (\omega_1 = 1), f_1 = 0.36, \omega_2 = 3.58, (f_2 = 0.04)$ | −1617.42 |
| γ | $(\omega_0 = 0), f_0 = 0.52, (\omega_1 = 1), f_1 = 0.006, \omega_2 = 0.57, (f_2 = 0.47)$ | −1592.80 |
| ε | $(\omega_0 = 0), f_0 = 0.33, (\omega_1 = 1), f_1 = 0.11, \omega_2 = 0.16, (f_2 = 0.56)$ | −2100.22 |
| M3 discrete | | |
| β | $\omega_0 = 0.02,\ f_0 = 0.65, \omega_1 = 0.57, f_1 = 0.26, \omega_2 = 2.02, (f_2 = 0.09)$ | −1608.57 |
| γ | $\omega_0 = 0.001,\ f_0 = 0.52, \omega_1 = 0.42, f_1 = 0.18, \omega_2 = 0.66, (f_2 = 0.31)$ | −1592.80 |
| ε | $\omega_0 = 0.04,\ f_0 = 0.66, \omega_1 = 0.27, f_1 = 0.24, \omega_2 = 0.89, (f_2 = 0.11)$ | −2099.60 |
| M7 beta | | |
| β | $p = 0.11, q = 0.29$ | −1612.60 |
| γ | $p = 0.23, q = 0.55$ | −1593.31 |
| ε | $p = 0.34, q = 1.50$ | −2101.40 |
| M8 beta and $\omega$ | | |
| β | $p = 0.16, q = 0.061,\ f_0 = 0.93, \omega_1 = 2.19, (f_1 = 0.07)$ | −1608.76 |
| γ | $p = 0.03, q = 0.64,\ f_0 = 0.57, \omega_1 = 0.60, (f_1 = 0.43)$ | −1592.79 |
| ε | $p = 0.95, q = 7.81,\ f_0 = 0.88, \omega_1 = 0.85, (f_1 = 0.12)$ | −2099.67 |

we found some notable misplacements (double lines in Fig. 2): (i) the rabbit ε and γ sequences were sister to the primate ε and γ genes, respectively, rather than sister to rodent ε- and γ-globins; (ii) the cow εII and εIV genes and goat εII comprised a monophyletic clade sister to the γ-globins instead of being within the ε clade; (iii) the mouse ε gene (a single-copy gene traditionally called y) did not appear within the ε clade but was sister to a clade including the cow εII and εIV and the goat εII genes; (iv) tarsier and bushbaby δ-globin genes were sister to tarsier and bushbaby β-globin genes, respectively; (v) the genes traditionally labeled as γ-globins in sheep, cow, and goat were placed within the β-globin clade; (vi) chicken ε was sister to chicken ρ instead of being more closely related to duck ε-globin; and (vii) *Cebus* Gγ and Aγ were more closely related to each other than to their respective human and chimpanzee orthologs.

All misplacements were supported by high bootstrap proportions (>70%) with the exception of the *Cebus* $A_\gamma$ and $G_\gamma$, the rabbit ε, and the mouse y branches, where there was low bootstrap support. We used the SH test to compare the expected placements with the estimated topology (Fig. 2). SH tests indicated significantly greater support for five misplacements (bushbaby δ, $p < 0.0001$; tarsier δ, $p = 0.002$; cow, sheep, and goat γ, $p = 0.000$; echidna β, $p = 0.053$; and *Cebus* $A_\gamma$ and $G_\gamma$, $p = 0.000$). The remaining misplacements did not fit these data significantly better than the expected phylogenetic placements (mouse ε, $p = 0.095$; rabbit γ, $p = 0.217$; rabbit β, $p = 0.59$; cow ε, $p = 0.193$; rabbit ε, $p = 0.289$; and mouse y, goat εII, and cow εII and εIV, $p = 0.225$; chicken ρ, $p = 0.397$). Results under KH and RELL tests were the same as with SH tests (data not shown).

A potential source of conflict between the gene tree and the species tree could be gene conversion (Ohta 1980, 1990). Hence, we used the misplacements to guide our tests of gene conversion. Tests were conducted on alignments of third codon positions only, by using different software programs (Grassly and Holmes 1997; Worobey 2001; Sawyer 1999; Jakobsen and Eastel 1996; Jakobsen et al. 1997; Maynard Smith and Smith 1998). We found evidence for two gene conversion events that are not reported previously: (i) among the duplicates in the goat β-globin cluster between nucleotide 12 and nucleotide 75 (site numbering refers to the human β-globin gene; PDB file 2hhb) (PLATO z-score = 4.85) and (ii) among the mouse β-globin cluster genes between nucleotide 210 and nucleotide 235 (PLATO z-score = 3.87). Our analysis corroborated gene conversions previously suggested for tarsier and bushbaby δ-globin genes (Koop et al. 1989; Grassly and Holmes 1997) between nucleotides 45–63 and nucleotides 357–375, and in cow εII and cow εIV between nucleotide 12 and nucleotide 30, in agreement with Schimenti and Duncan (1985a). However, we found no evidence for gene conversions between mouse β genes βh0 and βh1 or between mouse βh0 and mouse y, (see Figs. 1

**Table 2.** Likelihood ratio test statistics for comparing site-specific models for the β-, γ-, and ε-globin genes

| Globin gene | 2δ | df | *p* value |
|---|---|---|---|
| M0 (one ratio) vs. M3 (discrete) | | | |
| β | 135.02 | 2 | < 0.0001 |
| γ | 33.97 | 2 | < 0.0001 |
| ε | 76.47 | 2 | < 0.0001 |
| M7 (beta) vs. M8 (beta and $\omega$) | | | |
| β | 7.68 | 2 | 0.020 |
| γ | 1.03 | 2 | 0.600 |
| ε | 3.46 | 2 | 0.177 |

and 2) reported by Hill et al. (1984), or between *Cebus* $G_\gamma$ and $A_\gamma$.

