Large-Scale Analysis of Orthologs and Paralogs under Covarion-Like and Constant-but-Different Models of Amino Acid Evolution

Romain A. Studer[1,2] and Marc Robinson-Rechavi*[,1,2]

[1]Department of Ecology and Evolution, Biophore, University of Lausanne, Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

***Corresponding author:** E-mail: marc.robinson-rechavi@unil.ch.

Associate editor: Andrew Roger

Abstract

Functional divergence between homologous proteins is expected to affect amino acid sequences in two main ways, which can be considered as proxies of biochemical divergence: a "covarion-like" pattern of correlated changes in evolutionary rates, and switches in conserved residues ("conserved but different"). Although these patterns have been used in case studies, a large-scale analysis is needed to estimate their frequency and distribution. We use a phylogenomic framework of animal genes to answer three questions: 1) What is the prevalence of such patterns? 2) Can we link such patterns at the amino acid level with selection inferred at the codon level? 3) Are patterns different between paralogs and orthologs? We find that covarion-like patterns are more frequently detected than "constant but different," but that only the latter are correlated with signal for positive selection. Finally, there is no obvious difference in patterns between orthologs and paralogs.

Key words: covarion, positive selection, whole-genome duplication, vertebrate evolution, rate shift, heterotachy.

Research article

Introduction

Gene function changes during evolution, including changes in biochemical function that are expected to be reflected in the amino acid sequence. Most evolutionary models assume that duplication plays a major role in the evolution of such changes (Ohno 1970; Semon and Wolfe 2007; Conant and Wolfe 2008). We have previously shown that positive selection affects vertebrate protein-coding genes relatively frequently, but that there is no increase in its prevalence after duplication (Studer et al. 2008). A general trend of higher divergence after duplication is in fact not so well supported as expected and needs more investigation at all levels of divergence (Studer and Robinson-Rechavi 2009b).

Positive selection is expected to be correlated with functional changes, although a direct link remains to be established (Eyre-Walker 2006). In this work, we explore the divergence between homologous proteins, this time directly in the amino acid sequences. We used two different measures, considered as proxies of biochemical divergence: the "covarion-like" pattern and the "conserved-but-different" pattern (Anisimova and Liberles 2007; Liberles 2007; Studer and Robinson-Rechavi 2009a). Covarion-like evolution means that several sites experience acceleration or deceleration in their evolutionary rate in a correlated way (i.e., in the same evolutionary period), presumably due to variation in selective pressure. Conserved but different refers to a pattern of change from one amino acid to another in a specific evolutionary period for a site that is conserved the rest of the time. It is generally assumed that both these patterns are linked to functional change and driven by positive selection. It is unclear whether this would be the same type of selective events detected by codon models because these amino acid patterns do not necessarily imply a high rate of nonsynonymous to synonymous change.

We answer three questions: 1) What is the prevalence of such patterns? 2) Can we link such patterns at the amino acid level with selection inferred at the codon level? 3) Are patterns different between paralogs and orthologs?

Materials and Methods

Data

We used gene sequences and trees from the database HomolEns release 4 (Penel et al. 2009), which is based on Ensembl release 49 (March 2008; Hubbard et al. 2009) March 2008 being the date of release 49. Genes are organized in families, which include precalculated alignments and phylogenies. The main advantage of the Homolens system is the FamFetch query system and the TreePattern function (Dufayard et al. 2005). We can scan for specific topologies among the 23,155 family trees. We selected four different branches of the evolution of animals (fig. 1): 1) the whole-genome duplication at the basis of teleost fishes (3R) (549 subfamilies extracted), 2) the split between teleost fishes and tetrapods (4,024 subfamilies), 3) the whole-genome duplications at the base of the vertebrates (1R/2R) (1,014 subfamilies), and 4) the split between Protostomia

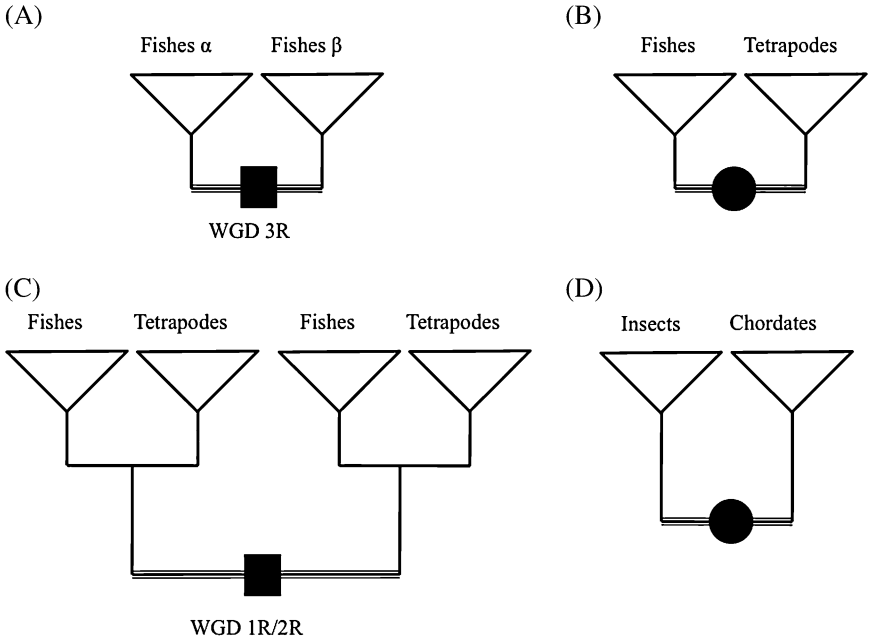

Fig. 1. Tree topologies studied. Boxes represent speciation events and circles represent duplication events. The branches tested are the 3R genome duplication at the basis of teleost fishes (*A*), the speciation between fishes and tetrapods (*B*), the 1R/2R genome duplications at the origin of vertebrates (*C*), and the speciation between insects and chordates (*D*).

