J Mol Evol (2004) 59:121–132
DOI: 10.1007/s00239-004-2597-8

JOURNAL OF MOLECULAR EVOLUTION

A Maximum Likelihood Method for Detecting Functional Divergence at Individual Codon Sites, with Application to Gene Family Evolution

Joseph P. Bielawski, Ziheng Yang

Department of Biology, University College London, London, WC1E 6BT, UK

Received: 27 June 2003 / Accepted: 29 December 2003

Abstract. The tailoring of existing genetic systems to new uses is called genetic co-option. Mechanisms of genetic co-option have been difficult to study because of difficulties in identifying functionally important changes. One way to study genetic co-option in protein-coding genes is to identify those amino acid sites that have experienced changes in selective pressure following a genetic co-option event. In this paper we present a maximum likelihood method useful for measuring divergent selective pressures and identifying the amino acid sites affected by divergent selection. The method is based on a codon model of evolution and uses the nonsynonymous-to-synonymous rate ratio (ω) as a measure of selection on the protein, with $\omega = 1$, < 1, and > 1 indicating neutral evolution, purifying selection, and positive selection, respectively. The model allows variation in ω among sites, with a fraction of sites evolving under divergent selective pressures. Divergent selection is indicated by different ω's between clades, such as between paralogous clades of a gene family. We applied the codon model to duplication followed by functional divergence of (i) the ε and γ globin genes and (ii) the eosinophil cationic protein (ECP) and eosinophil-derived neurotoxin (EDN) genes. In both cases likelihood ratio tests suggested the presence of sites evolving under divergent selective pressures. Results of the ε and γ globin analysis suggested that divergent selective pressures might be a consequence of a weakened relationship between fetal hemoglobin and 2,3-diphosphoglycerate. We suggest that empirical Bayesian identification of sites evolving under divergent selective pressures, combined with structural and functional information, can provide a valuable framework for identifying and studying mechanisms of genetic co-option. Limitations of the new method are discussed.

Key words: Maximum likelihood — Functional divergence — Codon model — ECP — EDN — Globins

Introduction

Evolutionary novelty appears to arise more frequently through changes in existing patterns of gene regulation, the function of existing proteins, or both, rather than by invention of completely new genes (Betrán and Long 2002; True and Carrol 2002). The tailoring of existing genetic systems to new uses is called genetic co-option (True and Carrol 2002). Although new genes have been created by assembling normally unrelated genomic segments (e.g., Long 2001; Long and Langley 1993), the observation that the total gene number in complex organisms does not differ greatly from simpler organisms suggests the importance of co-option of pre-existing genetic systems (e.g., Claverie 2001; Betrán and Long 2002). Furthermore, genetic co-option events have been associated with major changes in organism ecology and life history (e.g., Chen et al. 1997; Harris et al. 2002).

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

Although the molecular basis of genetic co-option has been studied only in a few cases, the process is thought to have played a role in the major episodes of adaptive divergence of multicellular organisms (Lynch and Conery 2000; Lynch and Force 2000; Taylor et al. 2001).

Gene duplication is an important mechanism for genetic co-option. It provides a mechanism for evolution of divergent protein functions (Piatigorsky and Wistow 1991; Ohta 1993; Hughes 1994), novel gene expression patterns (Force et al. 1999; Lynch and Force 2000), or both (Gibert 2002; Hughes 2002). For example, a single gene that is expressed in different tissues might experience conflicting selective pressures, the result being a compromise between optimal adaptations for any one tissue. Duplication of such a locus can lead to specialized patterns of expression among gene copies (Force et al. 1999), providing natural selection the freedom to promote tissue-specific functional divergence (e.g., Gibert 2002). The tremendous diversity of extant gene families, taken together with the observation that there is often an acceleration of amino acid substitution rates following gene duplication (Li 1985; Lynch and Conery 2000), suggests that gene duplication has been an important mechanism for functional divergence of genetic systems. However, it is generally difficult to identify the functionally important amino acid changes associated with these events.

In this paper we implement a new method for detecting functional divergence of proteins following a gene duplication event and for identifying the specific amino acid sites involved. The approach is an extension of the model of codon evolution developed by Goldman and Yang (1994; see also Muse and Gaut 1994), Modeling evolution among codons allows maximum likelihood (ML) estimation of the relative rates of nonsynonymous (d_N) and synonymous (d_S) changes. The ratio of these rates ($\omega = d_N/d_S$) is a measure of selective pressure on the protein product of a gene (Yang and Bielawski 2000). For example, if nonsynonymous mutations are deleterious, purifying selection will reduce their fixation rate and d_N/d_S will be less than 1, whereas if nonsynonymous mutations are advantageous they will be fixed at a higher rate than synonymous mutations, and d_N/d_S will be greater than 1. A d_N/d_S ratio equal to one is consistent with neutral evolution. The original model (Goldman and Yang 1994) averaged d_N/d_S over all sites of a gene and lineages of a phylogenetic tree. It was subsequently extended to allow variation in d_N/d_S among sites (Nielsen and Yang 1998; Yang et al. 2000) and among branches (Yang 1998). A recent approach allows variation in d_N/d_S among sites, with additional variation at some sites in a prespecified branch (Yang and Nielsen 2002). Here we describe a model that allows variation in d_N/d_S among sites, with a fraction of sites evolving under divergent selective pressures between two clades following a co-option event. We implement the model in the maximum likelihood framework and apply it to two cases of evolution by gene duplication: (i) the ε and γ globin gene family (Meireles et al. 1995; Johnson et al. 1996; Fitch et al. 1991) and (ii) the eosinophil cationic protein (ECP) and eosinophil-derived neurotoxin (EDN) gene family (Zhang et al. 1998; Zhang and Rosenberg 2002).

Theory

We assume that the phylogeny is given or independently estimated and that there has been some change in selective constraints following some point in evolutionary time that can be specified *a priori*. In this paper we are specifically interested in testing if functional constraints differ significantly between two paralogous clades of genes following a gene duplication event. A duplication event is a point in evolutionary time that can be easily identified *a priori* on a phylogeny.

The model of codon substitution of Goldman and Yang (1994) describes the substitution rate from one sense codon, i, to another, j, as

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions} \\ \mu\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion} \\ \mu\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition} \\ \mu\omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion} \\ \mu\omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition} \end{cases}$$

Parameter κ the transition–transversion rate ratio, π_j is the equilibrium frequency of codon j, and ω ($= d_N/d_S$) is a measure of the selective pressure acting on the protein product of the gene.