### *Variable Selective Pressure Among Sites*

In order to minimize the effect of gene conversion, we excluded the converted genes. Given that gene conversion tends to have a direction in globins, we knew for instance that δ-globins are generally converted by β-globin, and not vice versa. This prior knowledge allowed us to minimize gene conversion effects to some extent, although eliminating conversion altogether is impossible, as numerous events have characterized the evolution of β-globin genes. We also compared tests of variable selective pressure using different datasets, both with and without misplaced sequences. We obtained similar results for the different datasets, confirming that gene conversion, although probably present, did not greatly affect our results.

We expected selective pressure to vary among sites and among the genes of the β-globin family. We used codon models to detect among-site variability in selective pressure in the β-, ε-, and γ-globin genes. From the one-ratio model (M0) we found that the $\omega$ ratio averaged over all sites is 0.27, 0.26, and 0.17 for β-, γ-, and ε-globin genes, respectively, when the three genes were analyzed as separate data sets (Table 1). The estimates suggested that, on average, the ε-globin is more constrained than the γ and β. However, an $\omega$ ratio averaged over sites is a crude measure of selective pressure. Therefore we used models that allow selective pressure to vary among sites. The discrete model (M3), with three site classes, revealed considerable variation in selective pressure among sites (Table 1). For example, β-globin had 65% of sites under strong purifying selection ($\omega = 0.02$), 26% of sites were less constrained ($\omega = 0.57$), and 9% of sites were under positive selection ($\omega = 2.02$) (Table 1). Interestingly, neither γ nor ε showed evidence of sites evolving under positive selection (Table 1). Evolution of the majority of sites in all three paralogs was dominated by strong purifying selection, with 65% of sites in β, 52% of sites in γ, and 66% of sites in ε evolving with $\omega < 0.05$.

We tested for variable selective pressure among sites by conducting an LRT comparing the one-ratio model (M0) with the discrete model (M3); results were highly significant for all three genes (Table 2). In general, β-globin was the most variable gene in the family, having an additional class of sites evolving under positive Darwinian selection.

We were interested in identifying regions that are conserved in all three genes in the cluster, which presumably indicate functionally important residues in the protein product. For β-, γ-, and ε-globin genes separately, we plotted the approximate posterior mean of the $\omega$ ratio at each site (Fig. 4). Four regions are highly conserved in all three genes: (i) residues 28 to 38, located in helices B and C; (ii) residues 57 to 63, located in helix E; (iii) residues 79 to 81, located in helix F; and (iv) residues 87 to 101, located in helices F and G. When mapped onto the three-dimensional structure of the β-chain in hemoglobin, we found that sites within these four constrained regions were located mostly on the inner hydrophobic core of the subunit, the area around the heme pocket and the $\alpha_1\beta_1$ interface. In all cases we used the human β-globin chain structure (PDB: 2hhb) as reference to map sites into the three-dimensional structure. Residues 28 to 38 are distributed among the hydrophobic core, the $\alpha_1\beta_1$ interface between monomers, and part of the heme pocket.

The site-specific codon models were also used to identify positive selection at sites, indicated by $\omega > 1$. The selection model (M2), the discrete model (M3), and the beta and $\omega$ model (M8) allow $\omega > 1$ at a fraction of sites (Yang et al. 2000). All three models were generally consistent in suggesting a small fraction of sites (4 to 9%) evolving under positive Darwinian selection ($\omega$ between 2.02 and 3.58) in the β-globin gene (Table 1). We tested significance of sites evolving under positive selection by an LRT comparing M7, which does not allow for such sites, with M8, which has an additional parameter that can accommodate sites with $\omega > 1$. The test is highly significant for the β-globin gene (Table 2).

### *Variable Selective Pressures Among Branches*

A burst of nonsymoymous evolution is often observed following gene duplication, and positive Darwinian selection is frequently invoked to explain this pattern. An LRT was used to test whether selective pressure is significantly different between postduplication (PD) and postspeciation (PS) branches in the β-globin gene phylogeny; i.e.,