(limited to insects in our data set) and Deuterostomia (vertebrates and the two chordates *Ciona intestinalis* and *C. savignyi* in our data set) (1,234 subfamilies). This represents a total of 6,821 branches to test. These events were chosen because they are important in the evolution of animals, and they contain at least four sequenced genomes on each side of the branch, which appears from pilot studies to be a minimum requirement to be able to detect shifts in a significant manner.

For the families thus recovered, we removed species with 2× genome coverage (mostly coming from the Mammalian Genome Project of the the National Institute of Health). The restricted alignments were refined with MUSCLE (Edgar 2004). Computations were then done on the new alignment after removing all columns with at least one gap and extracting the well-aligned part using GBLOCKS (Castresana 2000). Phylogenetic subtrees were extracted from the global trees, and branch lengths reestimated with PhyML release 3.0 (Guindon and Gascuel 2003). For the manipulations of sequences and trees, we combined scripts in Python, BioPython (Cock et al. 2009), Jalview (Waterhouse et al. 2009), and the R library APE (Paradis et al. 2004).

Our data set includes ten species of tetrapods: the frog *Xenopus tropicalis* (Hellsten et al. 2010), the chicken *Gallus gallus* (International Chicken Genome Sequencing Consortium 2004), and the nine mammals: *Monodelphis domestica* (Mikkelsen et al. 2007), *Bos taurus* (Elsik et al. 2009), *Canis familiaris* (Lindblad-Toh et al. 2005), *Equus caballus* (Wade et al. 2009), *Mus musculus* (Waterston et al. 2002), *Rattus norvegicus* (Rat Genome Sequencing Project Consortium 2004), *Pongo pygmaeus abelii* (unpublished data), *Macaca mulatta* (Gibbs et al. 2007), *Pan troglodytes* (The Chimpanzee Sequencing and Analysis Consortium 2005), and *Homo sapiens* (International Human Genome Sequencing Consortium 2001, 2004); five species of teleost fishes: *Danio rerio* (unpublished), *Gasterosteus aculeatus* (unpublished), *Oryzias latipes* (Kasahara et al. 2007), *Tetraodon nigroviridis* (Jaillon et al. 2004), and *Takifugu rubripes* (Aparicio et al. 2002); the two Ciona: *C. intestinalis* (Dehal et al. 2002) and *C. savignyi* (Hill et al. 2008); and four species of insects: *Aedes aegypti* (Nene et al. 2007), *Anopheles gambiae* (Holt et al. 2002), *Apis melifera* (Honeybee Genome Sequencing Consortium 2006), and *Drosophila melanogaster* (Celniker et al. 2002).

Detection of Shifts in Evolutionary Rate (Covarion-Like)

A shift in evolutionary rate can be observed when a particular amino acid is constrained in one part of the phylogenetic tree (subtree) and is relaxed in the other subtree. This has been called "heterotachy" (Lopez et al. 2002; Philippe et al. 2003), "Type I of functional divergence" (Gu 1999, 2001), or "rate-shifting sites" (Abhiman and Sonnhammer 2005). A particular case is the "concomitantly variable codons" (covarions) process (Fitch 1971; Miyamoto and Fitch 1995; Pupko and Galtier 2002). This pattern of evolution postulates that a subset of sites shifting at the same time are more likely to be structurally linked in the protein (Pupko and Galtier 2002). For simplicity, we will use the term covarions for the rest of the paper. Although it should be noted that we do not study structural linkage, we only report cases where many sites have shifted at the same time (same branch of the tree). To detect such covarions, we used Checkcov (Galtier N, personal communication), an implementation of the algorithm described by Pupko and Galtier (2002). Checkcov performs statistical tests

for each residue and can be implemented in an automatic pipeline. The method performs a likelihood ratio test between a null model with only one evolutionary rate per site [rate-among-sites (RAS) model] against an alternative model with two evolutionary rates, one for each subtree. It manages repetition test inside an alignment by using a binomial distribution $B(n, P)$. Briefly, Checkcov assumes that if a covarion process has affected an alignment of length n, we should detect significantly more that 1% of sites at P value <0.01. We used the binomial test from R (binom.test(i, n, P)) to compute the corresponding P value.

We also used Procov to validate the results of Checkcov (Wang et al. 2009). Procov is a general method to detect covarions anywhere in a tree, whereas Checkcov focuses only on a single branch in the tree. To detect whether an alignment includes significant evidence of covarion-like evolution, Procov performs a likelihood ratio test between the RAS model and the general covarion model. For each site, Procov provides the log-likelihood value under the RAS model (1 degree of freedom) and the log-likelihood value under the COV model (3 degrees of freedom). To choose the cutoff to assign a site under a COV model, we used the Akaike information criterion. The difference in log likelihood should be higher than 2 (3 degrees of freedom for COV − 1 degree of freedom for RAS) to be significant. This is an approximation relative to the parametric bootstrap method used in the original paper of Procov (Wang et al. 2009); the parametric bootstrap necessitates a new computation for each data set, resulting in a slightly different cutoff in each case (1.62 for the data set of Wang et al.).

Detection of Change of Conservation Pattern (Constant but Different)

A change in conservation pattern can be seen when a residue is constrained in one subtree for a given property (e.g., a specific amino acid or hydrophobicity) and constrained in the other subtree for a different property (e.g., a different amino acid or polarity). This has been called "Type II of functional divergence" (Gu 2006), "conservation-shifting sites" (Abhiman and Sonnhammer 2005), or "constant but different" (CBD) (Gribaldo et al. 2003). We will use the term CBD for the rest of the paper. We focus only on radical changes in physicochemical properties, such as an acidic to a basic amino acid or a polar to a hydrophobic amino acid. We use Burst after Duplication with Ancestral Sequence Prediction (BADASP) (Edwards and Shields 2005) (release 1.3), which implements the Burst after Duplication (BAD) algorithm (Caffrey et al. 2000). BAD computes the observed differences between two subtrees by the comparison between ancestral conservation and recent conservation. The cutoff value and the minimum number of sites per family were chosen after simulations and after comparison with a positive selection data set (see Results).

