Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes

Romain A. Studer,[1,2] Simon Penel,[3] Laurent Duret,[3] and Marc Robinson-Rechavi[1,2,4]

[1]*Department of Ecology and Evolution, Biophore, Lausanne University, CH-1015 Lausanne, Switzerland;* [2]*Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland;* [3]*Université de Lyon, CNRS, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne F-69622, France*

A stringent branch-site codon model was used to detect positive selection in vertebrate evolution. We show that the test is robust to the large evolutionary distances involved. Positive selection was detected in 77% of 884 genes studied. Most positive selection concerns a few sites on a single branch of the phylogenetic tree: Between 0.9% and 4.7% of sites are affected by positive selection depending on the branches. No functional category was overrepresented among genes under positive selection. Surprisingly, whole genome duplication had no effect on the prevalence of positive selection, whether the fish-specific genome duplication or the two rounds at the origin of vertebrates. Thus positive selection has not been limited to a few gene classes, or to specific evolutionary events such as duplication, but has been pervasive during vertebrate evolution.

[Supplemental material is available online at www.genome.org.]

How important has positive selection been in the evolution of vertebrate genes? While polymorphism data indicate high levels of positive selection in the *Drosophila* genus (Eyre-Walker 2006), much lower levels are found in mammals, which have smaller population sizes (Zhang and Li 2005; Gojobori et al. 2007). Moreover, the most recent use of likelihood tests of codon evolution has identified only two genes out of 13,888 under positive selections in the human lineage (Bakewell et al. 2007). Less stringent studies had found as many as 9% (Bustamante et al. 2005; Jorgensen et al. 2005; Nielsen et al. 2005; Arbiza et al. 2006). It is not obvious whether these results can be extended to deeper vertebrate evolution.

The study of vertebrate evolution includes two complicating factors relative to mammals or flies. First, the time scale is much larger, with the divergence between ray finned fishes and tetrapodes estimated at 416–422 million years ago (Mya) (Benton and Donoghue 2006). Second, whole genome duplications may have induced changes in selective regimes, either at the origin of vertebrates (2R) (Putnam et al. 2008) or at the origin of teleost fishes, the largest clade of vertebrates (Jaillon et al. 2004). It is expected that for both copies of a gene to be kept after duplication, there should be either fixation of a new function or complementary loss of subfunctions (Force et al. 1999; Lynch et al. 2001). The first implies increased positive selection; the second, relaxation of purifying selection. Although some studies have taken asymmetry of selective pressure as some degree of support for the former (Brunet et al. 2006; Byrne and Wolfe 2006), the evidence is not conclusive (He and Zhang 2005; Scannell and Wolfe 2007).

In this work, we use a rigorous branch-site specific likelihood test (Zhang et al. 2005) to quantify positive selection during several episodes of bony vertebrate evolution. To evaluate the role of duplication, we contrast the evidence for positive selection after duplication to the evidence after speciation. We find that although it affects a small proportion of sites, positive selection is pervasive in vertebrate evolution, and surprisingly, whole genome duplication has no measurable effect on its incidence.

[4]Corresponding author.
E-mail marc.robinson-rechavi@unil.ch; fax 41-21-6924165.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.076992.108. Freely available online through the *Genome Research* Open Access option.

Results

To investigate the impact of positive selection in vertebrate evolution, we have analyzed all gene families that include orthologs from chicken, *Xenopus*, five fish species, and at least four mammalian species (Fig. 1). Within this data set (884 gene families), we have distinguished strict one to one orthologs, with no duplication detected ("singletons"), and paralogs from the fish-specific whole genome duplication. We tested three to five internal branches, chosen because of their biological relevance and because they separate at least four sequences on each side, giving us sufficient power for statistical testing.

We use a branch-site model of positive selection for which the branch to be tested needs to be specified a priori (Yang 1998; Yang and Nielsen 2002). It is also possible in the absence of a specific biological hypothesis to use this test to scan for positive selection over several branches, on condition of correcting for multiple testing (Anisimova and Yang 2007). We use the q-value method to control for false discovery rates (Storey and Tibshirani 2003), as it is well adapted to large numbers of tests and has been shown to be powerful when applied to the branch-site test of positive selection (Anisimova and Yang 2007). To evaluate the specificity and power of this methodology, we performed simulations that reproduce the original data as closely as possible (Table 1). We obtain 59%–99.7% of true positives when positive selection is simulated and 0%–8.5% of false positives on simulations with no positive selection. Using this test on the real data set, we find that positive selection has affected most genes during bony vertebrate evolution: 77% with the commonly used $q = 10\%$ threshold of false positives, and still 45% with a stringent $q = 1\%$ (Table 2). In the following analyses, we use $q = 10\%$ ($P \leq 0.078$), as it provides more data, and trends in all parameters considered are consistent if we use the more restrictive cut-off. The complete data, including all *P*- and *q*-values, are available as Supplemental material (Supplemental Table 1).

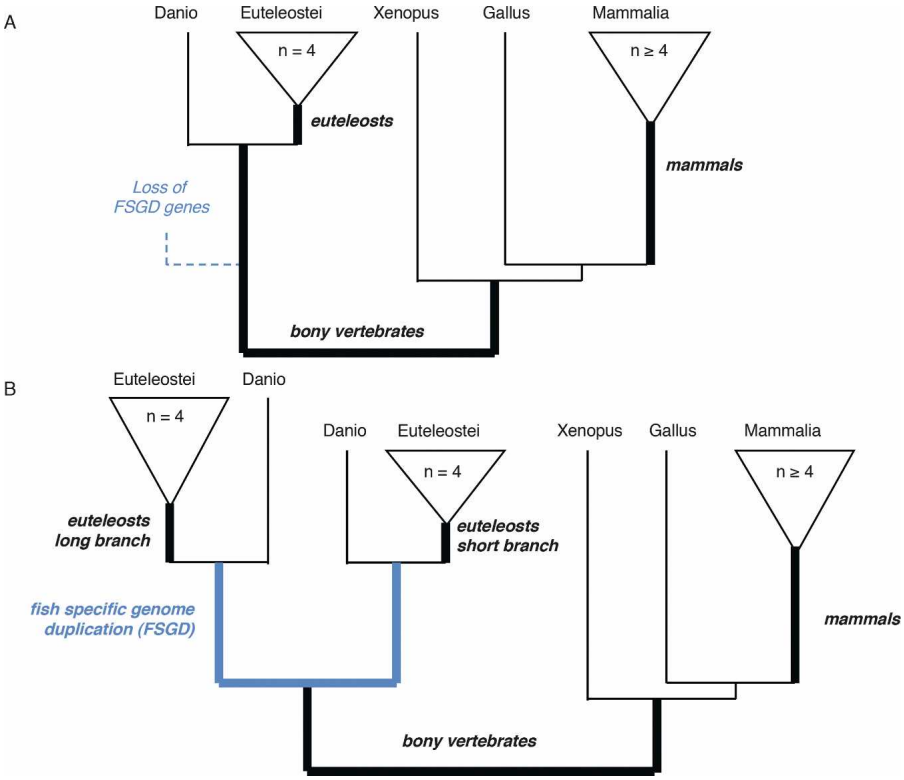

Figure 1. Tree topologies studied. Schematic representation of the tree topologies selected. (Black) speciation branches; (blue) duplication branches. Branches in bold were used as "foreground branches" in branch-site tests for positive selection. Bold italics indicate branch names used in the text and the tables. (*A*) "Singleton" tree type, with no duplication allowed. The dotted line represents gene loss after whole genome duplication. (*B*) "Fish-specific whole genome duplication" tree type. Duplication on other branches may be excluded or allowed. Because differences in d_N/d_S ratio have been reported between paralogs (Brunet et al. 2006; Scannell and Wolfe 2007), with higher d_N/d_S in the fast evolving paralog, we distinguish the longer and shorter branch after duplication, based on the PhyML protein phylogeny.