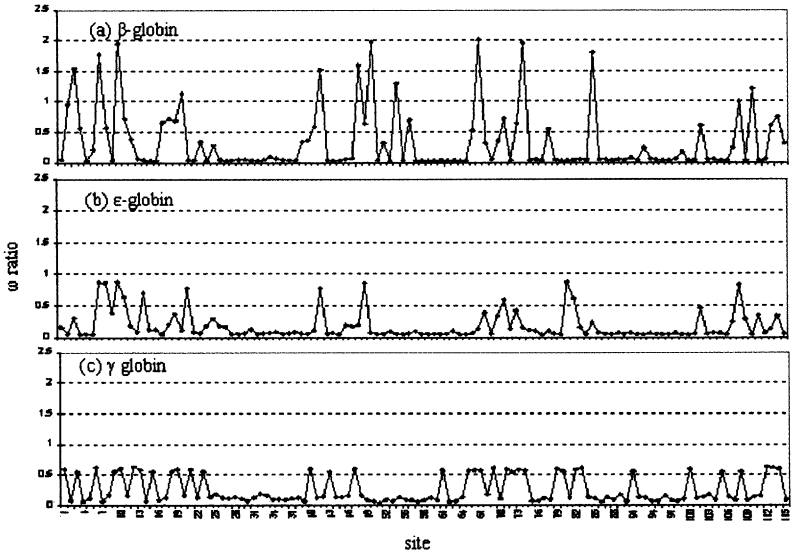


**Fig. 4.** Approximate posterior mean of the $\omega$ ratio for each site calculated under model M3 (discrete) for the (**a**) β-globin, (**b**) ε-globin, (**c**) γ-globin genes.

($\omega_{\beta(PD)} = \omega_{\varepsilon(PD)} = \omega_{\gamma(PD)}$) ≠ ($\omega_{\beta(PS)} = \omega_{\varepsilon(PS)} = \omega_{\gamma(PS)}$). The LRT was not significant (Table 3), suggesting no difference between PD branches and PS branches. Furthermore, estimates of $\omega$ suggested strong purifying selection in both the PD and the PS branches ($\omega_{(PD)} = 0.34$, $\omega_{(PS)} = 0.23$). We also fitted a more general four-ratio model in which the branches postdating the three duplication events in the phylogeny were assigned independent $\omega$ ratios ($\omega_{\beta(PD)}$, $\omega_{\varepsilon(PD)}$, $\omega_{\gamma(PD)}$ $\omega_{(PS)}$) and compared it with the one ratio model. Again, the LRT was not significant (Table 3), and none of the parameter estimates suggested positive Darwinian selection: $\omega_{\beta(PD)} = 0.41$, $\omega_{\varepsilon(PD)} = 0.22$, $\omega_{\gamma(PD)} = 0.08$, $\omega_{(PS)} = 0.24$. Note that in both PD–PS models tested, $d_N$ values averaged 0.024 and $d_S$ values averaged 0.101.

The above analysis averages rates over all sites in the gene and may lack power in detecting positive selection. Thus we also used branch-site models A and B (Yang and Nielsen 2002) to detect positive selection at a subset of sites along specific lineages. We tested each postduplication branch in the β-globin phylogeny as defined in Fig. 3. We found no evidence for positive selection at branches immediately following the duplication event that gave rise to proto-β and proto-ε, or after the duplication that created ε and γ (data not shown). The duplication event that resulted in $A_\gamma$- and $G_\gamma$-globins is hypothesized to have occurred along the branch leading to the simian primates (Slightom et al. 1985), but cannot be resolved on a gene tree because of frequent gene conversion events. When we used a specific dataset comprising ε- and γ-globins (Fig. 5) and tested the branch where the duplication is thought to have occurred we found an increase in nonsynonymous substitutions (M1 vs MA, $2\delta = 37.16$, df = 2, $p < 0.0001$; M3 vs MB, $2\delta = 18.66$, df = 2, $p < 0.0001$). The $d_N$ value was 0.021 and the $d_S$ value was 0.039, as measured as an average over all branches of the ε- and γ-globin tree. Parameter estimates under models A and B suggested positive selection at a few sites along the branch leading to simian primates ($\omega_{2(MA)} = 10.0$, $\omega_{2(MB)} = 4.58$ in Table 4). Interestingly, this branch is also thought to coincide with the recruitment of γ-globins for fetal expression (double line in Fig. 5).

Globin genes are expressed at different developmental stages, so each gene might be subject to different selective pressures. To test for paralog-specific differences in selective pressure, we fitted the "Paralog" model, where β-, γ-, and ε-globins have independent selective pressures (i.e., $\omega_\beta \neq \omega_\varepsilon \neq \omega_\gamma$). This model fits the data significantly better than the one-ratio model, with parameter estimates $\omega_\beta = 0.29$, $\omega_\varepsilon = 0.16$, $\omega_\gamma = 0.23$ (Table 3). The average $d_N$ value was 0.024 and the average $d_S$ value was 0.103. Those estimates are consistent with the $\omega$ estimates from the separate analysis of the paralogs, with ε-globin more constrained than γ- and β-globins (Table 1). Fitting additional models with two of the three ratios ($\omega_\beta$, $\omega_\varepsilon$, $\omega_\gamma$) forced to be identical suggests that $\omega_\gamma$ is different from $\omega_\beta$ and $\omega_\varepsilon$, while $\omega_\beta$ and $\omega_\varepsilon$ are not significantly different (Table 3).