Simulations

We performed simulations to test the accuracy and the power of the tests we use. For each subfamily from Homolens, we simulated five alignments of amino acid sequences. We specified the sequence length after removing gaps from the original alignment. We simulated the first 90% of the alignment under a nearly neutral process and the last 10% under a covarion process. The nearly neutral process was based on a RAS model with four categories and the alpha parameter estimated from the real data, using Evolver from PAML (Yang 2007), with a fixed tree topology. The covarion process was simulated by applying for each column a randomly chosen amino acid constraint in one subtree (100% conservation) and random amino acids in the other subtree (no conservation). Using Checkcov, we expect to find 0–1% of covarion sites in the first 90% of the simulated alignments and up to 100% of covarion sites in the last 10%.

We also used simulations derived from the 767 nonduplication trees of our positive selection data set of vertebrates (Studer et al. 2008). This was used to test the CBD under the neutral hypothesis and to define the minimal percentage of sites expected. These simulated multiple alignments followed a RAS model using Evolver.

Statistical Analysis

Over- and underrepresentation in gene ontology (GO) terms was estimated using the TopGO library (Alexa et al. 2006) from Bioconductor (Gentleman et al. 2004). We used false discovery rate to correct for repetition test for enrichment. All other statistical analyses were performed using R (R Development Core Team 2007) and the QVALUE package for test repetition (Storey and Tibshirani 2003).

Results

Comparison with Data of Positive Selection

The data from our previous study of positive selection consists in 767 families of singleton genes of vertebrates (Studer et al. 2008). The analysis has been performed by CodeML from the PAML package (Yang 2007). The model used was the branch-site model (Zhang et al. 2005), able to predict episodic positive selection, and identify amino acids position by providing a Bayes empirical Bayes (BEB) score (Yang et al. 2005).

First, we focused on sites. The BEB scores per site were retrieved for the test on the branch between tetrapods and teleost fishes. We assigned a value of 0 for sites with a BEB $<$ 50% and used the real BEB value otherwise. We then associated to these sites the corresponding chi-square value from Checkcov (COV) and the BAD score from BADASP. We have data for a total of 330,067 sites. We found a weak positive correlation between BEB and COV (Pearson's $r_{xy} = 0.07$, $P < 2.2 \times 10^{-16}$), a higher positive correlation between BEB and BAD (Pearson's $r_{xy} = 0.26$, $P < 2.2 \times 10^{-16}$), and, as expected, a weak negative correlation between COV and BAD (Pearson's $r_{xy} = -0.13$, $P < 2.2 \times 10^{-16}$). To verify that results are not influenced by potential saturation of synonymous changes, we repeated the analysis excluding branches with $dS > 1$. All trends are similar (BEB–COV $r_{xy} = 0.04$, BEB–BAD $r_{xy} = 0.19$, and COV–BAD $r_{xy} = -0.16$) consistent with the lack of saturation reported in Studer et al. (2008). In a second step, we made four different classes of sites (BEB $<$ 50%, 50% $\leq$ BEB $<$ 95%, 95%

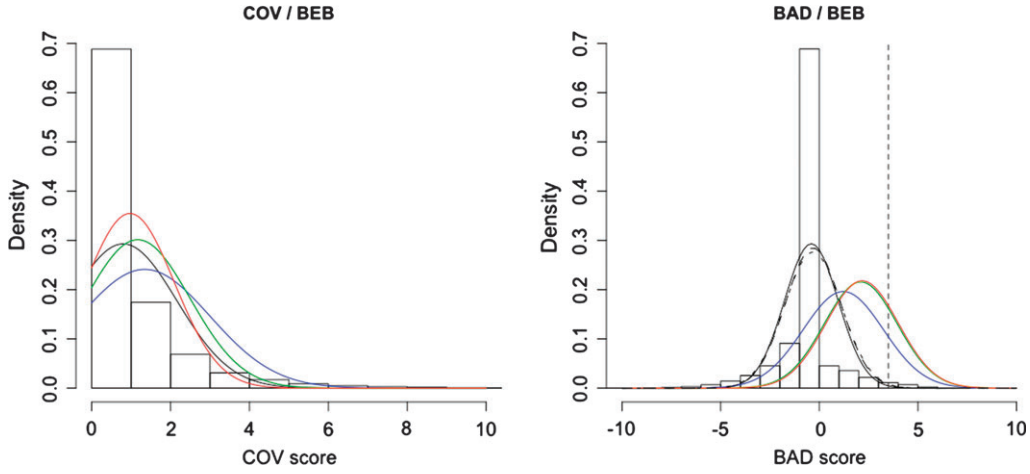

FIG. 2. Correlation of sites under positive selection with covarions and CBDs. (*A*) The histogram presents the values of the chi-square score per site for covarions. The curves are the chi-square values for different BEB intervals (posterior probability for a site to be under positive selection): black for BEB $<$ 50%, blue for BEB $<$ 95%, green for 99% $<$ BEB, and red for BEB $\geq$ 99%. (*B*) The histogram represents the values of BAD scores per site for CBD. The curves are the BAD values for different BEB intervals: black for BEB $<$ 50%, blue for BEB $<$ 95%, green for 99% $<$ BEB, and red for BEB $\geq$ 99%. The dashed curve represents BAD scores of nearly neutral simulated sequences. The dashed line at 3.5 is the 99th percentile of BAD scores for BEB $<$ 50%.

$\leq$ BEB $<$ 99%, and BEB $\geq$ 99%), and we plotted the distribution of covarion and CBD scores among these classes. These distributions show that the correlation between covarions and positive selection is not biologically significant (fig. 2*A*) because sites have the same distribution of covarion scores whatever their posterior probability of positive selection. On the other hand, there is a clear shift in CBD scores between the different classes of positive selection (fig. 2*B*). Of note, these results are consistent with the detection of a "bona fide" signal by the test for positive selection.