Detection of positive selection on 77% of genes corresponds to only 45% of phylogenetic branches tested. Most genes appear to have experienced positive selection during some periods of their evolution, but not during others. Evidence for positive selection is not evenly distributed across the branches we tested but is higher on the longer branches: The more evolutionary change accumulates, the greater the chance of detecting an episode of positive selection. Thus the most positive selection in our data set is detected for the divergence of tetrapodes and teleost fishes (~352 million years [Myr] of cumulated evolution); the less, for the base of the mammals (~150 Myr) (Fig. 1). Over all branches tested, there is a significant correlation between branch length (in amino acid substitutions/site) and the result of the likelihood test for positive selection (Spearman's $\rho = -0.37$, $P < 2.2 \times 10^{-16}$).

For branches on which positive selection is detected, it only concerns a minority of sites (Table 3). The mean over all significant branches gives 5.6% of sites under positive selection. Assuming that zero sites are positively selected when the test is not significant and weighting by gene length, we obtain a mean of 2.7% sites under positive selection (30,210 sites under positive selection/1,129,328 sites analyzed). This number corresponds to an "average branch," but the number of sites predicted to be under positive selection varies among branches (Table 3). Computing the mean for each branch separately, we obtain between 0.9% for the mammalian branch and 4.7% for the bony vertebrate branch. It is difficult to compute the proportion of sites that have been under positive selection over all of vertebrate evolution, since (1) we do not test all possible branches; (2) our data provide insufficient power to identify all the specific sites under positive selection, so we cannot determine whether the same sites are repeatedly under selection or not; and (3) the bony vertebrate branch is unrooted, thus combining selection that occurred in the ancestor of tetrapodes and in the ancestor of teleost fishes. But while we cannot from these data give a specific number of sites affected by positive selection overall, it is clear that it is a small proportion during each evolutionary period tested.

We tested for the influence of saturation of synonymous sites in two ways: by removing extreme data points and by simulation. First, we removed all genes with at least one branch of $d_S \geq 2$ synonymous substitutions/codon. This leads to removal of 237 genes (25% of the total), but the proportion of genes with positive selection detected is almost unchanged, at 79%; results per branch are also not affected (data not shown). Repeating this with a limit of $d_S \geq 1$ leads to removal of 88% of genes, but the proportions are again hardly changed, with positive selection detected on 77% of genes left (data not shown).

Second, we performed simulations that reproduce the distribution of gene families from the real data set (see Methods). Reassuringly, we recover the simulated total tree length with good accuracy, both in d_N (mean absolute value of error 14%) and in d_S (mean absolute value of error 13%). This stands in contrast to saturation problems if we were to use pairwise comparisons instead of computation over the whole tree. For example, on the simulated stickleback–rat data corresponding to ENSGACG00000006060 and ENSRNOG00000014240 (family HBG000007 of Supplemental Table 1), the simulated "true" d_S of 4.158 is estimated at 16.2 by pairwise comparison. But a whole tree analysis yields a value of 4.155. The longest single branch in this case has a true value of 1.52 and is estimated at 1.51. Thus the use of more sequences does help to break long branches with saturation issues. Moreover, if the results were due in large part to saturation of d_S, we expect spurious detection of positive selection on long branches. This is not the case, with only 0.6% of the 2673 branches tested significant for positive selection ($q = 10\%$) (Table 1), when simulated without positive selection. To investigate further the effect of potential saturation, we conducted additional simulations with modified branch lengths. Dividing all branch lengths by 2 led to zero branches significant for positive selection (data not shown), unsurprisingly. Multiplying all branch lengths by 1.5 (data not shown) or 2 (Table 1) led to 2.0% and 2.9%, respectively, of branches significant for positive selec-

Table 1. Evaluation of the accuracy and power of the likelihood tests for positive selection using simulated data

Tree type[a]	No. of trees	Foreground branch[a]	Significant LRT[b], q = 0.10			
			Nearly Neutral model		Branch site positive selection model	
			Original branch lengths	Branch lengths × 2	$\omega_2 = 4$	$\omega_2 = 9$
Singleton	767	Bony vertebrates	15 (2.0%)	53 (6.9%)	453 (59%)	718 (94%)
		Mammals	0 (0%)	2 (0.3%)	496 (65%)	723 (94%)
		Euteleosts	0 (0%)	14 (1.8%)	522 (68%)	736 (96%)
Fish-specific whole-genome duplication						
No other duplication	46	Bony vertebrates	1 (2.2%)	0 (0%)	30 (65%)	42 (91%)
		Mammals	0 (0%)	0 (0%)	30 (65%)	43 (93%)
		Euteleosts short branch	0 (0%)	1 (2.2%)	29 (63%)	45 (98%)
		Euteleosts long branch	0 (0%)	0 (0%)	34 (74%)	46 (100%)
		FSGD	0 (0%)	0 (0%)	29 (63%)	45 (98%)
Some other duplication(s)	71	Bony vertebrates	0 (0%)	5 (7.0%)	44 (62%)	65 (92%)
		FSGD	1 (1.4%)	2 (2.8%)	45 (63%)	65 (92%)
All	884	Bony vertebrates	16 (1.8%)	58 (6.6%)	527 (60%)	825 (93%)
	813	Mammals	0 (0%)	2 (0.2%)	526 (65%)	766 (94%)
	859	Euteleosts	0 (0%)	15 (1.7%)	585 (68%)	827 (96%)
	117	FSGD	1 (0.8%)	2 (1.7%)	74 (63%)	110 (94%)
Total overall branches N = 2673 branches			17 (0.6%)	77 (2.9%)	1712 (64%)	2528 (95%)
Total overall trees[c]	884		17 (1.9%)	75 (8.5%)	800 (90%)	881 (99.7%)

Data simulated under the same global parameters as the real data, with sites evolving under a mix of purifying selection and neutrality (Nearly Neutral model); plus sites under positive selection on the foreground branch for the positive selection model. Simulations under the Nearly Neutral model with original branch lengths divided by 2 are not shown, because 0% of tests are significant for all branches.
[a]Classification of topologies and branches following Figure 1. FSGD, fish-specific genome duplication.
[b]Number of trees where the model A with foreground $\omega_2 > 1$ is significantly more likely than the model A with $\omega_2 = 1$, with multiple test correction by the q-value of Storey and Tibshirani (2003). (LRT) Likelihood ratio test.
[c]The number of trees with at least one branch where positive selection was significant.

tion, largely under the accepted q = 10% of false positives. To also verify the power of the branch site test on such divergent sequences, we added 6% of sites under positive selection on one branch at a time. With $\omega = 4$ (as in Anisimova and Yang 2007), 64% of the branches tested were found significant. With $\omega = 9$, which is our observed median value, 95% of the branches tested were found significant (Table 1). Thus it appears that the test used is robust to large sequence divergence, and our results are not due to d_S saturation.