## Discussion

Gene conversion plays an important role in the evolution of multigene families, as it brings about the exchange of genetic material between related sequences (Schimenti 1994; Posada et al. 2002). It is a frequent mechanism of evolutionary change in globins and can act both to homogenize genes through concerted evolution (e.g., $A_\gamma$ and $G_\gamma$ in simian primates) and to introduce novelty among

**Table 3.** Maximum likelihood estimates of $\omega$ ratios under branch-specific models and likelihood ratio test statistics when the model is compared with the null model M0 (one ratio)

| Alternative model | Parameter estimate(s) | 2δ | df | *p* value |
|---|---|---|---|---|
| | *Postduplication and postspeciation (PD–PS) models* | | | |
| 2 ratios | $\omega_{(PD)} = 0.34$, $\omega_{(PS)} = 0.23$ | 1.10 | 1 | 0.29 |
| 4 ratios | $\omega_{\beta(PD)} = 0.41$, $\omega_{\varepsilon(PD)} = 0.22$, $\omega_{\gamma(PD)} = 0.08$, $\omega_{(PS)} = 0.24$ | 1.10 | 3 | 0.78 |
| | *Paralog models* | | | |
| 3 ratios | $\omega_\beta = 0.29$, $\omega_\gamma = 0.23$, $\omega_\varepsilon = 0.16$ | 11.66 | 2 | 0.003 |
| 2 ratios | $\omega_\beta = 0.28$, $\omega_\varepsilon = \omega_\gamma = 0.19$ | 7.89 | 1 | 0.005 |
| 2 ratios | $\omega_\varepsilon = 0.16$, $\omega_\gamma = \omega_\beta = 0.27$ | 9.88 | 1 | 0.002 |
| 2 ratios | $\omega_\gamma = 0.23$, $\omega_\beta = \omega_\varepsilon = 0.23$ | 0.045 | 1 | 0.832 |

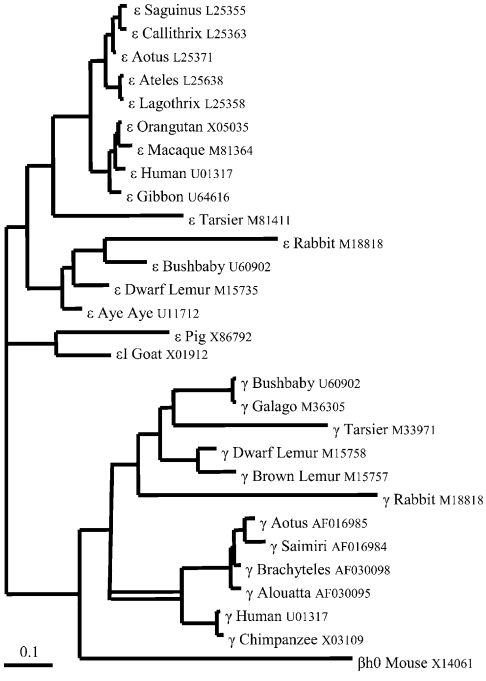


**Fig. 5.** Maximum likelihood tree of the ε- and γ-globin genes from eutherian mammals. The double line corresponds to the branch where the $G_\gamma$ and $A_\gamma$ split is hypothesized to have occurred, in the ancestor of simian primates.

homologous genes (e.g., cow εII and εIV). Gene conversion is known to affect gene phylogenies (Slatkin and Maddison 1989; Hudson et al. 1992; Maddison 2000). Given the general importance of the mechanism, its pervasiveness, and its effects on phylogeny reconstruction, it is essential to test for gene conversion when topological discrepancies arise in a gene family tree (Drouin 2002). By using statistical methods, we found evidence of two unreported gene conversion events in β-globins—(i) among duplicates in the goat β-globin cluster and (ii) among duplicates in the mouse β-globin cluster—and we confirmed many previously suggested cases. Furthermore, we suggest that the majority of misplacements in our gene tree are the result of gene conversion events.

The traditional model of evolution by gene duplication predicts an increase in nonsynonymous substitution rate immediately after genes duplicate. It is a matter of debate whether this rate increase is due to a relaxation of selective pressure or to the action of positive selection for advantageous mutations (for a review see Massingham et al. 2001; Mazet and Shimeld 2002). Previous studies of the β-globin family supported the positive selection model, with this mode of evolution being suggested following the split of myoglobin and hemoglobin (Goodman 1981) and following the divergence of α- and β-hemoglobins (Czelusniak et al. 1982). Accelerated amino acid evolution also occurred after the *en bloc* duplications within the ruminant artiodactyl lineage (Li and Gojobori 1983). In contrast to these examples, we found no significant evidence for a burst of nonsynonymous evolution in the branches postdating the initial duplications of the proto-β and proto-ε genes, or after the duplication giving rise to the β and δ or to the ε and γ clades, which correspond to the major duplication events within the gene family. We also tested for an increase in nonsynonymous substitutions at particular sites along the postduplication branches using branch-site models but failed to detect an evolutionary burst. Interestingly, a recent study of the early stages of evolution of duplicate genes within the human genome found that most genes exhibit an accelerated rate of nonsynonymous evolution in one duplicate (Zhang et al. 2003). Our data suggest that the early divergences within the β-globin family of genes do not fit this pattern, as we found no such changes in the evolutionary rate during the early stages of divergence.