Because BAD scores are not associated to a statistical test but are associated to positive selection, we used BEB values to identify a cutoff for BAD scores. We identified the 99th percentile for the distribution of sites in the class of BEB $<$ 50%. We found a BAD score value of 3.50, which will be defined as our cutoff to detect CBD. This is consistent with the cutoff value of 3+ recommended by the authors (BADASP manual, p. 23). It is also consistent with the distribution of BAD scores for sequences simulated under nearly neutral evolution (dashed curve in fig. 2). To fix the expected proportion of CBD sites (BAD score $\geq$3.5) in an alignment under the null hypothesis, we used these simulated alignments and selected the 99th percentile of the distribution (fig. 3). We obtain a limit of 3.9% (rounded

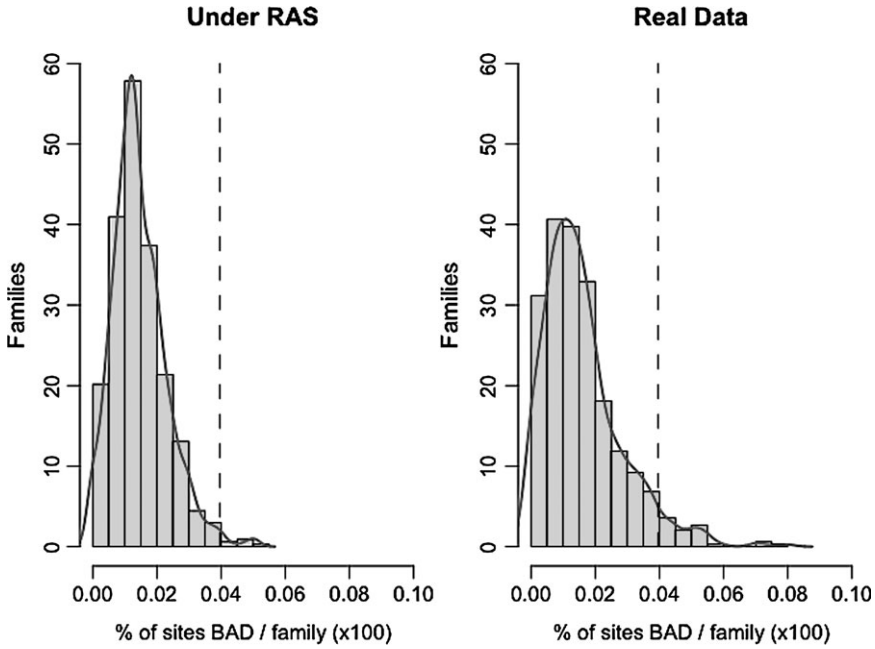

FIG. 3. Percentage of CBDs per families under RAS model and real data. Histograms of proportion of CBD sites with BAD score $>$3.5 per subtree in data simulated under a neutral model of RAS (*A*) and in real data (*B*). The dashed line at 3.9% is the 99th percentile under the neutral model.

Table 1. Evaluation of the Accuracy and Power of the Test for Covarions.

Event Studied	Number of Families[a]	Average Number of Taxa	Average Sequences Length	Accuracy under Neutral Evolution (=1 − Percentage Covarion) (%)	Power under Covarion Process (=Percentage Covarion) (%)
Duplication 3R	2,745	9.1	390.7	99.6	75.2
Speciation tetrapods–fishes	20,120	16.2	353.8	99.4	94.5
Duplication 2R	5,070	31.3	229.5	99.4	99.2
Speciation chordates–insects	6,170	28.0	254.9	99.2	94.5

[a] Five simulations per set of parameters derived from one family of real data.

to 4% for further analyses). These cutoff values (BAD $\geq$ 3.5; sites $>$4%) provide us with a stringent test of CBD for further analyses.

Covarion-Like Sites

We first performed tests on simulated data (table 1). For fish specific duplicates (3R), we recovered 75% of covarion sites. This power is not very high, but the accuracy is more than 99.6%. In other branches, we recovered 94–99% of covarion sites without losing accuracy, which is always above 99.2%.

On real data, various percentages of families under covarion processes were found, above the 1% threshold (table 2). In the most recent event tested, the 3R duplication, only two families (0.4%) are significant, and we suspect misalignment for one case (ENSDARP00000012602; Dup3R/HBG059441-2 in our Web supdata: http://bioinfo.unil.ch/supdata/). The 3R is the event with the lowest number of sequences per family (nine in average), and this could provide less power, as seen in simulations. For the speciation between tetrapods and fishes, there are twice as many genes per family (16 on average), and we find 33 families (0.8%). This result is coherent with a preliminary study (Studer and Robinson-Rechavi 2009a), where we found a similar trend using another tool (ShiftFinder; http://sites.univ-provence.fr/evol/phylogenomics-lab/PageWeb/SHIFTFINDER.htm) to detect shifts in evolutionary rates.

The highest number of significant families (8.5%) is found for the genes associated to the 2R duplication at the base of vertebrates. This is also the data set with the most sequences per family, 31 on average. The oldest event tested is the speciation event between insects and chordates, where we found 5.9% of families with an average of 28 sequences per family.

A different approach to the detection of covarion-like sites is implemented in Procov (Wang et al. 2009). Procov detects shifts that occurred on any branch of the tree and expectedly detects more shifts than the targeted approach of Checkcov. In our data set, Checkcov identifies 3% of significant families, whereas Procov identifies 66% of significant families. Among the 194 families identified by Checkcov, Procov also detected 97% of them. Of the 17,787 sites detected by Checkcov at the 1% threshold, 95% present some signal in Procov ($\ln(L\_cov) - \ln(L\_RAS) > 0$). Moreover, there is a strong correlation of Checkcov and Procov scores, when comparing sites that are significant under Checkcov (Spearman's $\rho = 0.41$, $P < 2.2 \times 10^{-16}$). Finally, using very conservative thresholds (ten for Checkcov and two for Procov), more than 50% of the 2,961 sites from Checkcov are recovered by Procov, whereas detecting exact covarion sites is a difficult problem.