An issue that is not covered by the simulations is the impact of alignment uncertainty (Landan and Graur 2007; Wong et al. 2008). We expect this impact to be limited in our study because

Table 2. Number of genes for which positive selection is detected

Tree type[a]	No. of trees	Foreground branch[a]	Significant LRT[b]	
			q = 0.10	q = 0.01
Singleton	767	Bony vertebrates	451 (59%)	240 (31%)
		Mammals	176 (23%)	62 (8%)
		Euteleosts	397 (52%)	195 (25%)
Fish-specific whole-genome duplication				
No other duplication	46	Bony vertebrates	27 (59%)	12 (26%)
		Mammals	12 (26%)	6 (13%)
		Euteleosts short branch	23 (50%)	8 (17%)
		Euteleosts long branch	21 (46%)	11 (24%)
		FSGD	14 (30%)	3 (6.5%)
Some other duplication(s)	71	Bony vertebrates	44 (62%)	20 (28%)
		FSGD	28 (39%)	10 (14%)
All	884	Bony vertebrates	522 (59%)	272 (31%)
	813	Mammals	188 (23%)	68 (8.4%)
	859	Euteleosts	441 (51%)	214 (25%)
	117	FSGD	42 (36%)	13 (11%)
Total overall branches N = 2673 branches			1193 (45%)	567 (21%)
Total overall trees[c]	884		677 (77%)	401 (45%)

[a]Classification of topologies and branches following that in Figure 1. FSGD, fish-specific genome duplication.
[b]Number of trees where the model A with foreground $\omega_2 > 1$ is significantly more likely than the model A with $\omega_2 = 1$, with multiple test correction by the q-value of Storey and Tibshirani (2003) over all tests performed. (LRT) Likelihood ratio test.
[c]The number of trees with at least one branch where positive selection was significant.

Table 3. Substitution parameters for branches with positive selection

Tree type[a]	Foreground branch[a]	No. of branches[b]	Mean length	Mean ω_0[c]	Sites background $\omega_0 < 1$		Sites background $\omega_1 = 1$	
					Percent of sites foreground $\omega_0 < 1$	Percent of sites foreground $\omega_2 > 1$	Percent of sites foreground $\omega_1 = 1$	Percent of sites foreground $\omega_2 > 1$
Singleton	Bony vertebrates	451	0.145	0.104	82%	6.5%	11%	1%
	Mammals	176	0.090	0.102	86%	3%	11%	0.5%
	Euteleosts	397	0.121	0.103	85%	3.5%	11%	0.5%
Fish-specific whole-genome duplication								
No other duplication	Bony vertebrates	27	0.077	0.080	86%	5.2%	8.4%	0.6%
	Mammals	12	0.118	0.072	88%	4.6%	7.1%	0.4%
	Euteleosts short branch	23	0.049	0.067	91%	2.3%	6.1%	0.2%
	Euteleosts long branch	21	0.140	0.073	87%	5.3%	7.2%	0.5%
	FSGD	14	0.059	0.075	86%	4.8%	8.3%	0.7%
Some other duplication(s)	Bony vertebrates	44	0.108	0.093	83%	7.8%	8.5%	0.7%
	FSGD	28	0.098	0.099	82%	8.3%	8.7%	1%

[a]Classification of topologies and branches following that in Figure 1. FSGD, fish-specific genome duplication.
[b]Number of cases where the model A with foreground $\omega_2 > 1$ is significantly more likely than the model A with $\omega_2 = 1$.
[c]ω_0, the purifying selection pressure, has the same value for foreground and background branches; only the proportion of sites under ω_0 varies.

(1) we selected very conserved gene families, (2) we realigned selected sequences using one of the best available multiple alignment algorithms, and (3) the detection of positive selection is done using only columns without gaps. To verify this, we performed two controls. For each control, results were recomputed on the "bony vertebrate" branch, the most divergent of our study. First, we realigned selected sequences with MAFFT (Katoh and Toh 2008). Results are strongly correlated between the MAFFT and MUSCLE alignments (correlation of ΔLnL values: $r = 0.87$, $P < 10^{-16}$). The significance of the test changes for only 6% of genes. Second, we filtered both MUSCLE and MAFFT alignments with Gblocks (Castresana 2000) to exclude poorly aligned sites from the computations. On average, only 4.4% of sites are excluded. The correlation between results with and without Gblocks is high (MUSCLE: $r = 0.88$; MAFFT: $r = 0.94$), with changes in significance for 7% of genes based on either MUSCLE or MAFFT alignments. Changes in significance occur in both directions, so that the proportion of genes with significant positive selection on the "bony vertebrate" branch is approximately the same (59%–64%) with MUSCLE or MAFFT, with or without Gblocks. In summary, alignment errors do not appear to have a major impact on our results.

What can these results teach us about the occurrence of positive selection? First, we contrasted the Gene Ontology and Panther ontology categories of genes that show positive selection on a given branch, to those that do not. Although the usual suspects of positive selection are slightly overrepresented (e.g., GO terms "response to external stimulus," "defense response," "immune system process," Panther terms "signal transduction," "immunity and defense"), no term varies significantly according to positive selection, on any branch (for GO, all adjusted *P*-values = 1; for Panther, all *q*-values > 0.53; data not shown). Second, we contrasted the categories of our data set to the complete set of genes from the human genome. This tests whether our procedure of data selection may have introduced a functional bias leading to excess detection of positive selection. On the contrary, our data set appears enriched in genes implicated in core cellular and physiological processes (Supplemental Tables 2, 3), and depleted in categories most often reported as targets of positive selection (Vallender and Lahn 2004; Yang 2006). This is consistent with our relatively conservative selection procedure and indicates that positive selection in vertebrates is not restricted to a small subset of fast evolving genes.