We assume that selective pressure varies among the amino acids encoded by a gene. Moreover, we assume that a subset of sites experience a change in selective pressure at a point in evolutionary history

such as a duplication event. We do not know the history of selective pressure at each site, and we wish to identify which sites have experienced a change following the duplication event and estimate the level of selective pressure in each paralogous clade at such sites. To achieve this we considered two "branch-site" models, which we refer to as Models C and D, as they follow two earlier branch-site models (called A and B) implemented by Yang and Nielsen (2002).

Model C is an extension of the site-specific "neutral" model (M1) of Nielsen and Yang (1998). M1 assumes two classes of sites; one class is completely conserved, with $\omega_0 = 0$, and the other class is completely neutral, with $\omega_1 = 1$. Only the proportion of sites under $\omega_0 = 0$ (f_0) is estimated via ML, as $f_1 = 1 - f_0$. Although M1 allows different selective pressure among site classes, it assumes that the same selective pressure (for each site class) acts over all branches of a phylogeny. Model C extends M1 by adding a third class of sites where selective pressure differs in different parts of a phylogeny. The ω parameters of this third class of sites are estimated from the data via ML. Because assumptions under M1 and Model C are too simplistic for many datasets (Yang 2001), we used Model D only to analyze data sets considered in this paper. Model D is an extension of the site-specific "discrete" model (M3) of Yang et al. (2000). For example, M3 ($k = 2$ categories), assumes two classes of sites with proportions f_0, f_1 and ratios ω_0, ω_1 estimated from the data (Fig. 1A). Model D extends M3 by allowing selective pressure at one class of sites to differ in different parts of a phylogeny (Fig. 1B). For gene families, this means a class of sites having two independent ω parameters, one for each clade of paralogous genes that arose by duplication (Fig. 1B: ω_{1A}, ω_{1B}).

Note that Models A and B of Yang and Nielsen (2002) assume four classes of sites in a sequence: the first two classes have different but uniform selective pressure over all branches of the phylogeny (ω_0, ω_1); the third and fourth classes have ω_0 and ω_1 in all but a few "foreground branches," where selective pressure is assumed to have changed (i.e., $\omega_0 \rightarrow \omega_2$ and $\omega_1 \rightarrow \omega_2$). Model A fixes $\omega_0 = 0$ and $\omega_1 = 1$, whereas they are free parameters in Model B. Models A and B were designed for cases where a certain event caused some sites to evolve under positive selection along prespecified branches (i.e., $\omega_2 > 1$). In Model D we are interested in those sites that have evolved under divergent selective pressures (with $\omega_{1A} \neq \omega_{1B}$ in Fig. 1B), and not necessarily in sites under positive selection.

Let h be a codon site and n the number of such sites in a dataset. The observed data $\mathbf{x}_h$ at a site is a vector of codons across the alignment. Let y_h be the site class to which site h belongs, and assume that there are $k = 2$ classes of sites. In the first site class

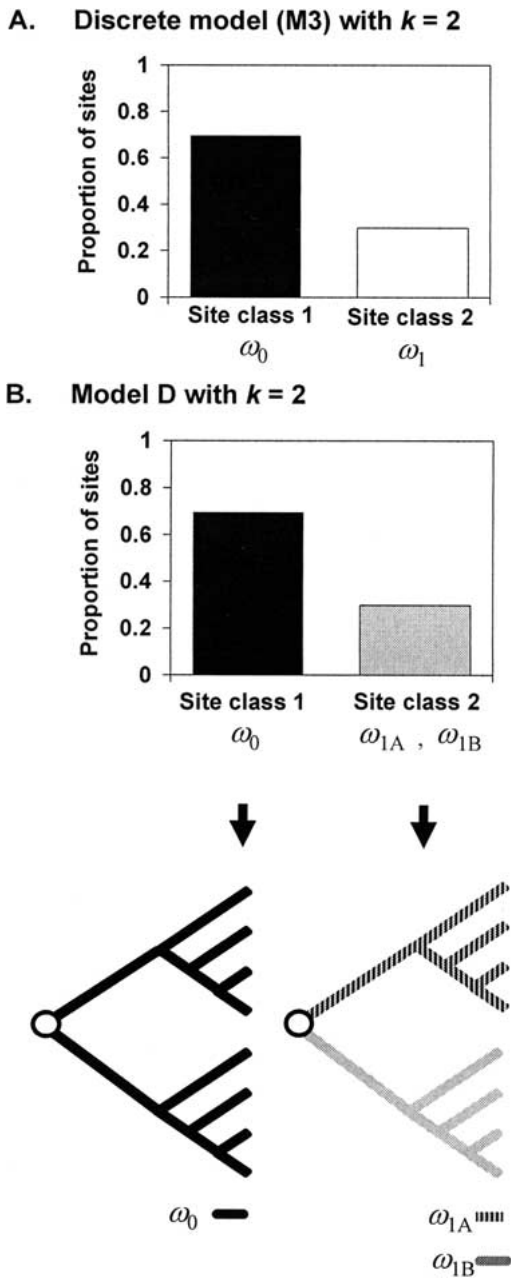

Fig. 1. **A** The discrete model (M3) of Yang et al. (2000) with $k = 2$ site classes assumes that two classes of sites are evolving under different levels of selective pressure (ω_0 and ω_1), with proportions f_0 and f_1. **B** Model D extends M3 by allowing selective pressure at one class of sites to differ in different parts of the phylogeny. A gene duplication is indicated on the topology in B by an open circle. Divergent selection pressure following the duplication event is accommodated by one site class with two independent ω parameters (ω_{1A} and ω_{1B}) for each clade of paralogous genes.

($y_h = 0$) all branches have ω_0, while in the second class ($y_h = 1$) the two clades have two independent ω parameters for paralogous clades (ω_{1A} and ω_{1B}, respectively). The probability of the data at site h, conditioned on the site class $p(\mathbf{x}_h|y_h)$, can be calculated according to Goldman and Yang (1994) if $y_h = 0$, or Yang (1998) if $y_h = 1$. The unconditional probability is an average over the site classes:

$$p(\mathbf{x}_h) = \sum_{k=0}^{1} f_k p(\mathbf{x}_h | y_h = k)$$

We assume that the substitution process at individual codon sites is independent, so that the log likelihood is a sum over all sites in the sequence:

$$l = \sum_{h=1}^{n} \log\{p(\mathbf{x}_h)\}$$

The model was also implemented with $k = 3$ site classes. In this case, if $y_h = 0$ or 1, all branches have the same ω ratio, ω_0 or ω_1, respectively. If $y_h = 2$, the two paralogous clades have ω_{2A} or ω_{2B}, respectively.