CBD Sites

We found no significant CBD sites in 3R duplicates and very few in other events (table 3): tetrapode–teleost speciation (0.5%), 2R duplication (0.4%), and chordate–insect speciation (1.4%). This is probably due to our stringent cutoff and control for false discovery rate. Interestingly, the average of CBD sites in significant families is quite high (≈10%), supporting the relevance of those shifts that we do detect.

Global Trends

A potential confounding effect in the analysis of 2R duplicate genes is the evolutionary rate of fish genes, which

Table 2. Results of the Detection of Covarions.

Event Studied	Number of Families	Average Number of Sequences	Average Number of Sites	Families without Significant[a] Signal for Covarions		Families with Significant[a] Signal for Covarions		
				Number of Families (%)	Mean Branch Length[b]	Number of Families (%)	Percentage of Covarions Sites (%)	Mean Branch Length
Duplication 3R	549	9.1	391.8	547 (99.6)	0.134	2 (0.4)	3.2	0.201
Speciation tetrapods–fishes	4,024	16.2	355.2	3991 (99.2)	0.247	33 (0.8)	3.5	0.445
Duplication 2R	1,014	31.3	231.0	928 (91.5)	0.290	86 (8.5)	4.8	0.522
Speciation chordates–insects	1,234	28.0	256.2	1161 (94.1)	0.426	73 (5.9)	4.4	0.667

[a] P value = 1% and Q value threshold at 10%.
[b] In amino acid substitutions.

Table 3. Results of the Detection of CBD.

				Families without Significant[a] Signal for CBD		Families with significant[a] signal for CBD		
Event Studied	Number of Families	Average Number of Sequences	Average Number of Sites	Number of Families (%)	Mean Branch Length[b]	Number of Families (%)	Percentage of BADASP Sites (%)	Mean Branch Length
Duplication 3R	549	9.1	391.8	549 (100)	0.134	0 (0)	NA	NA
Speciation tetrapods–fishes	4,024	16.2	355.2	4003 (99.5)	0.245	21 (0.5)	9.7	0.900
Duplication 2R	1,014	31.3	231.0	1010 (99.6)	0.305	4 (0.4)	12.7	1.592
Speciation chordates–insects	1,234	28.0	256.2	1217 (98.6)	0.430	17 (1.4)	10.3	1.149

[a] Cutoff at 4%, based on simulation data, and Q value threshold at 10%.
[b] In amino acid substitutions.

evolve faster than tetrapods (Brunet et al. 2006; Steinke et al. 2006). We recomputed the 2R families after removing all fish sequences. The first observation, as expected, is a global increase of the mean length in the tested branch, from 0.257 amino acid substitutions per site to 0.439, because we measure the branch between the 2R event and the amphibian–amniote speciation, instead of the branch between the 2R event and the earlier teleost–tetrapode speciation. More interesting, there is a decrease in the number of families with significant covarion patterns (from 8.5% to 4.4%) and inversely a large increase in the number of families significant for CBD (from 0.4% to 16.8%). The fast-evolving fish genes probably increased the global heterogeneity in amino acid alignments, decreasing the signal for paralog-specific conservation of amino acids (CBD).

GO enrichment shows that our data set is biased toward slow-evolving genes, as in Studer et al. (2008), for similar reasons of stringent data collection. Significant results in covarions revealed only a few categories, typical of fast-evolving proteins; no GO categories were enriched in CBD genes. With these results, we can only say that it seems that covarions and CBDs can be found among most categories of genes.

Finally, we evaluated possible confounding factors by computing a linear model (analysis of variance) testing the effect of different parameters describing gene families (tables 4 and 5). The number of sequences and the branch length are correlated with the percentage of significant sites for both tests, whereas the number of sites analyzed has no impact. It should be noted that the branch length explains up to half of the variance in the CBD analyses. However, at least 72% of the variance is explained by none of these parameters for covarions and at least 44% of the variance for CBD sites. Presumably, most of this remaining variance is due to shifts in function and selection.

Discussion

Covarion-Like and CBD patterns as Proxies of Functional Divergence

We found a variable proportion of protein families to be significantly under covarion process, depending of the branch tested, from 0.8% to 8.5% (table 2), and very few families with a CBD pattern. The proportion of 2% found by Gruenheit et al. (2008) in their balanced data set between two monophyletic groups is within the same range.

Table 4. Effect of Potential Confounding Factors on the Detection of Covarions.

	Duplication 3R		Speciation Tetrapods–Fishes		Duplication 2R		Speciation Vertebrates–Insects	
Variable	Variance Explained (%)	*P* Value[a]	Variance Explained (%)	*P* Value[a]	Variance Explained (%)	*P* Value[a]	Variance Explained (%)	*P* Value[a]
Number of genes	2	3.8×10^{-04}	2	2.2×10^{-16}	2	1.5×10^{-07}	3	5.5×10^{-12}
Number of sites	0	1.5×10^{-01}	0	2.1×10^{-01}	0	8.8×10^{-01}	0	7.2×10^{-01}
Branch length separating subtrees	2	2.0×10^{-04}	12	2.2×10^{-16}	16	2.2×10^{-16}	6	2.2×10^{-16}
Number of branches in subtree alpha	2	1.1×10^{-04}	0	8.6×10^{-04}	0	8.3×10^{-01}	0	2.9×10^{-01}
Sum of branch lengths in subtree alpha	8	2.9×10^{-13}	3	2.2×10^{-16}	1	5.6×10^{-05}	1	1.7×10^{-03}
Number of branches in subtree beta[b]	NA	NA	0	6.1×10^{-01}	0	3.7×10^{-01}	NA	NA
Sum of branch lengths in subtree beta	10	2.2×10^{-16}	5	2.2×10^{-16}	1	8.5×10^{-04}	1	3.3×10^{-04}
Median_diff[c]	2	2.8×10^{-05}	4	2.2×10^{-16}	0	1.3×10^{-02}	5	2.2×10^{-16}
Residuals	74		72		79		84	

NOTE.—NA, non-available.
[a] Italic values indicate significant after a Bonferroni correction ($\alpha = 0.05/4 = 0.0125$).
[b] The values NA were removed from the analysis of variance because the number of branches in the subtree beta is identical to the number of branches in the subtree alpha.
[c] Median_diff represents the difference between the medians of all branch lengths for both trees.