Another surprising observation is that the branches following the fish-specific whole genome duplication do not stand out in our results. For each parameter computed, they are within the range observed for the other branches tested (Tables 2–4). We searched for weaker but potentially significant effects, first by contrasting parameter values for the duplication branches ("FSGD") to those for all other branches, on the same trees (Table 5; column FSGD vs. Other Branches). The only significant effect is weaker purifying selection in some cases. Next, we contrasted speciation branches that follow a duplication event to the same speciation branches without a prior duplication (Table 5) to quantify longer term effects. In addition to the fish-specific genome duplication, we verified for such effects due to the whole genome duplications at the origin of vertebrates (2R in Table 5). Consistent with the previous results, most comparisons are not significant. It should be noted that the "singleton" branches include duplication followed by loss, as well as cases where one paralog was not detected by the methods used. The only significant differences indicate more purifying selection for genes kept in double after the duplication, consistent with known biased retention (Davis and Petrov 2004; Brunet et al. 2006). If there is more positive selection after duplication, as expected for neofunctionalization of sequences, we expect a better fit of the positive selection model. This should translate into higher ΔLnL values, the difference in log likelihood between the models with and without positive selection. But these values never differ significantly according to the presence or absence of duplication, on or before the branches tested, thus providing no evidence for protein sequence neofunctionalization.

Table 4. Substitution parameters for branches without positive selection

Tree type[a]	Foreground branch[a]	No. of branches[b]	Mean length	Mean ω_0[c]	Sites background $\omega_0 < 1$: Percent of sites foreground $\omega_0 < 1$	Sites background $\omega_0 < 1$: Percent of sites foreground $\omega_2 = 1$	Percent of sites background $\omega_1 = 1$[d]
Singleton	Bony vertebrates	316	0.112	0.077	85%	7.6%	7.2%
	Mammals	591	0.062	0.090	84%	6%	10%
	Euteleosts	370	0.077	0.081	87%	3.9%	8.9%
Fish-specific whole-genome duplication							
No other duplication	Bony vertebrates	19	0.058	0.064	88%	5.8%	5.8%
	Mammals	34	0.046	0.074	88%	3.8%	8%
	Euteleosts short branch	23	0.069	0.079	88%	2.8%	9.5%
	Euteleosts long branch	25	0.130	0.073	78%	14%	8.1%
	FSGD	32	0.067	0.072	76%	15%	8.5%
Some other duplication(s)	Bony vertebrates	27	0.0852	0.069	78%	14%	8%
	FSGD	43	0.074	0.093	82%	9.7%	8.5%

[a]Classification of topologies and branches following that in Figure 1. FSGD indicates fish-specific genome duplication.
[b]Number of cases where the model A with foreground $\omega_2 = 1$ is not significantly less likely than the model A with $\omega_2 > 1$.
[c]ω_0, the purifying selection pressure, has the same value for foreground and background sites; only the proportion of sites under ω_0 varies.
[d]Corresponding foreground sites are all $\omega_1 = \omega_2 = 1$.

Discussion

Detection of positive selection in vertebrate evolution

This is to our knowledge the first such scan for ancient positive selection using the rigorous branch-site model as improved by Zhang et al. (2005). We find that while only 0.9%–4.7% of codons have experienced positive selection on different branches of vertebrate evolution, such episodes of positive selection have affected 77% of genes investigated. This high number is found, although we use a conservative test (e.g., Bakewell et al. 2007) and do not test all branches of the vertebrate tree. It should be noted that the gene data set analyzed here is not representative of the whole genome. Indeed, to avoid misalignments artifacts, we restricted our analyses to gene families for which orthologs could be aligned over at least 80% of their length. Moreover, to limit problems of saturation, we selected only gene families for which orthologs could be identified in at least 11 species (five fishes, *Xenopus*, chicken, and at least four mammals). Hence gene families with frequent gene duplications and losses or with a high rate of amino acid substitution are underrepresented in our data set. Because of this bias toward highly conserved genes, our results are probably an underestimate of the true frequency of genes that are subject to positive selection.

Our results stand in some contrast to reports of rare positive selection in mammals (Endo et al. 1996; Clark et al. 2003; Jorgensen et al. 2005; Zhang and Li 2005; Arbiza et al. 2006; Bakewell et al. 2007; Gibbs et al. 2007) and are more reminiscent of results in lineages with larger population sizes (Nielsen and Yang 2003; Eyre-Walker 2006; Drosophila 12 Genomes Consortium 2007; Sawyer et al. 2007). Interestingly, the most recent *Drosophila* study (Drosophila 12 Genomes Consortium 2007) found 2% of codons under positive selection, a very similar proportion to our observations. Estimates of the proportion of changes driven by positive selection in *Drosophila* vary consider-

Table 5. Influence of duplication on substitution parameters

Branch behavior[a]	Parameter	FSGD vs. other branches[b]: *P*-value[c]	FSGD vs. other branches[b]: Difference[d]	Euteleosts branch: FSGD topology vs. singleton: *P*-value[c]	Euteleosts branch: FSGD topology vs. singleton: Difference[d]	Bony vertebrates branch: 2R detected vs. not detected: *P*-value[c]	Bony vertebrates branch: 2R detected vs. not detected: Difference[d]
LRT significant	ΔLnL	0.14	8.9–10.7	0.57	10–11	0.94	11–11
	Branch length[e]	0.017	0.085–0.11	0.014	0.093–0.12	0.035	0.13–0.14
	Mean ω_0	0.93	0.075–0.075	0.015	0.061–0.075	0.0098	0.072–0.079
	Percent of sites ω_0[f]	0.92	83%–84%	0.0029	89%–85%	0.67	83%–82%
	Percent of sites ω_1[f]	0.27	8.6%–10%	**3.4×10^{-4}**	**6.6%–11%**	0.035	9.3%–11%
	Percent of sites ω_2[f]	0.12	8.1%–5.6%	0.56	4.1%–4.1%	0.075	8.1%–7.2%
LRT nonsignificant	ΔLnL	0.41	0.69–0.65	0.30	0.79–0.65	0.33	0.91–1.0
	Branch length[e]	0.43	0.082–0.069	0.46	0.10–0.077	0.018	0.095–0.12
	Mean ω_0	0.13	0.054–0.062	0.65	0.061–0.061	**3.8×10^{-6}**	**0.048–0.064**
	Percent of sites ω_0[f]	**0.00026**	**80%–86%**	0.28	84%–87%	0.011	88%–83%

Boldface indicates that the difference is significant.
[a]Classification according to whether the LRT for positive selection is significant on each branch ($q = 0.10$).
[b]Only branches from tree topologies for which the FSGD branch exists were used.
[c]Nonpaired Wilcoxon test. In bold if the difference is significant after Bonferroni correction ($\alpha = 0.05/30 = 0.0017$).
[d]Duplication or post-duplication mean − nonduplication mean.
[e]In amino acid substitutions/site.
[f]Values for the foreground branch of each test.
(LRT) Likelihood ratio test.

ably according to methodology (Sawyer et al. 2007; Shapiro et al. 2007).