Parameters of the model include κ, the ω's, the f's, and the $(2N - 3)$ branch lengths of the phylogeny. These are estimated by numerical maximization of the log likelihood. The branch length measures the expected number of nucleotide substitutions per codon and is defined as an average across site classes (Nielsen and Yang 1998). Codon frequencies (π_i's) are estimated by using observed base or codon frequencies. Since an analytical solution is not possible, an iterative, hill-climbing, algorithm is used to maximize the likelihood function. At each iterative step the algorithm computes a search direction and does a one-dimensional search along that direction. The process is repeated at the best point along each search direction. The iteration continues until there is no improvement in the log-likelihood value, and changes to the parameter values are very small. All parameters, with the exception of codon frequencies, are updated simultaneously.

The likelihood ratio test (LRT) is used to compare the null model (M3) with Model D, which differs only by the assumption of divergent selective pressures at one class of sites following a duplication event. If the assumed topology is unrooted, twice the difference in log likelihood (2δ) under Models M3 and D with the same number of site classes is compared with a χ^2 distribution having one degree of freedom. Note that the example in Fig. 1 shows rooted topologies. In this case there are two degrees of freedom because there is an extra branch length at the root under Model D that is not an identifiable parameter under M3. Significance of the LRT indicates the presence of sites evolving under significantly different selective pressures between the two clades. An empirical Bayes approach, based on ML estimates of model parameters, is used to infer to which class an individual codon site is most likely to belong (Nielsen and Yang 1998). The empirical Bayes approach uses ML parameter estimates in the prior distribution without accounting for their sampling errors. As a result, the accuracy of prediction may be influenced. An alternative is to use the more computationally costly hierarchical Bayesian approach, integrating over the uncertainty in the prior distribution. This is not pursued in this paper.

Data Analysis

We compiled data for two presumed examples of genetic co-option: (i) the divergence of ε and γ globins and (ii) the divergence of the eosinophil cationic protein (ECP) and the eosinophil-derived neurotoxin (EDN). Both these gene families have been well studied. Evidence has been found for the action of positive Darwinian selection since the divergence of ECP and EDN (Zhang et al. 1998; Bielawski and Yang 2003) but not since the divergence of the ε and γ globins (G. Aguileta, pers. commun.). In neither case has there been a specific test of the hypothesis that a long-term shift in selective constraints has occurred at a fraction of sites following gene co-option. We tested this hypothesis in these two cases.

We implemented three types of codon models. The first was the site- and time-homogeneous model, M0 (one ratio), of Goldman and Yang (1994), which averages selective pressure over codon sites and branches. The second was the site heterogeneous model, M3 (discrete), of Yang et al. (2000), with $k = 2$ and $k = 3$ site categories. The third was the new branch-site model, Model D, also with $k = 2$ and $k = 3$ site categories. All models were implemented under two different tree topologies: (i) the expected species tree, derived from the literature (Goodman et al. 1998; Goodman 1999; Meireles et al. 1999; Page et al. 1999), and (ii) the estimated gene tree. Gene trees were estimated using ML under the HKY85 substitution model (Hasegawa et al. 1985) combined with a discrete gamma model of rate variation among sites (Yang 1994). Tree searches were conducted by using the PAUP* computer program (Swofford 2000). All ML analyses of codon models were performed using the codeml program of the PAML package (Yang 1997; http://abacus.gene.ucl.ac.uk/software/paml.html).

ε *and* γ *Globins*

Transport of oxygen from lungs to tissues in vertebrates is accomplished via reversible binding with hemoglobin. In all vertebrates but cylostomes, hemoglobin is a tetramer comprised of two pairs of subunits, with the adult subunits designated α and β. In placental mammals, two paralogs (ε and γ) are expressed during early development instead of β.

ε and γ arose about 80–100 MYA via a tandem duplication of an embryonic ε-type globin (Koop and Goodman 1988). Expression of ε is embryonic in all placental mammals, while γ expression is embryonic only in nonprimate placental mammals and prosimian primates, being delayed to fetal expression in simian primates (Johnson et al. 1996).

Persistence of both ε and γ over 80 to 100 million years of evolution implies strong selective pressure for both gene products, presumably due to some form of genetic co-option and divergence (Fitch et al. 1991). If functions of ε and γ had not diverged, it is likely that one copy would have become nonfunctional via the accumulation of deleterious mutations. Fitch et al. (1991) suggested that the initial gene duplication event was followed by divergence of ε and γ for different "embryonic niches." Later (35 to 55 MYA), a second case of genetic co-option occurred when embryonically expressed γ was recruited for fetal expression in the early simian lineage (Koop and Goodman 1988; Tagle et al. 1988; Meireles et al. 1995; Johnson et al. 1996). The objective of our analysis was, first, to test for divergence in selective pressure between ε and γ and, second, to identify sites consistent with this type of selective pressure if they existed.

The ε globin gene sequences were from *Alouatta seniculus* (GenBank accession number = L25367), *Aotus azarai* (L25371), *Ateles geoffroyi* (L25368), *Brachyteles arachnoids* (L25366), *Callithrix jacchus* (L25363), *Cebus olivaceus* (U18610), *Cheirogaleus medius* (U11712), *Eulemer macaco* (M15735), *Galago crassicaudatus* (M36304), *Homo sapiens* (U01317), *Hylobates syndactylus* (U64616), *Lagothrix lagothrica* (L25358), *Macaca mulatta* (M81364), *Otolemur crassicaudatus* (U60902), *Pan paniscus* (M81362), *Pongo pygmaeus* (X05035), *Saimiri sciureus* (L25354), and *Tarsius syrichta* (M81411). γ globin gene sequences were from *Alouatta seniculus* (AF030097), *Aotus azarai* (U57044), *Ateles paniscus* (AF030093), *Brachyteles arachnoides* (AF030089), *Callithrix jacchus* (AF321384), *Cebus apella* (U57043), *Cheirogaleua medius* (M15758), *Eulemur macaco* (M15757), *Galago crassicaudatus* (M36305), *Homo sapiens* (U01317), *Hylobates lar* (J05174), *Lagothrix lagotricha* (AF030094), *Macaca mulatta* (M19434), *Otolemur crassicaudatus* (U60902), *Pongo pygmaeus* (M16208), *Pan troglodytes* (X03109), *Saimiri ustus* (AF016984), and *Tarsius bancanus* (AF0726810).