Table 5. Effect of Potential Confounding Factors on the Detection of CBD Sites.

	Duplication 3R		Speciation Tetrapods–Fishes		Duplication 2R		Speciation Vertebrates–Insects	
Variable	Variance Explained (%)	*P* Value[a]	Variance Explained (%)	*P* Value[a]	Variance Explained (%)	*P* Value[a]	Variance Explained (%)	*P* Value[a]
Number of genes	3	2.2×10^{-09}	4	2.2×10^{-16}	1	1.5×10^{-06}	4	2.2×10^{-16}
Number of sites	0	5.3×10^{-01}	0	2.1×10^{-02}	0	5.2×10^{-01}	0	4.6×10^{-01}
Branch length separating subtrees	51	2.2×10^{-16}	52	2.2×10^{-16}	39	2.2×10^{-16}	39	2.2×10^{-16}
Number of branches in subtree alpha	0	9.6×10^{-01}	0	8.0×10^{-01}	0	5.7×10^{-01}	0	2.4×10^{-01}
Sum of branch lengths in subtree alpha	0	2.3×10^{-01}	0	6.6×10^{-02}	0	2.6×10^{-02}	0	5.7×10^{-01}
Number of branches in subtree beta[b]	NA	NA	0	6.8×10^{-01}	0	2.1×10^{-01}	NA	NA
Sum of branch lengths in subtree beta	1	1.8×10^{-04}	0	9.0×10^{-02}	0	2.4×10^{-01}	0	9.1×10^{-04}
Median_diff[c]	0	5.2×10^{-01}	1	6.9×10^{-16}	1	5.2×10^{-04}	2	8.1×10^{-13}
Residuals	44		44		59		55	

NOTE.—NA, non-available.

[a] Italic values indicates significant after a Bonferroni correction ($\alpha = 0.05/4 = 0.0125$).

[b] The values NA were removed from the ANOVA because the number of branches in the subtree beta is identical to the number of branches in the subtree alpha.

[c] Median_diff represents the difference between the medians of all branch lengths for both trees.

This suggests that covarion-like patterns may reflect relatively frequent small refinements in function or even compensatory mutations without change in function. Importantly, most cases of covarion that we detect with Checkcov are also detected by the more recent Procov method (Wang et al. 2009). The rarity of CBD is consistent with more radical functional changes. It is also possible that CBD sites are more difficult to discriminate from other slow-evolving sites, resulting in rare detection under stringent criteria. Interestingly, the proportion of sites found when a branch is significant for a family is quite high for both patterns, and highest for CBD, which is consistent with the detection of radical functional shifts in protein function in at least some cases. Of note, removing fast-evolving fish genes increased our capacity to detect CBD patterns between tetrapode paralogs. Thus, we should be careful to take into account evolutionary rate differences between species in the study of paralog divergence.

We have tried to apply strict cutoffs for both covarion and CBD, but it is difficult to estimate the true level of changes in amino acid evolution that will affect function in a biologically meaningful way (Levasseur et al. 2007). Case studies present both examples of the evolution of a new function (Braasch et al. 2006) and of more subtle optimization of the original function (Christin et al. 2008). In general, minor sequence changes can affect structural properties that are directly linked to the biochemical function (Tokuriki and Tawfik 2009). At the extreme, the bacterial melamine deaminase shares 98% of identity with the atrazine chlorohydrolase (Seffernick et al. 2001). A correlation of covarions as detected in sequences with structural and functional divergence has been found in several case studies. This is the case for bacterial and eukaryotic elongation factors, which differ in function (Gaucher et al. 2001), and for which covarions were confirmed by crystal structure (Gaucher, Das, et al. 2002). It is also the case in caspases (Wang and Gu 2001). In vertebrate hemoglobin α and β, CBD sites seem to be a more reliable predictor of function divergence than covarions (Gribaldo et al. 2003). Other case studies have highlighted the importance of studying covarion or CBD patterns in correlation with protein function (Gaucher, Gu, et al. 2002; Philippe et al. 2003; Liberles 2007). A difficulty in comparing results from various studies is the use of different methods, with no common standard, and notably differences in the treatment of radical versus conservative changes (Liberles 2007).

Both covarions and CBD sites can result from either gain or loss of function. A covarion pattern may indicate a site, which was kept functionally constrained in one subtree but lost this constraint in the other, or the recruitment of a site (newly constrained) for a novel function. And a CBD pattern may indicate a site conserved in the ancestor which changed by positive selection to get a new function, or a partition of the ancestral protein into different functions, especially through escape from adaptive conflict (Conant and Wolfe 2008).

The covarion model has been recently extended to detect shifts in any branch in a phylogenetic tree (Penn et al. 2008). However, this method needs dozens of sequences, which is not yet applicable to the comparison of completely sequenced animal genomes. It would be interesting in the future to develop methods to analyze such patterns between three subtrees: the two lineages of interest and an outgroup, which will help to discriminate between gain and loss of function (Studer and Robinson-Rechavi 2009b).

Can We Link These Shifts at the Amino Acid Level with Positive Selection Inferred at the Codon Level?