Most codon model studies in mammals have used relatively few sequences per gene, testing for selection either by pairwise comparison or on a tree with few sequences. Simulations indicate that likelihood tests tend to be overly conservative when few sequences are used (Anisimova et al. 2002; Anisimova and Yang 2007). The design of our tree patterns allowed us to test exclusively internal branches with at least four sequences on each side (Fig. 1). Moreover, we tested longer branches than intramammalian studies. We observe a positive correlation between detection of selection and branch length, which is consistent with previous reports that more positive selection can be detected when longer branches are tested (Anisimova and Yang 2007; Gibbs et al. 2007), as long as saturation is not reached. A few previous reports have illustrated the power of likelihood tests to detect positive selection in ancient vertebrate evolution (e.g., Bielawski and Yang 2004). Saturation of d_S would be problematic in pairwise comparisons of sequences as divergent as human and zebrafish. But our simulations, as well as those of Anisimova and colleagues (Anisimova et al. 2002; Anisimova and Yang 2007), show that the maximum likelihood estimate as we use it is robust to d_S saturation. This appears to be due to the use of more sequences, which break the long branches of the gene tree.

Our results are consistent with a model of recurrent adaptive amino acid substitutions, driven by weak positive selection, as modeled recently in fly (Andolfatto 2007). This model notably predicts more selective substitutions in rapidly evolving genes, which is consistent with the correlation with d_N that we observe. It also predicts that such selection will be difficult to detect in genome scans, as is the case. This model, and simulation results (Anisimova et al. 2002; Anisimova and Yang 2007), indicate that our results are not contradictory to the reports of rare positive selection in mammals. Rather, testing over relatively short time intervals provides evidence for the strongest signals of positive selection only, on a small subset of genes, whereas testing over longer time enables us to detect events which are rarer, but eventually affect most genes. It remains to be seen whether the same sites are repeatedly under positive selection in different lineages or whether different sites are affected. Our data provide insufficient power to test this, as specific sites under weak positive selection are difficult to identify.

Several investigators have noted that a high d_N/d_S ratio can be caused by low d_S, due to local constraints on synonymous substitutions (Pond and Muse 2005; Chamary et al. 2006; Schattner and Diekhans 2006; Mayrose et al. 2007; Parmley and Hurst 2007), while Friedman and Hughes (2007) have also argued for an impact of GC content. To evaluate more in detail the influence of such potential confounding factors on our results, we adjusted a linear model explaining the test results (ΔLnL) for each branch by global characteristics of the tree (Supplemental Table 4). An ANOVA on this model shows that (1) for all branches, the most significant contributors to our results are the number of sites and the d_N value. This confirms that with more amino acid substitutions, more positive selection can be detected. (2) In some cases GC content and d_S contribute significantly to test results, but always explain little variance. (3) More than 60% of variance in test results is explained by none of the global parameters included in the model. Of note, two measures of alignment quality, the proportion of gaps and the number of sites excluded by Gblocks, have little to no effect on test results. The variation in GC content along each branch, which could indicate changes in codon usage or in recombination rates, also has no effect. In contrast to a report of bias of d_N/d_S tests (Wyckoff et al. 2005), we find the expected weak negative correlation of global ω with d_S length of the tree ($r = -0.19$, $P = 1.1 \times 10^{-8}$), and strong positive correlation with d_N ($r = 0.82$, $P < 2.2 \times 10^{-16}$). Excluding genes with at least one very high d_S branch did not change results. Thus, as far as we can tell, we seem to be effectively detecting branch-site specific positive selection, not some bias of the alignment or the tree.

Two functional categories of genes tend to be overrepresented in reports of positive selection (Vallender and Lahn 2004; Yang 2006): genes involved in host defense and immunity or in evading these defenses, and genes involved in sexual reproduction. Such trends have been confirmed in several genomic scans for positive selection, notably in primates, where they typically also include neuronal function and perception (Bustamante et al. 2005; Nielsen et al. 2005; Biswas and Akey 2006; Voight et al. 2006; Wang et al. 2006; Gibbs et al. 2007). This seems contradictory with our results: Positively selected genes do not differ in functional categories from other genes, while our total sample is in fact biased toward basic cellular processes (Supplemental Tables 2, 3). First, we note that in some previous studies, functional categories such as metabolism genes were reported as under positive selection (Roth and Liberles 2006; Voight et al. 2006; Petersen et al. 2007). Second, different studies may measure different selection modes. Polymorphism studies in primates typically report genes that are under recent selection, whereas we have scanned for selection in more ancient vertebrate evolution. The branch-site test is not intended to detect continuous selective pressure, which would likely characterize arms race genes. Moreover, most interspecific studies have used less stringent model testing and may report as being under positive selection genes that are under weak purifying selection (Zhang et al. 2005; Bakewell et al. 2007). Indeed, the most stringent primate study so far found positive selection mostly in genes involved in basic cellular functions (Bakewell et al. 2007). Finally, in a recent study of *Drosophila* genomes (Drosophila 12 Genomes Consortium 2007), positive selection was found to affect all functional categories, albeit more strongly "defense response." Similarly, our results do not exclude that positive selection be strongest on fast evolving (e.g., immunity) genes, absent from our data set. But they do indicate that episodes of positive selection affect all categories of genes, in vertebrates as in flies.

The impact of genome duplication on patterns of selection

In addition to the functional categories already discussed, duplicate genes are also overrepresented in reports of positive selection (Yang 2006). There have been several reports of higher d_N/d_S ratios on branches following duplications (Jordan et al. 2004; He and Zhang 2005; Brunet et al. 2006; Byrne and Wolfe 2006; Johnston et al. 2007) and even on branches preceding duplications (Johnston et al. 2007). But these studies used global measures of d_N/d_S, which do not distinguish between relaxed negative selection, and positive selection on a few sites. In the analysis of the macaque genome, an excess of duplicate genes among those with positive selection was noted, but on quite a small sample (Gibbs et al. 2007). Thus this study is to our knowledge the first large-scale quantification of positive selection after duplication. By using only duplicates from whole genome duplication, we constrain the duplication branches to represent the same divergence time. And by contrasting branches of the same

gene tree, we control for biased retention of duplicates (Davis and Petrov 2004; Brunet et al. 2006).

The result of this careful testing is striking: The substitution parameters after duplication differ very little, if at all, from those after speciation (Table 5). The few significant differences are consistent with previous reports of biased retention of genes under stronger purifying selection, after whole genome duplication (Davis and Petrov 2004; Brunet et al. 2006), and of some relaxation of purifying selection, rather than with any increase in positive selection. Of note, the faster evolving paralog does not show evidence of more positive selection (Table 2), as would be expected if the asymmetry were due to neofunctionalization at the protein level (Brunet et al. 2006; Byrne and Wolfe 2006). This result shows the importance of controlling with speciation branches before attributing an effect to duplication. A study solely conducted on the duplication branch might have concluded erroneously that 36% of genes were subject to positive selection because of the duplication, whereas this proportion is not higher than in comparable branches without duplication. We note that previous studies that contrasted duplication and speciation branches, while not explicitly identifying positive selection, also found that differences are more slight than expected (Seoighe et al. 2003; Conant et al. 2007; Hellsten et al. 2007; Johnston et al. 2007; Scannell and Wolfe 2007). The weak difference between protein coding gene evolution after speciation and after duplication may reflect the importance of other levels of function, such as expression, on the divergence of duplicates (Hellsten et al. 2007; Hughes 2007). Indeed, a recent study of whole genome duplication in yeast (Wapinski et al. 2007) found that duplicates diverged mostly in regulation, and much less in biochemical function, which is what we can detect at the level of amino acid sequences.