The estimated phylogeny for the ε and γ sequences is shown in Fig. 2. That gene tree and a "species" tree assuming the expected species relationships within the ε and γ clades were used in all analyses. Results obtained from the two trees were very similar and only those obtained under the gene tree (Fig. 2) are presented.

The one-ratio model (M0) yielded an estimated $\omega = 0.19$ (Table 1), indicating that purifying selection dominated the evolution of these globins. However,

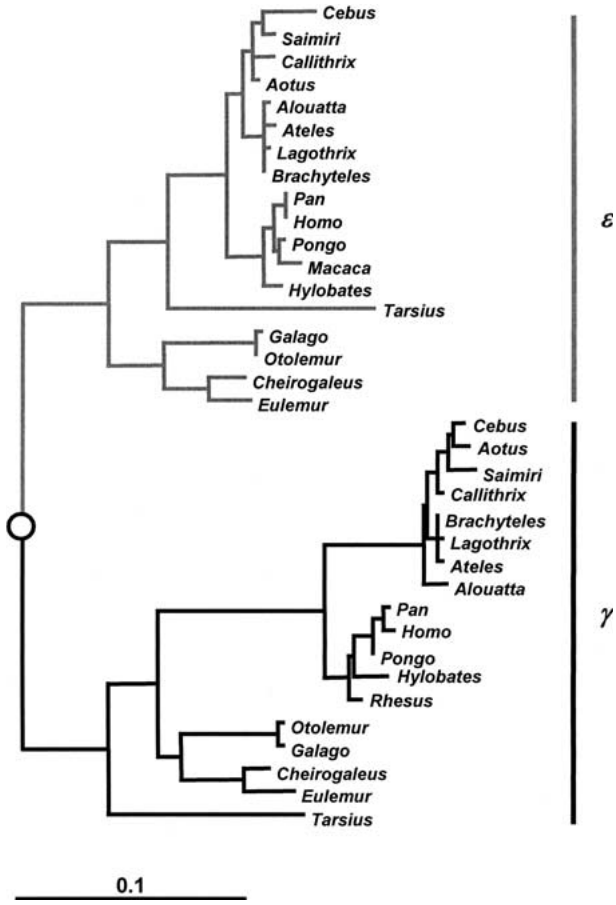

Fig. 2. Gene tree for the 36 sequences from the ε and γ gene family. The topology was obtained by using maximum likelihood analysis under the HKY85 substitution matrix combined with a correction for among sites rate variation (discrete gamma model). The scale bar indicates the mean number of substitutions per nucleotide site. The open circle indicates the duplication event that gave rise to the ε and γ genes. Under Model D, a fraction of sites was allowed to evolve under divergent selection pressure, with ω_{1A} and ω_{1B} for the two paralogous clades, respectively. *Macaca f.* and *Macaca n.* indicate *M. fascicularis* and *M. nemestrina*, respectively.

this finding is based on an average over all sites and branches. Next we tested for heterogeneous selective pressure among sites. The discrete model (M3), which assumes variation among sites but no variation among branches, was applied to these sequences (Table 1). Likelihood ratio tests (LRTs) of M0 against M3 indicated significant variation in selective pressure among sites (Table 2). An LRT of M3 with $k = 2$ site classes against M3 with $k = 3$ site classes was not significant (Table 2). M3 with $k = 2$ site classes suggested a large fraction of sites (70%) evolving under very strong purifying selection ($\omega = 0.05$), and a small fraction of sites (30%) evolving more quickly, under much weaker purifying selection ($\omega = 0.55$). We note that estimates under M3 with $k = 2$ are quite different from those under M3 with $k = 3$. However, both models provide strong evidence that the ω ratio and selective pressure are highly variable among sites.

Table 1. Parameter estimates and log likelihood scores for the ε and γ globin gene family[a]

Model	p	Estimate of ω parameter	Positive selection	l
M0: One ratio	1	$\omega = 0.19$	None	−2472.83
Site-specific models				
M3: Discrete ($k = 2$)	3	$\omega_0 = 0.05$, $f_0 = 0.70$ $\omega_1 = 0.55$, ($f_1 = 0.30$)	None	−2443.66
M3: Discrete ($k = 3$)	5	$\omega_0 = 0.00$, $f_0 = 0.26$ $\omega_1 = 0.48$, $f_1 = 0.11$ $\omega_2 = 0.60$, ($f_2 = 0.25$)	None	−2443.46
Branch-site models				
Model D ($k = 2$)	4	$\omega_0 = 0.05$, $f_0 = 0.70$ $\omega_{1\varepsilon} = 0.38$, $\omega_{1\gamma} = 0.75$ ($f_1 = 0.30$)	None	−2440.72
Model D ($k = 3$)	6	$\omega_0 = 0.04$, $f_0 = 0.65$ $\omega_1 = 0.61$, $f_1 = 0.19$ $\omega_{2\varepsilon} = 0.008$, $\omega_{2\gamma} = 0.79$ ($f_2 = 0.16$)	None	−2436.52

[a]All analyses conducted using the rooted tree shown in Fig. 2. Frequency parameters, f, shown in parentheses were obtained by subtraction. Equilibrium codon frequencies were obtained under the F3×4 model. p is the number of free parameters in the ω distribution.

Table 2. Likelihood ratio test statistics (2δ) for the ε and γ globin family

	2δ	df	P value
LRTs of variable ω's among sites			
One ratio vs. M3 ($k = 2$)	58.3	2	< 0.0001
One ratio vs. M3 ($k = 2$)	58.7	4	< 0.0001
M3 ($k = 2$) vs. M3 ($k = 3$)	0.4	2	0.82
Model D ($k = 2$) vs. Model D ($k = 3$)	8.40	2	0.01
LRTs of variable ω's among sites and branches			
M3 ($k = 2$) vs. Model D ($k = 2$)	5.88	2	0.053
M3 ($k = 3$) vs. Model D ($k = 3$)	13.88	2	0.001
M3 ($k = 2$) vs. Model D ($k = 3$)	14.28	4	0.006

In order to test for divergence in selective pressure between ε and γ, we applied the new model (Model D), which accommodates both the heterogeneity among sites and divergent selective pressures. Here, we allowed one class of sites to evolve under divergent selective pressures following the duplication of the ancestral ε-type globin (Fig. 2). Significance of the LRT for divergent selective pressures at a fraction of sites was borderline when we assumed $k = 2$ site classes ($2\delta = 5.88$, df = 2, $P = 0.05$), and was unmistakable when we assumed $k = 3$ site classes ($2\delta = 13.88$, df = 2, $P = 0.001$). Parameter estimates under Model D with $k = 3$ site classes suggested a large set of sites (~65%) evolving under strong purifying selection ($\omega = 0.04$), a small set of sites (~19%) evolving under much weaker selective pressure ($\omega = 0.61$), and a small set of sites (~16%) evolving under divergent selective pressures, with very strong purifying selection in the ε-clade ($\omega_{2\varepsilon} = 0.008$), and weak purifying selection in the γ clade ($\omega_{2\gamma} = 0.79$). Model D with $k = 2$ also suggested sites under divergent selection with $\omega_{2\varepsilon} = 0.38$ for the ε clade and $\omega_{2\gamma} = 0.75$ for the γ clade.