Although the models of amino acid changes are not able to discriminate between a relaxation of purifying selection and positive selection, this is the main focus of codon models (Anisimova and Liberles 2007). But an advantage of

amino acid patterns is the potential for comparisons between sequences for which synonymous substitutions are saturated. Thus, a correlation between amino acid shifts and the detection of positive selection would be interesting. We found only a weak positive correlation between scores of positive selection and of covarions. Among sites identified as under positive selection on a specific branch, there are more sites under purifying selection on background branches (type K2a of CodeML; Yang 2007) (~6.5% in Studer et al. 2008), than sites under neutral evolution on background branches (type K2b) (~1% in Studer et al. 2008). The type K2a of codon model with its slow evolutionary rate except on one branch is closest to a CBD pattern. The covarion pattern may be more similar to the type K2b but is not quite the same. CodeML K2b sites are expected to be variable in all background branches, whereas covarion sites are only variable in one subtree. In fact, the covarion pattern is most similar to the clade model of CodeML (Bielawski and Yang 2004), which has not been used in any large scan to our knowledge.

In any case, it appears that by detecting CBD sites, we do detect a signal of positive selection similar to that inferred by the branch-site model. This opens the possibility to scan for such shifts in very ancient events, where synonymous substitutions are clearly saturated.

Is the Incidence of Covarion-Like and CBD Patterns Different between Paralogs and Orthologs?

In a study of vertebrate hemoglobin α and β, metrics of protein divergence were compared in their capacity to distinguish paralog divergence from ortholog divergence (Gribaldo et al. 2003). But on a larger data set, there is no obvious excess of either covarion-like or CBD sites on duplication branches relative to speciation branches (tables 2 and 3). The main explanatory variables for test results are branch length and the number of genes analyzed (tables 4 and 5). If anything, there seems to be a slight deficit of CBD sites after duplication, but this could be due to other factors, such as the evolutionary rate acceleration in teleost fishes. This is consistent with a previous report of similar levels of amino acid variability in orthologs and paralogs on a smaller sample (Conant et al. 2007). It has been suggested that the divergence of paralogs owes more to expression patterns than to biochemical function (Wapinski et al. 2007), although a stronger divergence of expression between paralogs than between orthologs remains to be established (Studer and Robinson-Rechavi 2009b). At the protein level, in any case, it appears that functional shifts could be frequent not only between paralogs but also between orthologs and that we should be careful when applying orthology to define the function of new genes (Lynch 2009; Studer and Robinson-Rechavi 2009b).

Conclusions

In our study, we found that the CBD pattern correlates well with sites predicted by a branch-site model. This pattern could be used as a proxy of positive selection in very ancient evolutionary events, avoiding the classical problem of saturation of synonymous sites. Positive selection is expected to be involved in strong changes that affect the protein structure and function, which could be causing the CBD pattern. Most evolutionary models predict a burst of functional change after duplication (Force et al. 1999; Conant and Wolfe 2008). We find that changes at the amino acid level, while not infrequent, affect both orthologs and paralogs similarly. Thus, it seems that the same pattern of functional divergence affects both paralogs and orthologs genes, and this might be, rather than an exception, a general trend (Studer and Robinson-Rechavi 2009b). If there are any differences in gene evolution after duplication, it is possible that they affect expression patterns rather than protein biochemical function as reflected in sequences (Wapinski et al. 2007).

Acknowledgments

We acknowledge funding from Etat de Vaud and Swiss National Science Foundation (grant 116798). The computations were performed at the Vital-IT (http://www.vital-it.ch) center for High Performance Computing of the Swiss Institute of Bioinformatics. We thank Nicolas Galtier, Pierre Pontarotti, Huaichun Wang, Tal Pupko, Adi Stern, and anonymous reviewers for helpful discussions.

References

Abhiman S, Sonnhammer EL. 2005. Large-scale prediction of function shift in protein families with a focus on enzymatic function. *Proteins* 60:758–768.

Alexa A, Rahnenfuhrer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22:1600–1607.

Anisimova M, Liberles DA. 2007. The quest for natural selection in the age of comparative genomics. *Heredity* 99:567–579.

Aparicio S, Chapman J, Stupka E, et al. (41 co-authors). 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 297:1301–1310.

Bielawski JP, Yang Z. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J Mol Evol.* 59:121–132.

Braasch I, Salzburger W, Meyer A. 2006. Asymmetric evolution in two fish-specifically duplicated receptor tyrosine kinase paralogons involved in teleost coloration. *Mol Biol Evol.* 23:1192–1202.

Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol.* 23:1808–1816.

Caffrey DR, O'Neill LA, Shields DC. 2000. A method to predict residues conferring functional differences between related proteins: application to MAP kinase pathways. *Protein Sci.* 9:655–670.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17:540–552.

Celniker SE, Wheeler DA, Kronmiller B, et al. (32 co-authors). 2002. Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. *Genome Biol.* doi:10.1186/gb-2002-3-12-research0079.

Christin PA, Salamin N, Muasya AM, Roalson EH, Russier F, Besnard G. 2008. Evolutionary switch and genetic convergence