Conclusion

An important conclusion of this work is that the most stringent test does detect positive selection in significant amounts in vertebrate evolution. Thus, adaptive evolution at the molecular level does appear significant (Hoekstra and Coyne 2007), although we note that our results are in no way exclusive of functional evolution by regulatory mutations (Hughes 2007; Nei 2007; Prud'homme et al. 2007). Our results are supportive of a model of widespread but transient positive selection. Finally, we do not find a large difference of evolutionary modes of protein coding sequences after duplication, relative to speciation. It remains to be tested whether this is due to divergence at other levels or to a lesser impact of duplication on gene evolution than expected.

Methods

Data

Gene families were obtained from the database HOMOLENS version 3 (http://pbil.univ-lyon1.fr/databases/homolens.html), which is based on Ensembl release 41 (October 2006) (Hubbard et al. 2007). HOMOLENS is built on the same model as HOVERGEN (Duret et al. 1994) or HOBACGEN (Perriere et al. 2000), with genes organized in families, which include precalculated alignments and phylogenies. In HomolEns version 3, alignments are computed with MUSCLE (Edgar 2004) (with default parameters), and phylogenetic trees with PHYML (substitution model = JTT, estimated proportion of invariable sites, four categories, estimated gamma, initial tree with BIONJ) (Guindon and Gascuel 2003). Phylogenies are computed on conserved blocks of the alignments selected with Gblocks (Castresana 2000).

Using the TreePattern functionality of the FamFetch client for HOMOLENS, which allows scanning for gene tree topologies (Dufayard et al. 2005), we selected three sets (Fig. 1): (1) a set of "singleton" genes, where duplication is strictly forbidden along the tree; (2) a strict set of "fish-specific genome duplication" genes, where paralogs are retained in all fishes after the whole genome duplication but other duplication is forbidden; and (3) a relaxed set of "fish-specific genome duplication" genes, where paralogs are retained in all fishes after the whole genome duplication and other duplication is allowed. In all cases, we imposed that all five fishes, the *Xenopus*, the chicken, and at least four mammals be represented in the tree. In addition, to clarify an eventual effect of older whole genome duplications at the origin of vertebrates (known as 2R for two rounds of duplication), we selected all genes with duplications specific to vertebrates, predating the teleost fish–tetrapode split. This allowed us to define the subset of genes in our data that were kept in duplicate after these older genome duplications.

For the families thus recovered, we restricted alignments and trees to the selected phylogenetic pattern, notably excluding more distant paralogs (e.g., from duplications basic to vertebrates). We removed species with low genome coverage. 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. DNA alignments are calculated from the protein alignments, with RevTrans (Wernersson and Pedersen 2003). To evaluate the impact of alignment uncertainty, the restricted alignments were also refined with MAFFT (Katoh and Toh 2008), and high-quality alignments were selected from both MUSCLE and MAFFT alignments using Gblocks (type = codons) (Castresana 2000). For the manipulations of sequences and trees, we combined scripts in Python, BioPython, Jalview (Clamp et al. 2004), and the R library APE (Paradis et al. 2004).

Our data set includes 10 species of tetrapodes: the frog *Xenopus tropicalis* (DoE Joint Genome Institute, unpubl.); the chicken *Gallus gallus* (International Chicken Genome Sequencing Consortium 2004); the seven mammals *Monodelphis domestica* (Mikkelsen et al. 2007), *Bos taurus* (HGSC at Baylor College of Medicine, unpubl.), *Canis familiaris* (Lindblad-Toh et al. 2005), *Mus musculus* (Waterston et al. 2002), *Rattus norvegicus* (Rat Genome Sequencing Project Consortium 2004), *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); and five species of teleost fishes: the zebrafish *Danio rerio* (Zebrafish Sequencing Group at the Sanger Institute, unpubl.); and the four euteleosts *Gasterosteus aculeatus* (The Broad Institute, unpubl.), *Oryzias latipes* (Kasahara et al. 2007), *Tetraodon nigroviridis* (Jaillon et al. 2004), and *Takifugu rubripes* (Aparicio et al. 2002).

All alignments and trees can be viewed and downloaded at http://bioinfo.unil.ch/supdata/.

Detection of positive selection

We used the branch-site model A (Yang and Nielsen 2002; Zhang et al. 2005), which allows to detect positive selection that acts on a subset of sites in a specific lineage. Positive selection is detected by a d_N/d_S ratio $\omega > 1$. This model has been reported to be more sensitive for the detection of positive selection than previous models (Yang 2006), such as branch models (Yang 1998) or site models (Yang et al. 2000).

The application of this model necessitates providing a phylogenetic tree and defining a priori the branch we want to test for positive selection. This branch is called the foreground branch

(Fig. 1, bold), where positive selection may be allowed. All other branches in the tree represent the background branches, where sites are only allowed to evolve under purifying or neutral selection. The original formulation (Yang and Nielsen 2002) compared this branch-site model A (alternative hypothesis) to the Neutral site model M1 (null hypothesis). The problem in this test is that it does not discriminate between positive selection and relaxation of purifying selection (Zhang 2004). To avoid this problem of false positives, Zhang et al. (2005) proposed a stricter test, which contrasts the branch-site model A with $\omega_2 \geq 1$ (alternative hypothesis) to the model A with $\omega_2 = 1$ fixed (null hypothesis). The test is done by comparing the difference of likelihood values $2 \times \Delta$LnL to a χ^2 distribution of 1 degree of freedom. This test has been reported to be very conservative (e.g., Bakewell et al. 2007), and it is the only one we used in our analysis. Of note, the model distinguishes two components of the positively selected set of sites: sites that are under purifying selection in the rest of the tree, and sites that are neutral in the rest of the tree. All computations are done using CODEML from the PAML package (v3.15) (Yang 1997).

We use the value $2 \times \Delta$LnL as the best measure of outcome of the test for positive selection in all subsequent statistical analyses.

Statistical analysis

We used the web server FatiGO+ to perform statistical analysis on Gene Ontology terms (Al-Shahrour et al. 2007), with FDR correction for multiple testing (Benjamini and Hochberg 1995). We also used the PANTHER Classification System (Thomas et al. 2003), followed by QVALUE correction for multiple testing (A. Dabney and J.D. Storey, unpubl.).