We examined the sensitivity of the analysis to tree topology and to assumptions about codon usage. Results shown above were obtained under model F3×4, which uses the nucleotide frequencies at the three positions of the codon to compute the expected codon frequencies (Goldman and Yang 1994). We also examined parameter estimates under two other models of codon frequencies: the Fequal model, which assumes all codons are used equally, and the F61 model, which uses the 61 empirical codon frequencies as parameters. Parameter estimates under Model D with $k = 3$ site classes are similar under all three models of codon frequencies and under both tree topologies, and the qualitative conclusions about selective pressure acting at the three site classes are the same (Table 3).

We identified 12 codon sites with posterior probabilities ≥75% of evolving under divergent selective pressures in ε and γ (2V, 6A, 66K, 70T, 87A, 88K, 117T, 118H, 119F, 127V, 144H; *Callithrix* ε used as amino acid reference). We mapped those sites to the three-dimensional structure of hemoglobin. Most sites had a nonrandom distribution on the tetramer; four were located at or within one residue of the $\alpha_1\,\beta_1$ interface (116, 117, 118, and 126), two were located at binding sites for 2,3-diphosphoglycerate (DPG) (1 and 143), and four were located in the region of the heme pocket (65, 69, 86, 87).

Divergent selective pressures might be related to either co-option associated with the duplication of

Table 3. Parameter estimates for the ε and γ globin family under Model D with $k = 3$ and gene tree (and species tree)[a]

Parameter	Model of codon frequencies		
	Fequal	F3×4	F61
ω_0	0.05 (0.05)	0.04 (0.05)	0.05 (0.05)
f_0	0.65 (0.67)	0.65 (0.69)	0.71 (0.71)
ω_1	0.72 (0.73)	0.61 (0.62)	0.58 (0.54)
f_1	0.19 (0.18)	0.19 (0.16)	0.18 (0.18)
$\omega_{2\varepsilon}$	0.00 (0.02)	0.008 (0.05)	0.03 (0.06)
$\omega_{2\gamma}$	0.90 (1.09)	0.79 (1.06)	1.04 (1.46)
[f_2]	[0.16 (0.15)]	[0.16 (0.15)]	[0.10 (0.11)]

[a]Values in brackets are not free parameters and were obtained by subtraction.

embryonic ε-type globin or recruitment of γ for fetal expression. Recruitment of γ is thought to be associated with a gene duplication event that resulted in A_γ and G_γ globins, but this event cannot be resolved on a gene tree because frequent gene conversion among A and G copies has made them virtually identical in DNA sequence. However, considerable information is available concerning biochemical consequences of fetal expression of γ. Higher oxygen affinity of fetal blood is required for efficient oxygen transfer from mother to fetus (Poyart et al. 1992). In primates with fetal expression of γ globin, this is accomplished through partial loss of the regulatory effect of 2,3-DPG, which normally binds sites in the two adult β chains of hemoglobin and shifts the equilibrium to low oxygen affinity (Poyart et al. 1992). 2,3-DPG interacts with the hemoglobin tetramer at just seven residues: residues 1, 2, and 82 on both chains and residue 143 on one chain (Perutz and Imai 1980). Posterior probabilities for sites 1 and 143 were highest (>75%) for divergent selective pressures. Thus three of the seven residues relevant to the affinity of fetal hemoglobin to 2,3-DPG were identified to be under divergent selective pressures in our analysis.

We note that our high estimate of $\omega_{2\gamma} = 0.79$ for sites evolving under divergent selection pressures in the γ clade is an average over all lineages in that clade. There are at least two scenarios for the high ratio. The first is that the high ω ratio in γ represents relaxed functional importance at those sites relative to ε. Relaxation of selective pressure, especially at sites that interact with 2,3-DPG, could have been an indirect result of amino acid changes at other sites that increased the oxygen affinity of γ-chain hemoglobin relative to maternal hemoglobin. A second scenario is a short burst of positive Darwinian selection somewhere in the γ clade, yielding an elevated ω ratio when averaged over all the γ clade lineages. It is tempting to speculate that such a burst of adaptive evolution was associated with duplication of embryonic ε-type globin or with recruitment of γ for fetal expression, with positive selection acting directly on those amino acid changes that increased oxygen affinity of γ-chain hemoglobin. However, altered oxygen affinity in γ could have evolved via just a few amino acid substitutions (Poyart et al. 1992), which might not be detected by a long-term average of ω. In either case the results support functional divergence between ε and γ.

ECP and EDN

ECP and EDN are RNase genes that arose about 31 million years ago through a gene duplication event in the ancestor to Old World monkeys and hominoids (Hamann et al. 1990; Zhang et al. 1998). Both ECP and EDN have host-defense roles, but their specific functions differ. ECP is a nonspecific toxin to bacteria and parasites, probably through cell membrane disruption, whereas EDN acts as a potent antiviral agent through degradation of viral RNA (Rosenberg and Domachowske 1999). ECP also has antiviral activity, but it is substantially less effective than Old World monkey EDN (Domachowske et al. 1998).