There was one exception to the general pattern described above. In the lineage of stem-simians, which represents the transition from embryonic to fetal expression of γ-globins (Tagle et al. 1988; Fitch et al. 1991), we detected an acceleration in nonsynonymous substitution rates and identified positively selected sites. Whereas previously used methods employed raw counts of synonymous and nonsynony-

**Table 4.** Parameter estimates and log-likelihood scores for the γ-globin gene under different sites and branch-site models

| Model | $p$ | Parameter estimates | Positive selection | $\ell$ |
|---|---|---|---|---|
| | | *Site-specific models* | | |
| M1 (neutral) | 1 | $(\omega_0 = 0.00), f_0 = 0.47\ (\omega_1 = 1.00), (f_1 = 0.53)$ | No | −3170.38 |
| M3 (discrete)($K = 2$) | 3 | $\omega_0 = 0.06,\ \ f_0 = 0.74\ \ \omega_1 = 0.63, (f_1 = 0.26)$ | No | −3094.02 |
| | | *Branch-site models* | | |
| Model A | 3 | $(\omega_0 = 0),\ \ \ f_0 = 0.45\ (\omega_1 = 1), f_1 = 0.44\ \omega_2 = 10, (f_{2+3} = 0.1)$ | Yes | −3151.80 |
| Model B | 5 | $\omega_0 = 0.05, f_0 = 0.60\ \omega_1 = 0.63,\ f_1 = 0.22\ \omega_2 = 4.58, (f_{2+3} = 0.18)$ | Yes | −3084.69 |

*Note*. $p$ is the number of parameters in the $\omega$ ratio distribution. The foreground branch in the branch-site models is the branch leading to simian primates.

mous substitutions, and were thus unable to determine the source of amino acid evolution acceleration, the branch-site models indicated that nonsynonymous rate acceleration in the lineage of stem-simian γ-globins was caused by positive Darwinian selection.

Although it is possible that undetected gene conversion affects our tests for variable $d_N/d_S$ rate ratios among branches, we believe that our results are not overly influenced by it. Our estimates are based on the comparison of silent and replacement changes and both are similarly affected by gene conversion events. Furthermore, a recent simulation study (Anisimova et al. 2003) showed that LRTs are robust to low or moderate levels of recombination, such as those we might not have been able to detect. It could also be that greater similarity among sequences reduced the power of our tests to detect an increase in $d_N/d_S$ rate ratios following gene duplication. However, we note that the tests were powerful enough in the case of the simian γ-globin amino acid replacement acceleration. Furthermore, if adaptive evolution occurs by a single or a small number of substitutions, it may not be detected by methods based on $d_N/d_S$ ratios (Bielawski and Yang 2003). It is known that large phenotypic changes in globins can be achieved by only one or a few amino acid changes (Perutz 1983). A good example of the latter is provided by the deletion of the NA1 valine residue from the protein chain encoded by γ-globin in some artiodactyls, which increases the oxygen affinity of the hemoglobin monomer (Poyart et al. 1992). Hence, in cases where we did not detect positive selection or even an increase in amino acid replacement rates, our findings do not exclude the possibility of neofunctionalization in β-globin genes by a few adaptive substitutions with large phenotypic effects.

The DDC model of gene copy preservation does not require a burst of nonsynonymous substitutions and assumes that purifying selection continues to act on both gene copies following duplication (Force et al. 1999; Zhang 2003). Nonetheless, if subfunctions are partitioned among the functional domains of the encoded protein, a potential outcome of the DDC model is heterogeneity in purifying selection among the gene copies. Dermitzakis and Clark (2001) proposed that identification of heterogeneity in pattern of amino acid substitution between different domains of the proteins encoded by paralogous genes could lead to the discovery of genes under subfunctionalization. While the DDC model has traditionally centered on regulatory sequences, we extend the possibility of finding subfunctionalization to protein-coding sequences by identifying heterogeneous selection pressure among paralogs. In the case of mammalian β-globins, genes are linked in a specific arrangement which, in most species, is known to be related to the order of expression of the genes (Hardison 1998). If the arrangement of β-globin genes in the cluster corresponds to a domain-like partition of function, each domain of expression could be subject to different selective pressures. Hence, our results are in agreement with a subfunctionalization model, as we found that each paralogous clade (i.e., domain of expression) is subject to significantly different selective constraints. Our findings suggest a long-term process of divergence during which each paralog has been subject to different constraints by purifying selection, presumably related to differences in expression regulation. As described earlier, our findings do not exclude the possibility of brief episodes of increased amino acid replacement, in which case, other models (e.g., Ohta 1988) may still be relevant to the evolution of β-globins.

The β-globin gene is the only gene with sites predicted to be under positive selection in placental mammals. We identified 12 sites under Darwinian selection, consistent with the earlier study of Yang et al. (2001). These sites are located mostly at the exterior of the protein chain, with two sites located at the $\alpha_1\beta_1$ interface between the α and the β subunit of hemoglobin (116H and 111A). As a prelude to a more detailed analysis, we tested for positive selection in the α-globin genes currently available in GenBank and found at least one positively selected site (115A) located at the $\alpha_1\beta_1$ interface. Our results raise the interesting possibility of long-term coevolution of

some alpha and beta protein chain residues located in the $\alpha_1\beta_1$ interface. A more detailed study is necessary to rigorously examine this hypothesis.