To correct for CODEML testing on multiple branches in multiple phylogenetic trees, we control for false discovery by using the *q*-value (Storey and Tibshirani 2003). All *P*-values from our likelihood ratio tests were treated as one series of repetitions (m branches $\times$ n trees). Our *P*-values follow a bimodal distribution, because it is not rare that the alternative brings no improvement over the null hypothesis (thus $P = 1$), while in many other cases the difference of likelihoods is large (thus $P \approx 0$). As recommended for a bimodal distribution (documentation of the QVALUE library), we used the *bootstrap* method for estimating π_0 in the R package QVALUE (A. Dabney and J.D. Storey, unpubl.).

We used nhPhyml (topology fixed, transition/transversion [Ts/Tv] ratio estimated, alpha parameter estimated with four categories, GC equilibrium frequency optimized for each branch) (Boussau and Gouy 2006) to estimate the GC rate at third codon positions at each node of the phylogenetic trees. We then computed the ΔGC for each branch of interest, as the difference between GC at the nodes bracketing that branch.

All other statistical analyses were performed using R (R Development Core Team 2007).

Simulations

Simulated nucleotide alignments were generated using Evolver, from the PAML package (v3.15) (Yang 1997).

To test accuracy, we generated alignments under the null hypothesis of no positive selection. We used the Nearly Neutral model (M1a), allowing sites to be under purifying selection or neutral evolution. For each real data set, a simulated data set was generated using the same global parameters as the real data: number of sequences, sequence length, tree topology, branch lengths (defined as the number of nucleotide substitutions per codon), d_N/d_S ratio ω, Ts/Tv ratio κ, and codon usage. This procedure guaranties that the simulated data set has the same distribution of parameters as the real data set, including potential confounding factors such as codon usage or long branches. In addition, to control for the effect of potential underestimation of d_S (saturation), we conducted simulations modifying the total length of the tree. Dividing all branch lengths by 2 provides an estimate of the behavior of the test with less divergent sequences, while multiplying by 1.5 or 2 provides estimates of the behavior of the test with even more divergent sequences. Alternatively, the latter simulations could correct for underestimation of branch lengths in the original analysis.

To test power, we generated alignments using the same procedure, plus specifying branch-site–specific positive selection on one branch. Thus we performed as many simulations for each data set, as branches tested. In accordance with our results (Table 3), we simulated nucleotide alignments with 84% of sites under purifying selection (ω_0), 10% of sites under neutral evolution (ω_1) and 6% of sites under positive selection on the foreground branch (ω_2). Values of $\omega_2 = 4$ and $\omega_2 = 9$ were chosen, following, respectively, the method of Anisimova and Yang (2007), and the median value observed in our data.

Acknowledgments

We thank Darlene Goldstein, Jérôme Goudet, Tal Pupko, Nicolas Salamin, and Ken Wolfe for helpful discussions. We also thank Adam Eyre-Walker and anonymous reviewers for insightful remarks. R.S. and M.R.R. acknowledge funding from Etat de Vaud and Swiss National Science Foundation grant 116798, and L.D. from the Agence Nationale de la Recherche (GIP ANR JC05_49162). We thank the VITAL-IT project of the Swiss Institute of Bioinformatics for providing the computational resources.

References

Al-Shahrour, F., Minguez, P., Tarraga, J., Medina, I., Alloza, E., Montaner, D., and Dopazo, J. 2007. FatiGO +: A functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res.* **35:** W91–W96.

Andolfatto, P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res.* **17:** 1755–1762.

Anisimova, M. and Yang, Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol. Biol. Evol.* **24:** 1219–1228.

Anisimova, M., Bielawski, J.P., and Yang, Z. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol. Biol. Evol.* **19:** 950–958.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.-m., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297:** 1301–1310.

Arbiza, L., Dopazo, J., and Dopazo, H. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput. Biol.* **2:** e38. doi: 10.1371/journal.pcbi.0020038.

Bakewell, M.A., Shi, P., and Zhang, J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc. Natl. Acad. Sci.* **104:** 7489–7494.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. [Ser A]* **57:** 289–300.

Benton, M.J. and Donoghue, P.C.J. 2006. Paleontological evidence to date the tree of life. *Mol. Biol. Evol.* **24:** 26–53.

Bielawski, J.P. and 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.

Biswas, S. and Akey, J.M. 2006. Genomic insights into positive selection. *Trends Genet.* **22:** 437–446.

Boussau, B. and Gouy, M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst. Biol.* **55:** 756–768.