Evolution of this gene family has been well studied. Zhang et al. (1998) found an excess of nonsynonymous substitutions over synonymous substitutions in the branch leading to the ECP gene, a pattern consistent with positive Darwinian selection. Those authors suggested that strong selection for the antiparasitic function of ECP probably acted shortly after gene duplication. Zhang and Rosenberg (2002) later showed that the increased antiviral activity of EDN was predominately due to amino acid substitutions at two interacting sites. Using maximum likelihood methods, Bielawski and Yang (2003) confirmed an elevated rate of nonsynonymous substitution following the ECP–EDN duplication event and also, reported that ECP might have continued to evolve under positive Darwinian selection long after the initial period of functional divergence, whereas EDN evolution had been dominated by purifying selection. Bielawski and Yang (2003) used methods that did not

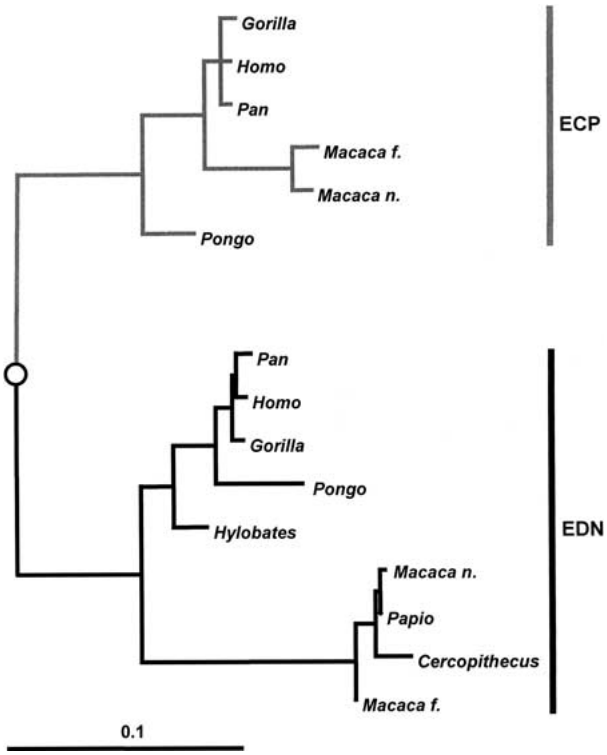

Fig. 3. Gene tree for 15 sequences from the ECP–EDN gene family. The topology was obtained by using maximum likelihood analysis under the HKY85 substitution matrix combined with a correction for among-site rate variation (discrete gamma model). The scale bar indicates the mean number of substitutions per nucleotide site. The open circle indicates the duplication event that gave rise to the ECP and EDN genes. Under Model D, a fraction of sites was allowed to evolve under divergent selection pressure, with ω_{1A} and ω_{1B} for the two paralogous clades, respectively.

account for variable selective pressure among sites. Here, we use Model D to specifically test for a subset of sites evolving under divergent selective pressures in ECP and EDN.

ECP gene sequences were from *Gorilla gorilla* (U24097), *Homo sapiens* (AF294019), *Macaca fascicularis* (U24098), *Macaca nemestrina* (AF479627), *Pan troglodytes* (AF294028), and *Pongo pygmaeus* (U24101). EDN gene sequences were from *Cercopithecus aethiops* (AF479630), *Gorilla gorilla* (U24100), *Homo sapiens* (AF294007), *Hylobates leucogenys* (AF479628), *Macaca fascicularis* (U24096), *Macaca nemestrina* (AF479631), *Pan troglodytes* (AF294081), *Papio hamadryas* (AF479629), and *Pongo pygmaeus* (U24104).

The estimated phylogeny for these ECP and EDN sequences is shown in Fig. 3. Results shown below were obtained by assuming the gene tree; results obtained by assuming the species tree were very similar and are not shown. The one ratio model (M0) yielded an estimated $\omega = 0.85$ (Table 4), indicating a high relative rate of amino acid evolution. Likelihood ratio tests (LRTs) of M0 against M3 indicated significant variation in selective pressure among sites (Table 5). Similar to the globin dataset above, an LRT of M3 with $k = 2$ site classes against M3 with $k = 3$ site classes was not significant (Table 5). M3 with $k = 2$ site classes (Table 4) suggested a large fraction of sites (72%) evolving under purifying selection ($\omega = 0.34$) and a small fraction of sites (28%) evolving under positive Darwinian selection with $\omega = 2.72$. Next, we tested for divergence in selective pressure between ECP and EDN by using Model D; we allowed one class of sites to evolve under divergent selective pressures following the gene duplication event (Fig. 3). LRTs for divergent selective pressures at a fraction of sites were significant whether we assumed $k = 2$ or $k = 3$ site classes (Table 5).

Parameter estimates under Model D vary depending on whether $k = 2$ or $k = 3$. This arises from (i) differences in averaging ω's over sites when assuming two versus three classes of sites; and (ii) sampling errors in ML parameter estimation. Because of sampling errors, particularly when the number of sampled lineages is low, individual parameter estimates must be interpreted cautiously. However, both models suggest the presence of sites evolving under positive selection with $\omega > 1$ in both ECP and EDN and sites evolving under very divergent selective pressures in these two paralogs. Estimates when $k = 3$ suggest a set of sites (~42%) evolving under strong purifying selection ($\omega = 0.07$), a small set of sites (~13%) evolving under positive Darwinian selection in both clades ($\omega = 3.76$), and a set of sites (~45%) evolving under purifying selection in the EDN clade ($\omega = 0.28$) and positive Darwinian selection in the ECP clade ($\omega = 3.21$). The sites evolving under positive selection just in ECP could reflect long-term selective pressure to maintain antiviral activity against respiratory viral pathogens (Bielawski and Yang 2003). We conducted sensitivity analyses and found that the LRTs and qualitative results of ML parameter estimation were robust to model of codon frequencies and tree topology (data not shown). Zhang and Rosenberg (2002) recently reported that additional gene duplication and conversion events occurred in the orangutan *Pongo pygmaeus*. We repeated our analyses on a dataset that excluded the orangutan and obtained very similar findings (data not shown). Because these findings are based on a relatively small sample of lineages, they need to be confirmed in a larger dataset. The empirical Bayes approach uses ML estimates of parameters to identify sites under divergent selection pressure but does not account for their sampling errors. Sampling errors of the ML estimates will be high in small datasets, such as ECP-EDN; as a result, the reliability of Bayesian site identification will be affected and may be sensitive to tree topology and model of codon frequencies.