on rbcL following the evolution of C4 photosynthesis. *Mol Biol Evol.* 25:2361–2368.
Cock PJ, Antao T, Chang JT, et al. (11 co-authors). 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423.
Conant GC, Wagner GP, Stadler PF. 2007. Modeling amino acid substitution patterns in orthologous and paralogous genes. *Mol Phylogenet Evol.* 42:298–307.
Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet.* 9: 938–950.
Dehal P, Satou Y, Campbell RK, et al. (87 co-authors). 2002. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science* 298:2157–2167.
Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perriere G. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 21:2596–2603.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
Edwards RJ, Shields DC. 2005. BADASP: predicting functional specificity in protein families using ancestral sequences. *Bioinformatics* 21:4190–4191.
Elsik CG, Tellam RL, Worley KC, et al. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 324:522–528.
Eyre-Walker A. 2006. The genomic rate of adaptive evolution. *Trends Ecol Evol.* 21:569–575.
Fitch WM. 1971. Rate of change of concomitantly variable codons. *J Mol Evol.* 1:84–96.
Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.
Gaucher EA, Das UK, Miyamoto MM, Benner SA. 2002. The crystal structure of eEF1A refines the functional predictions of an evolutionary analysis of rate changes among elongation factors. *Mol Biol Evol.* 19:569–573.
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.
Gaucher EA, Miyamoto MM, Benner SA. 2001. Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. *Proc Natl Acad Sci U S A.* 98:548–552.
Gentleman RC, Carey VJ, Bates DM, et al. (25 co-authors). 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5:R80.
Gibbs RA, Rogers J, Katze MG, et al. (176 co-authors). 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316:222–234.
Gribaldo S, Casane D, Lopez P, Philippe H. 2003. Functional divergence prediction from evolutionary analysis: a case study of vertebrate hemoglobin. *Mol Biol Evol.* 20:1754–1759.
Gruenheit N, Lockhart PJ, Steel M, Martin W. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. *Mol Biol Evol.* 25:1512–1520.
Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol.* 16:1664–1674.
Gu X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol Biol Evol.* 18:453–464.
Gu X. 2006. A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. *Mol Biol Evol.* 23:1937–1945.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.
Hellsten U, Harland RM, Gilchrist MJ, et al. (48 co-authors). 2010. The genome of the Western clawed frog Xenopus tropicalis. *Science* 328:633–636.
Hill MM, Broman KW, Stupka E, Smith WC, Jiang D, Sidow A. 2008. The C. savignyi genetic map and its integration with the reference sequence facilitates insights into chordate genome evolution. *Genome Res.* 18:1369–1379.
Holt RA, Subramanian GM, Halpern A, et al. (123 co-authors). 2002. The genome sequence of the malaria mosquito Anopheles gambiae. *Science* 298:129–149.
Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* 443:931–949.
Hubbard TJ, Aken BL, Ayling S, et al. (58 co-authors). 2009. Ensembl 2009. *Nucleic Acids Res.* 37:D690–D697.
International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695–716.
International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931–945.
Jaillon O, Aury JM, Brunet F, et al. (61 co-authors). 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* 431:946–957.
Kasahara M, Naruse K, Sasaki S, et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447:714–719.
Levasseur A, Orlando L, Bailly X, Milinkovitch MC, Danchin EG, Pontarotti P. 2007. Conceptual bases for quantifying the role of the environment on gene evolution: the participation of positive selection and neutral evolution. *Biol Rev Camb Philos Soc.* 82:551–572.
Liberles DA. 2007. Ancestral sequence reconstruction. New York: Oxford University Press.
Lindblad-Toh K, Wade CM, Mikkelsen TS, et al. (234 co-authors). 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438:803–819.
Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. *Mol Biol Evol.* 19:1–7.
Lynch VJ. 2009. Use with caution: developmental systems divergence and potential pitfalls of animal models. *Yale J Biol Med.* 82:53–66.
Mikkelsen TS, Wakefield MJ, Aken B, et al. 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* 447:167–177.
Miyamoto MM, Fitch WM. 1995. Testing the covarion hypothesis of molecular evolution. *Mol Biol Evol.* 12:503–513.
Nene V, Wortman JR, Lawson D, et al. 2007. Genome sequence of Aedes aegypti, a major arbovirus vector. *Science* 316: 1718–1723.
Ohno S. 1970. Evolution by gene duplication. Heidelberg: Springer-Verlag.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.
Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, Perriere G. 2009. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics.* 10(Suppl 6):S3.
Penn O, Stern A, Rubinstein ND, Dutheil J, Bacharach E, Galtier N, Pupko T. 2008. Evolutionary modeling of rate shifts reveals

specificity determinants in HIV-1 subtypes. *PLoS Comput Biol.* 4:e1000214.

Philippe H, Casane D, Gribaldo S, Lopez P, Meunier J. 2003. Heterotachy and functional shift in protein evolution. *IUBMB Life*. 55:257–265.

Pupko T, Galtier N. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. *Proc Biol Sci.* 269:1313–1316.

R Development Core Team. 2007. R: a language and environment for statistical computing. Vienna (Austria): Foundation for Statistical Computing.

Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428:493–521.

Seffernick JL, de Souza ML, Sadowsky MJ, Wackett LP. 2001. Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different. *J Bacteriol.* 183:2405–2410.

Semon M, Wolfe KH. 2007. Consequences of genome duplication. *Curr Opin Genet Dev.* 17:505–512.

Steinke D, Salzburger W, Braasch I, Meyer A. 2006. Many genes in fish have species-specific asymmetric rates of molecular evolution. *BMC Genomics*. 7:20.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*. 100:9440–9445.

Studer RA, Penel S, Duret L, Robinson-Rechavi M. 2008. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res.* 18:1393–1402.

Studer RA, Robinson-Rechavi M. 2009a. Evidence for an episodic model of protein sequence evolution. *Biochem Soc Trans.* 37:783–786.

Studer RA, Robinson-Rechavi M. 2009b. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet*. 25:210–216.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.

Tokuriki N, Tawfik DS. 2009. Protein dynamism and evolvability. *Science* 324:203–207.

Wade CM, Giulotto E, Sigurdsson S, et al. 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* 326:865–867.

Wang HC, Susko E, Roger AJ. 2009. PROCOV: maximum likelihood estimation of protein phylogeny under covarion models and site-specific covarion pattern analysis. *BMC Evol Biol.* 9:225.

Wang Y, Gu X. 2001. Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. *Genetics* 158:1311–1320.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449:54–61.

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25:1189–1191.

Waterston RH, Lindblad-Toh K, Birney E, et al. (222 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol.* 22:1107–1118.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol.* 22:2472–2479.