Brunet, F.G., Crollius, H.R., Paris, M., Aury, J.M., Gibert, P., Jaillon, O.,

Laudet, V., and Robinson-Rechavi, M. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.* **23:** 1808–1816.
Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Todd Hubisz, M., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437:** 1153–1157.
Byrne, K.P. and Wolfe, K.H. 2006. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics* **175:** 1341–1350.
Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17:** 540–552.
Chamary, J.V., Parmley, J.L., and Hurst, L.D. 2006. Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7:** 98–108.
The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.
Clamp, M., Cuff, J., Searle, S.M., and Barton, G.J. 2004. The Jalview Java alignment editor. *Bioinformatics* **20:** 426–427.
Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P.D., Kejariwal, A., Todd, M.A., Tanenbaum, D.M., Civello, D., Lu, F., Murphy, B., et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302:** 1960–1963.
Conant, G.C., Wagner, G.P., and Stadler, P.F. 2007. Modeling amino acid substitution patterns in orthologous and paralogous genes. *Mol. Phylogenet. Evol.* **42:** 298–307.
Davis, J.C. and Petrov, D.A. 2004. Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol.* **2:** e55. doi: 10.1371/journal.pbio.0020055.
Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:** 203–218.
Dufayard, J.F., Duret, L., Penel, S., Gouy, M., Rechenmann, F., and Perriere, G. 2005. Tree pattern matching in phylogenetic trees: Automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* **21:** 2596–2603.
Duret, L., Mouchiroud, D., and Gouy, M. 1994. HOVERGEN: A database of homologous vertebrate genes. *Nucleic Acids Res.* **22:** 2360–2365.
Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32:** 1792–1797.
Endo, T., Ikeo, K., and Gojobori, T. 1996. Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* **13:** 685–690.
Eyre-Walker, A. 2006. The genomic rate of adaptive evolution. *Trends Ecol. Evol.* **21:** 569–575.
Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.
Friedman, R. and Hughes, A.L. 2007. Likelihood-ratio tests for positive selection of human and mouse duplicate genes reveal nonconservative and anomalous properties of widely used methods. *Mol. Phylogenet. Evol.* **42:** 388–393.
Gibbs, R.A., Rogers, J., Katze, M.G., Bumgarner, R., Weinstock, G.M., Mardis, E.R., Remington, K.A., Strausberg, R.L., Venter, J.C., Wilson, R.K., et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316:** 222–234.
Gojobori, J., Tang, H., Akey, J.M., and Wu, C.-I. 2007. Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. *Proc. Natl. Acad. Sci.* **104:** 3907–3912.
Guindon, S. and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52:** 696–704.
He, X. and Zhang, J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169:** 1157–1164.
Hellsten, U., Khokha, M.K., Grammer, T.C., Harland, R.M., Richardson, P., and Rokhsar, D.S. 2007. Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog *Xenopus laevis*. *BMC Biol.* **5:** 31. doi: 10.1186/1741-7007-5-31.
Hoekstra, H.E. and Coyne, J.A. 2007. The locus of evolution: Evo-Devo and the genetics of adaptation. *Evolution Int. J. Org. Evolution* **61:** 995–1016.
Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., et al. 2007. Ensembl 2007. *Nucleic Acids Res.* **35:** D610–D617.
Hughes, A.L. 2007. Looking for Darwin in all the wrong places: The misguided quest for positive selection at the nucleotide sequence level. *Heredity* **99:** 364–373.
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, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431:** 946–957.
Johnston, C.R., O'Dushlaine, C., Fitzpatrick, D.A., Edwards, R.J., and Shields, D.C. 2007. Evaluation of whether accelerated protein evolution in chordates has occurred before, after, or simultaneously with gene duplication. *Mol. Biol. Evol.* **24:** 315–323.
Jordan, I.K., Wolf, Y.I., and Koonin, E.V. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol. Biol.* **4:** 22. doi: 10.1186/1471-2148-4-22.
Jorgensen, F., Hobolth, A., Hornshoj, H., Bendixen, C., Fredholm, M., and Schierup, M. 2005. Comparative analysis of protein coding sequences from human, mouse and the domesticated pig. *BMC Biol.* **3:** 2. doi: 10.1186/1741-7007-3-2.
Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447:** 714–719.
Katoh, K. and Toh, H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9:** 286–298.
Landan, G. and Graur, D. 2007. Heads or tails: A simple reliability check for multiple sequence alignments. *Mol. Biol. Evol.* **24:** 1380–1383.
Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., Zody, M.C., et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438:** 803–819.
Lynch, M., O'Hely, M., Walsh, B., and Force, A. 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics* **159:** 1789–1804.
Mayrose, I., Doron-Faigenboim, A., Bacharach, E., and Pupko, T. 2007. Towards realistic codon models: Among site variability and dependency of synonymous and non-synonymous rates. *Bioinformatics* **23:** i319–i327.
Mikkelsen, T.S., Wakefield, M.J., Aken, B., Amemiya, C.T., Chang, J.L., Duke, S., Garber, M., Gentles, A.J., Goodstadt, L., Heger, A., et al. 2007. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447:** 167–177.
Nei, M. 2007. The new mutation theory of phenotypic evolution. *Proc. Natl. Acad. Sci.* **104:** 12235–12242.
Nielsen, R. and Yang, Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol. Biol. Evol.* **20:** 1231–1239.
Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3:** e170. doi: 10.1371/journal.pbio.0030170.
Paradis, E., Claude, J., and Strimmer, K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20:** 289–290.
Parmley, J. and Hurst, L. 2007. How common are intragene windows with $K_A > K_S$ owing to purifying selection on synonymous mutations? *J. Mol. Evol.* **64:** 646–655.
Perriere, G., Duret, L., and Gouy, M. 2000. HOBACGEN: Database system for comparative genomics in bacteria. *Genome Res.* **10:** 379–385.
Petersen, L., Bollback, J.P., Dimmic, M., Hubisz, M., and Nielsen, R. 2007. Genes under positive selection in *Escherichia coli*. *Genome Res.* **17:** 1336–1343.
Pond, S.K. and Muse, S.V. 2005. Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.* **22:** 2375–2385.
Prud'homme, B., Gompel, N., and Carroll, S.B. 2007. Emerging principles of regulatory evolution. *Proc. Natl. Acad. Sci.* **104:** 8605–8612.
Putnam, N.H., Hellsten, U., Yu, J.S., Pennachio, L., Blow, M., Shoguchi, E., Robinson-Rechavi, M., Butts, T., Ferrier, D.E.K., Garcia-Fernàndez, J., et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453:** 1064–1071.
R Development Core Team. 2007. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
Rat Genome Sequencing Project Consortium. 2004. Genome sequence

of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.
Roth, C. and Liberles, D. 2006. A systematic search for positive selection in higher plants (Embryophytes). *BMC Plant Biol.* **6:** 12. doi: 10.1186/1471-2229-6-12.
Sawyer, S.A., Parsch, J., Zhang, Z., and Hartl, D.L. 2007. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proc. Natl. Acad. Sci.* **104:** 6504–6510.
Scannell, D.R. and Wolfe, K.H. 2007. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res.* **18:** 137–147.
Schattner, P. and Diekhans, M. 2006. Regions of extreme synonymous codon selection in mammalian genes. *Nucleic Acids Res.* **34:** 1700–1710.
Seoighe, C., Johnston, C.R., and Shields, D.C. 2003. Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation. *Mol. Biol. Evol.* **20:** 484–490.
Shapiro, J.A., Huang, W., Zhang, C., Hubisz, M.J., Lu, J., Turissini, D.A., Fang, S., Wang, H.-Y., Hudson, R.R., Nielsen, R., et al. 2007. Adaptive genic evolution in the *Drosophila* genomes. *Proc. Natl. Acad. Sci.* **104:** 2271–2276.
Storey, J.D. and Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100:** 9440–9445.
Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. 2003. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* **13:** 2129–2141.
Vallender, E.J. and Lahn, B.T. 2004. Positive selection on the human genome. *Hum. Mol. Genet.* **13:** R245–R254.
Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* **4:** e72. doi: 10.1371/journal.pbio.0040072.
Wang, E.T., Kodama, G., Baldi, P., and Moyzis, R.K. 2006. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc. Natl. Acad. Sci.* **103:** 135–140.
Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449:** 54–61.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Wernersson, R. and Pedersen, A.G. 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.* **31:** 3537–3539.
Wong, K.M., Suchard, M.A., and Huelsenbeck, J.P. 2008. Alignment uncertainty and genomic analysis. *Science* **319:** 473–476.
Wyckoff, G.J., Malcom, C.M., Vallender, E.J., and Lahn, B.T. 2005. A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. *Trends Genet.* **21:** 381–385.
Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.
Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15:** 568–573.
Yang, Z. 2006. *Computational molecular evolution*. Oxford University Press, Oxford, UK.
Yang, Z. and 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., and Pedersen, A.M. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431–449.
Zhang, J. 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol. Biol. Evol.* **21:** 1332–1339.
Zhang, L. and Li, W.-H. 2005. Human SNPs reveal no evidence of frequent positive selection. *Mol. Biol. Evol.* **22:** 2504–2507.
Zhang, J., Nielsen, R., and 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.

Received February 4, 2008; accepted in revised form June 5, 2008.