Table 4. Parameter estimates and log likelihood scores for the ECP–EDN gene family[a]

Model	p	Estimate of ω parameter	Positive selection	l
M0: One ratio	1	$\omega = 0.85$	None	−1729.679
Site-specific models				
M3: Discrete ($k = 2$)	3	$\omega_0 = 0.34$, $f_0 = 0.72$ $\boldsymbol{\omega_1 = 2.72}$ ($f_1 = 0.28$)	Yes	−1707.317
M3: Discrete ($k = 3$)	5	$\omega_0 = 0.00$, $f_0 = 0.34$ $\omega_1 = 1.06$, $f_1 = 0.56$ $\boldsymbol{\omega_2 = 4.56}$ ($f_2 = 0.0.09$)	Yes	−1705.319
Branch-site models				
Model D ($k=2$)	4	$\boldsymbol{\omega_0 = 3.49}$, $f_0 = 0.16$ $\omega_{1(\mathrm{EDN})} = 0.15$, $\boldsymbol{\omega_{1(\mathrm{ECP})} = 1.26}$ ($f_1 = 0.84$)	Yes	−1696.086
Model D ($k = 3$)	6	$\omega_0 = 0.07$, $f_0 = 0.42$ $\boldsymbol{\omega_1 = 3.76}$, $f_1 = 0.13$ $\omega_{2(\mathrm{EDN})} = 0.28$, $\boldsymbol{\omega_{2(\mathrm{ECP})} = 3.21}$ ($f_2 = 0.45$)	Yes	−1691.296

[a]All analyses conducted using the rooted tree shown in Fig. 3. Frequency parameters, f, shown in parentheses were obtained by subtraction. Equilibrium codon frequencies were obtained under the F3×4 model. p is the number of free parameters in the ω distribution.

Table 5. Likelihood ratio test statistics (2δ) for the ECP–EDN family

	2δ	df	P value
LRTs of variable ω's among sites			
One ratio vs. M3 ($k = 2$)	44.72	2	< 0.001
One ratio vs. M3 ($k = 3$)	48.72	4	< 0.001
M3 ($k = 2$) vs. M3 ($k = 3$)	4.00	2	0.136
Model D ($k = 2$) vs. Model D ($k = 3$)	9.58	2	0.008
LRTs of variable ω's among sites and branches			
M3 ($k = 2$) vs. Model D ($k = 2$)	22.46	2	< 0.0001
M3 ($k = 3$) vs. Model D ($k = 3$)	28.05	2	< 0.0001
M3 ($k = 2$) vs. Model D ($k = 3$)	32.04	4	< 0.0001

Discussion

Mechanisms of genetic co-option have been difficult to study, in part, because functionally important changes have been difficult to identify. Recently, this problem has received a lot of attention from the standpoint of amino acid evolution (reviewed by Massingham et al. 2001; and Gaucher et al. 2002). Several approaches have been developed based on the premise that site-specific shifts in rates of amino acid evolution are related to changes in selective pressure (e.g., Gu 2001; Knudsen and Miyamoto 2001; Susko et al. 2002). Such a framework provides an important tool for studying genetic co-option (Massingham et al. 2001; Gaucher et al. 2002), especially for ancient divergences, where saturation of synonymous substitutions excludes a reliable codon-based analysis. However, for more recent divergences, an amino acid-based analysis does not fully utilize the information content of nucleotide datasets (Massingham et al. 2001; Bielawski and Yang 2003). Amino acid rates are limited by an inability to differentiate between different types of selective pressure that give rise to an amino acid rate shift; i.e., positive selection, neutral evolution, and purifying selection. The branch-site model of codon evolution (Model D) presented in this paper should provide a valuable tool for studying genetic co-option when sequences are not too divergent.

In this paper we applied Model D to address two problems of gene family evolution. First, we asked if there was a fraction of sites evolving under divergent selective pressures following a gene duplication event. We expect that the LRT will be a powerful basis for answering such a question, as similar LRTs have been shown to be a powerful and reliable means of testing for site specific heterogeneity in selective pressure (Anisimova et al. 2001). Indeed, we found significant evidence for divergent evolution in both the ε and γ and the ECP and EDN gene families, two well-studied families thought to have undergone functional divergence following gene duplication. The second problem is identification of specific sites involved in functional divergence. We expect this to be a more difficult problem, as information about rates of synonymous and nonsynonymous changes must be divided among paralogous clades, thus increasing sampling errors. If such partitioning results in too few changes along the branches of a specific clade, parameter estimates will be less reliable, as will the posterior probabilities.

Application of Model D to ε and γ divergence demonstrated that important clues to the mode of

adaptive molecular evolution can be obtained from sequences which do not exhibit the characteristic marker of positive Darwinian selection ($\omega > 1$). In particular, our results support the notion that a weakened relationship between γ-chain hemoglobin and 2,3-DPG is connected with molecular adaptation for increased oxygen affinity (Poyart et al. 1992). Furthermore, we note that the majority of sites identified as evolving under divergent selective pressures in ε and γ were associated with the major structural and functional features of the hemoglobin tetramer. We believe that Bayesian site identification in well-sampled datasets, combined with structural and functional information, can provide a valuable framework for identifying and studying mechanisms of genetic co-option.

Model D is very simple in the sense that it allows for only one set of sites evolving under divergent selective pressures; however, two or more such classes of sites might exist. For example, two classes of sites might represent two different domains, one being released from purifying selection in one paralog and the other being released from purifying selection in the other paralog. Model D would be forced to average over both classes of sites, resulting in reduced power of the LRT and lower posterior probabilities. One solution to this problem is to use a bivariate distribution for selective pressure, similar to what has been implemented for amino acid rates (e.g., Gu 2001; Susko et al. 2002). For example, the gamma model of among sites variation in ω (Yang et al. 2000) could be extended by allowing two gamma distributions, one for each paralogous clade. This approach would have the advantage of greater flexibility in modeling functional divergence.

Both site models and branch-site models are computationally complex, with estimation of parameters for finite mixture distributions (such as those in M3 and Model D) being particularly difficult. For example, a small fraction of sites under strong selective pressure might fit a dataset nearly as well as a higher fraction of sites under lower selective pressure. This impacts Bayesian site identification, as ML parameter estimates are used to compute the posterior probabilities. Simulation studies showed that low accuracy in Bayesian site identification occurs when sequence divergence is very low or too few sequences are sampled because under such conditions the sampling errors in ML parameter estimates are too high (Anisimova et al. 2002). Similarly, suboptimal parameter estimates, based on local optimum, also could lead to low accuracy in Bayesian site identification. We found local optima in both datasets under Model D but not under Model M0 or M3. With these points in mind, we make the following recommendations: first, to avoid being trapped at a local optimum, users should run Model D multiple times using different initial values; second, we advise caution on Bayes site prediction when sequence divergence is very low or when few sequences are sampled; and third, different tree topologies and models of codon frequencies may be used to evaluate the robustness of parameter estimates.