Much is now known about what makes the globin fold a robust structure (Perutz et al. 1960; Bashford et al. 1987; Murzin and Finkelstein 1988; Brenner et al. 1997). Proteins whose secondary structures are mainly alpha helices, such as β-globin chains, are flexible and can easily accommodate many residues or prosthetic groups without disrupting tertiary or quaternary structural arrangements (Chothia et al. 1977; Efimov 1979). β-Globins share the canonical features of the globin fold and have maintained a robust structure despite 200 million years of evolutionary divergence (Efstratiadis et al. 1980; Czelusniak et al. 1982). Arguably, the most important feature that explains the preservation of the globin fold is the clear conservation of hydrophobic residues at buried positions in globin proteins (Lesk and Chothia 1980). In our study we identified regions conserved in all three β-globin genes, located in the interior or hydrophobic part of the subunit. Presumably, these conserved sites are involved in the maintenance of the secondary structures which in turn stabilize the tertiary and quaternary structures of hemoglobin. Furthermore, we found that some of the conserved sites are also part of empty concavities of the protein surface accessible to solvent (Liang et al. 1998). Concavities are particularly important, as they are often associated with binding and catalytic activity (Liang and Dill 2001). For example, of the 23 sites which participate in interactions with the heme group, 15 correspond to the conserved sites in our study, with 3 involved in hydrogen bonding. With the exception of site 38Thr, all sites that participate in interactions with the heme ligand have hydrophobic–hydrophobic contacts, which stabilize the structure. Hence, during the long evolutionary history of the genes encoding the β-globin chain of hemoglobin, these functionally and structurally important sites have been preserved, while at the same time a fraction of residues has been the target of divergent fine-tuning of the protein function.

Gene family evolution reflects a balance between homogenization by unequal crossing-over and gene conversion, and diversification by mutation (Ohta 2000). Both drift and selection play an important role in the evolutionary fate of duplicated genes, but only positive selection can account for the evolution of new functions (Ohta 1987). The dynamics of these forces are complicated (Ohta 2000), and our analysis of the β-globin family of genes illustrates this complexity. Gene conversion is clearly a frequent force for homogenization of some closely related members of this family (e.g., $A_\gamma$- and $G_\gamma$-globins). As expected, gene conversion is less important to the evolution of the more divergent members, as it is prohibited when sequence divergence is too high (Ohta 2000). In addition to the partitioning of β-globin paralogs into domains of expression, this gene family exhibits divergence both by positive Darwinian selection (β- and γ-globins) and by differential patterns of purifying selection pressure (γ-and ε-globins). While more tests are clearly necessary to fully discriminate between the DDC and Ohta models, we suggest that comparison between synonymous and nonsynonymous substitution rates provides a useful tool in studying relative roles of different evolutionary forces during the evolution of a gene family.

*Acknowledgments.* G.A. was supported by a grant from the Mexican Council for Science and Technology (CONACYT). J.P.B. and Z.Y. were supported by grant 31/G14969 from the Biotechnology and Biological Sciences Research Council (BBSRC, UK).

## References

Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164:1229–1236

Bashford D, Chothia C, Lesk AM (1987) Determinants of a protein fold. Unique features of the globin amino acid sequences. J Mol Biol 196:199–216

Bielawski JP, Yang Z (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics 3:201–212

Brenner SE, Chothia C, Hubbard TJ (1997) Population statistics of protein structures: Lessons from structural classifications. Curr Opin Struct Biol 7:369–376

Bunn HF (1981) Evolution of mammalian hemoglobin function. Blood 58:189–197

Chothia C, Levitt M, Richardson D (1977) Structure of proteins: packing of alpha helices and pleated sheets. Proc Natl Acad Sci USA 74:4130–4134

Clark AG (1994) Invasion and maintenance of a gene duplication. Proc Natl Acad Sci USA 91:2950–2954

Cleary ML, Schon EA, Lingrel JB (1981) Two related pseudogenes are the result of a gene duplication in the goat beta-globin locus. Cell 26:181–190

Cooper SJ, Murphy R, Dolman G, Hussey D, Hope RM (1996) A molecular and evolutionary study of the beta-globin gene family of the Australian marsupial Sminthopsis crassicaudata. Mol Biol Evol 13:1012–1022

Czelusniak J, Goodman M, Hewett-Emmett D, Weiss ML, Venta PJ, Tashian RE (1982) Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes. Nature 298:297–300

Dermitzakis ET, Clark AG (2001) Differential selection after duplication in mammalian developmental genes. Mol Biol Evol 18:557–562

DiLeone RJ, Russell LB, Kingsley DM (1998) An extensive 3′ regulatory region controls expression of Bmp5 in specific anatomical structures of the mouse embryo. Genetics 148:401–408

Drouin G (2002) Testing claims of gene conversion between multigene family members: Examples from echinoderm actin genes. J Mol Evol 54:138–139

Efimov AV (1979) Packing of alpha-helices in globular proteins: layer-structure of globin hydrophobic cores. J Mol Biol 134:23–40

Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, et al. (1980) The structure and evolution of the human beta-globin gene family. Cell 21:653–668

Farace MG, Brown BA, Raschella G, Alexander J, Gambari R, Fantoni A, Hardies SC, Hutchison CA III, Edgell MH (1984) The mouse beta h1 gene codes for the z chain of embryonic hemoglobin. J Biol Chem 259:7123–7128

Fitch DH, Mainone C, Goodman M, Slightom JL (1990) Molecular history of gene conversions in the primate fetal gamma-globin genes Nucleotide sequences from the common gibbon, Hylobates lar. J Biol Chem 265:781–793