Addendum

After submission of the manuscript for this paper, Forsberg and Christiansen (2003) published a similar codon model, which also allows position-specific changes in selection pressure in two different parts of a phylogeny. They applied Gu's (2001) model of functional divergence to the codon model of Goldman and Yang (1994). Their model assumes that the ω ratio varies among sites according to a discrete distribution with three classes. However, a proportion p_d of sites is under different selective pressures and has independent ω's, drawn from the same distribution, for two subclades of a phylogeny. The rest of the sites have the same ω, drawn from the same discrete distribution, for the whole tree. Our Models C and D are different in assuming different overall selective pressures (ω_{2A} versus ω_{2B}) for the two clades. Forsberg and Christiansen (2003) applied their codon model to a set of influenza A virus nucleoprotein sequences to study the changes in selection pressure after a shift from avian to human hosts. They used an LRT to test the hypothesis that $p_d > 0$ and an empirical Bayes procedure to predict which sites had experienced a shift in selection pressures. Both the analysis by Forsberg and Christiansen (2003) and that in this paper demonstrate the importance of codon-based approaches in studying genetic co-option events, whether they be functional divergence following gene duplication or adaptive alteration of existing genes to the functional requirements of parasitizing a new host.

Acknowledgments. Valuable discussions were contributed by Gabriela Aguileta. We thank Katherine A. Dunn and Gabriela Aguileta for constructive comments on the manuscript. This research was supported by a UK Biotechnology and Biological Sciences Research Council Grant.

References

Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18:1585–1592

Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of Bayesian prediction of amino acid sites under positive selection. Mol Biol Evol 19:950–958

Betrán E, Long M (2002) Expansion of genome coding regions by acquisition of new genes. Genetica 115:65–80

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

Chen L, DeVries AL, Cheng CH (1997) Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. Proc Natl Acad Sci USA 94:3811–3816

Claverie JM (2001) Gene number. What if there are only 30,000 human genes? Science 291:1255–1257

Domachowske JB, Bonville CA, Dyer KD, Rosenberg HF (1998) Evolution of antiviral activity in the ribonuclease A gene superfamily: Evidence for a specific interaction between eosinophil-derived neurotoxin (EDN/RNase 2) and respiratory syncytial virus. Nucleic Acids Res 26:5327–5332

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 Y-I, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545

Forsberg R, Christiansen FB (2003) A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. Mol Biol Evol 20:1252–1259

Gaucher EA, Gu X, Miyamoto MM, Benner SA (2002) Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci 27:315–321

Gibert JM (2002) The evolution of engrailed genes after duplication and speciation events. Dev Genes Evol 212:307–318

Goldman N, Yang Z (1994) A codon based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Goodman M (1999) The genomic record of Humankind's evolutionary roots. Am J Hum Genet 64:31–39

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, Groves CP (1998) Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol 9:585–598

Gu X (2001) Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol 18:453–464

Hamann KJ, Ten RM, Loegering DA, Jenkins RB, Heise MT, Schad CR, Pease LR, Gleich GJ, Barker RL (1990) Structure and chromosome localization of the human eosinophil-derived neurotoxin and eosinophil cationic protein genes: Evidence for intronless coding sequences in the ribonuclease gene superfamily. Genomics 7:535–546

Harris MP, Fallon JF, Prum RO (2002) Shh-Bmp2 signalling module and the evolutionary origin and diversification of feathers. J Exp Zool 294:160–176

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

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

Hughes AL (2002) Adaptive evolution after gene duplication. Trends Genet 18:433–434

Johnson RM, Buck S, Chiu C, Schneider H, Sampaio I, Gage DA, Shen TL, Schneider MP, Muniz JA, Gumucio DL, Goodman M (1996) Fetal globin expression in New World monkeys. J Biol Chem 271:14684–14691

Knudsen B, Miyamoto MM (2001) A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc Natl Acad Sci USA 98:14512–14517

Koop BF, Goodman M (1988) Evolutionary and developmental aspects of two hemoglobin beta-chain genes (epsilon M and beta M) of opossum. Proc Natl Acad Sci USA 85:3893–3897

Li W-H (1985) Accelerated evolution following gene duplication and its implications for the neutralist-selectionist controversy. In: Otha T, Aoki K (eds) Population genetics and molecular evolution. Japan Scientific Press, Tokyo, pp 333–352

Long M (2001) Evolution of novel genes. Curr Opin Genet Dev 11:673–680

Long M, Langley CH (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science 260:91–95

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

Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473

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

Meireles CM, Schneider MP, Sampaio MI, Schneider H, Slightom JL, Chiu CH, Neiswanger K, Gumucio DL, Czelusniak J, Goodman M (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

Meireles CM, Czelusniak J, Schneider MP, Muniz JA, Brigido MC, Ferreira HS, Goodman M (1999) Molecular phytogeny of ateline new world monkeys (Platyrrhini, atelinae) based on gamma-globin gene sequences: evidence that brachyteles is the sister group of lagothrix. Mol Phylogenet Evol 12:10–30

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. Mol Biol Evol 11:715–725

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

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

Page SL, Chiu Ch, Goodman M (1999) Molecular phytogeny of Old World monkeys (Cercopithecidae) as inferred from gamma-globin DNA sequences. Mol Phylogenet Evol 13:348–359

Perutz MF, Imai K (1980) Regulation of oxygen affinity of mammalian haemoglobins. J Mol Biol 136:183–191

Piatigorsky J, Wistow G (1991) The recruitment of crystallins: new functions precede gene duplication. Science 252:1078–1079

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

Rosenberg HF, Domachowske JB (1999) Eosinophils, riobnucleases and host defence: solving the puzzle. Immunol Res 20:261–274

Susko E, Inagaki Y, Field C, Holder ME, Roger AJ (2002) Testing for differences in rates-across-sites distributions in phylogenetic subtrees. Mol Biol Evol 19:1514–1523

Swofford DL (2000) PAUP* Phylogenetic analysis using parsimony (*and other methods) Version 4. Sinauer, Sunderland, MA

Taylor J, Van de Peer Y, Meyer A (2001) Genome duplication, divergent resolution and speciation. Trends Genet 17:299–301

Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT (1988) Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatus*). J Mol Biol 203:439–455

True JR, Carrol SB (2002) Gene co-option in physiological and morphological evolution. Annu Rev Cell Dev Biol 18:53–80

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314

Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. 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 Evolut 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 A-MK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449

Zhang J, Rosenberg HF (2002) Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. Proc Natl Acad Sci USA 99:5486–5491

Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA 95:3708–3713