Fitch DH, Bailey WJ, Tagle DA, Goodman M, Sieu L, Slightom JL (1991) Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. Proc Natl Acad Sci USA 88:7396–7400

Force A, Lynch M, Pickett FB, Amores A, Van YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545

Garner KJ, Lingrel JB (1989) A comparison of the beta A-and beta B-globin gene clusters of sheep. J Mol Evol 28:175–184

Gerhart J, Kirchner M (1997) Cells, embryos, and evolution. Blackwell Science, Cambridge, MA

Gilbert W (1978) Why genes in pieces? Nature 271:501

Gillemans N, McMorrow T, Tewari R, Wai AW, Burgtorf C, Drabek D, et al. (2002) A functional and comparative analysis of globin loci in pufferfish and man. Blood 101:2842–2849

Go M (1981) Correlation of DNA exonic regions with protein structural units in haemoglobin. Nature 291:90–92

Goodman M (1981) Globin evolution was apparently very rapid in early vertebrates: A reasonable case against the rate-constancy hypothesis. J Mol Evol 17:114–120

Goodman M, Koop BF, Czelusniak J, Weiss ML (1984) The eta-globin gene.Its long evolutionary history in the beta-globin gene family of mammals. J Mol Biol 180:803–823

Grassly NC, Holmes EC (1997) A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol Biol Evol 14:239–247

Hardison R (1998) Hemoglobins from bacteria to man: Evolution of different patterns of gene expression. J Exp Biol 201(8):1099–1117

Hardison R, Slightom JL, Gumucio DL, Goodman M, Stojanovic N, Miller W (1997) Locus control regions of mammalian beta-globin gene clusters: Combining phylogenetic analyses and experimental results to gain functional insights. Gene 205:73–94

Hill A, Hardies SC, Phillips SJ, Davis MG, Hutchison CA III, Edgell MH (1984) Two mouse early embryonic beta-globin gene sequences. Evolution of the nonadult beta-globins. J Biol Chem 259:3739–3747

Hosbach HA, Wyler T, Weber R (1983) The Xenopus laevis globin gene family: Chromosomal arrangement and gene structure. Cell 32:45–53

Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. Genetics 132:583–589

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc R Soc Lond B Biol Sci 256:119–124

Hutchison CA III, Hardies SC, Padgett RW, Weaver S, Edgell MH (1984) The mouse globin pseudogene beta h3 is descended from a premammalian delta-globin gene. J Biol Chem 259:12881–12889

Jakobsen IB, Easteal S (1996) A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput Appl Biosci 12:291–295

Jakobsen IB, Wilson SR, Easteal S (1997) The partition matrix: exploring variable phylogenetic signals along nucleotide sequence alignments. Mol Biol Evol 14:474–484

Johnson RM, Buck S, Chiu C, et al. (1996) Fetal globin expression in New World monkeys. J Biol Chem 271:14684–l4691

Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol 29:170–179

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. Genome Biol 3:Research 0008

Konkel DA, Maizel JV Jr, Leder P (1979) The evolution and sequence comparison of two recently diverged mouse chromosomal beta-globin genes. Cell 18:865–873

Koop BF, Siemieniak D, Slightom JL, Goodman M, Dunbar J, Wright PC, Simons EL (1989) Tarsius delta-and beta-globin genes: Conversions, evolution, and systematic implications. J Biol Chem 264:68–79

Krakauer DC, Nowak MA (1999) Evolutionary preservation of redundant duplicated genes. Semin Cell Dev Biol 10:555–559

Kretschmer PJ, Coon HC, Davis A, Harrison M, Nienhuis AW (1981) Hemoglobin switching in sheep: Isolation of the fetal gamma-globin gene and demonstration that the fetal gamma- and adult beta A-globin genes lie within eight kilobase segments of homologous DNA. J Biol Chem 256:1975–1982

Lacy E, Maniatis T (1980) The nucleotide sequence of a rabbit beta-globin pseudogene. Cell 21:545–553

Lacy E, Hardison RC, Quon D, Maniatis T (1979) The linkage arrangement of four rabbit beta-like globin genes. Cell 18:1273–1283

Lesk AM, Chothia C (1980) How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. J Mol Biol 136:225–270

Li WH (1997) Molecular evolution, 2nd ed. Sinauer Associates, Sunderland, MA

Li WH, Gojobori T (1983) Rapid evolution of goat and sheep globin genes following gene duplication. Mol Biol Evol 1:94–108

Li WH, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237–239

Liang J, Dill KA (2001) Are proteins well-packed?. Biophys J 81:751–766

Liang J, Edelsbrunner H, Woodward C (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci 7:1884–1897

Lingrel JB, Townes TM, Shapiro SG, Spence SE, Liberator PA, Wernke SM (1983) Organization, structure, and expression of the goat globin genes. Prog Clin Biol Res 134:131–139

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Maddison WP (2000) Testing character correlation using pairwise comparisons on a phylogeny. J Theor Biol 202:195–204

Martin SL, Vincent KA, Wilson AC (1983) Rise and fall of the delta globin gene. J Mol Biol 164:513–528

Massingham T, Davies LJ, Lio P (2001) Analysing gene function after duplication. Bioessays 23:873–876

Mazet F, Shimeld SM (2002) Gene duplication and divergence in the early evolution of vertebrates. Curr Opin Genet Dev 12:393–396

Maynard Smith J, Smith NH (1998) Detecting recombination from gene trees. Mol Biol Evol 15:590–599

Meireles CM, Schneider MP, Sampaio MI, Schneider H, Slightom JL, Chiu CH, et al. (1995) Fate of a redundant gamma-globin gene in the atelid clade of New World monkeys: Implications concerning fetal globin gene expression. Proc Natl Acad Sci USA 92:2607–2611

Murzin AG, Finkelstein AV (1988) General architecture of the alpha-helical globule. J Mol Biol 204:749–769

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936

O'Brien SJ, Eizirik E, Murphy WJ (2001) Genomics. On choosing mammalian genomes for sequencing. Science 292:2264–2266

Ohno S (1970) Evolution by gene duplication. Springer Verlag, Berlin

Ohta T (1980) Amino acid diversity of immunoglobulins as a product of molecular evolution. J Mol Evol 15:29–35

Ohta T (1983) On the evolution of multigene families. Theor Popul Biol 23:216–240

Ohta T (1984) Some models of gene conversion for treating the evolution of multigene families. Genetics 106:517–528

Ohta T (1987) Simulating evolution by gene duplication. Genetics 115:207–213

Ohta T (1988) Evolution by gene duplication and compensatory advantageous mutations. Genetics 120:841–847

Ohta T (1990) How gene families evolve. Theor Popul Biol 37:213–219

Ohta T (1993) Pattern of nucleotide substitutions in growth hormone-prolactin gene family: A paradigm for evolution by gene duplication. Genetics 134:1271–1276

Ohta T (1998) On the pattern of polymorphisms at major histocompatibility complex loci. J Mol Evol 46:633–638

Ohta T (2000) Evolution of gene families. Gene 259:45–52

Patthy L (1985) Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules. Cell 41:657–663

Perutz MF (1983) Species adaptation in a protein molecule. Mol Biol Evol 1:1–28

Perutz MF, et al. (1960) Structure of haemoglobin, A three-dimensional Fourier synthesis at 5.5 A resolution, obtained by X-ray analysis. Nature 185:416–422

Posada D, Crandall KA, Holmes EC (2002) Recombination in evolutionary genomics. Annu Rev Genet 36:75–97

Poyart C, Wajcman H, Kister J (1992) Molecular adaptation of hemoglobin function in mammals. Respir Physiol 90:3–17

Rowe T (1999) At the roots of the mammalian family tree. Nature 398:283–284

Saban J, King D (1994) Sequence of the sheep fetal beta globin gene and flanking region. Biochim Biophys Acta 1218:87–90

Satoh H, Inokuchi N, Nagae Y, Okazaki T (1999) Organization, structure, and evolution of the nonadult rat beta-globin gene cluster. J Mol Evol 49:122–129

Sawyer SA (1999) GENECONV: A computer package for the statistical detection of gene conversion. Distributed by the author, Department of Mathematics, Washington University, St. Louis (available at http://.math.wustl.edu/∼sawyer)

Schimenti JC (1994) Gene conversion and the evolution of gene families in mammals. Soc Gen Physiol Ser 49:85–91

Schimenti JC, Duncan CH (1985a) Concerted evolution of the cow epsilon 2 and epsilon 4 beta-globin genes. Mol Biol Evol 2:505–513

Schimenti JC, Duncan CH (1985b) Structure and organization of the bovine beta-globin genes. Mol Biol Evol 2:514–525

Schon EA, Cleary ML, Haynes JR, Lingrel JB (1981) Structure and evolution of goat gamma-, beta C-, and beta A-globin genes: Three developmentally regulated genes contain inserted elements. Cell 27:359–369

Shapiro SG, Schon EA, Townes TM, Lingrel JB (1983) Sequence and linkage of the goat epsilon I and epsilon II beta-globin genes. J Mol Biol 169:31–52

Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16:1114–1116

Slatkin M, Maddison WP (1989) A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics 123:603–613

Slightom JL, Chang LY, Koop BF, Goodman M (1985) Chimpanzee fetal G gamma and A gamma globin gene nucleotide sequences provide further evidence of gene conversions in hominine evolution. Mol Biol Evol 2:370–389

Springer MS, Murphy WJ, Eizirik E, O'Brien SJ (2003) Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc Natl Acad Sci USA 100:1056–1061

Swofford DC (1998) PAUP* 4.0-Phylogenetic analysis using parsimony (*and other methods). Version 4.0. Sinauer Associates, Sunderland, MA

Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*) Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203:439–455

Thompson TD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: Flexible strategies for multiple alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

Townes TM, Fitzgerald MC, Lingrel JB (1984) Triplication of a four-gene set during evolution of the goat beta-globin locus produced three genes now expressed differentially during development. Proc Natl Acad Sci USA 81:6589–6593

Wheeler D, Hope R, Cooper SB, Dolman G, Webb GC, Bottema CD, Gooley AA, Goodman M, Holland RA (2001) An orphaned mammalian beta-globin gene of ancient evolutionary origin. Proc Natl Acad Sci USA 98:1101–1106

Worobey M (2001) A novel approach to detecting and measuring recombination: New insights into evolution in viruses, bacteria, and mitochondria. Mol Biol Evol 18:1425–1434

Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496–503

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908–917

Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449

Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18:292–298

Zhang P, Gu Z, Li W-H (2003) Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol 4:R56