

The evolutionary fate of recently duplicated retrogenes in mice

Philippe Gayral, Pierre Caminade, Pierre Boursot, Nicolas Galtier

▶ **To cite this version:**

Philippe Gayral, Pierre Caminade, Pierre Boursot, Nicolas Galtier. The evolutionary fate of recently duplicated retrogenes in mice. Journal of Evolutionary Biology, 2007, 20 (2), pp.617-626. 10.1111/j.1420-9101.2006.01245.x . hal-02348108

HAL Id: hal-02348108

https://hal.umontpellier.fr/hal-02348108

Submitted on 27 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

doi:10.1111/j.1420-9101.2006.01245.x

The evolutionary fate of recently duplicated retrogenes in mice

P. GAYRAL,[1] P. CAMINADE, P. BOURSOT & N. GALTIER

CNRS UMR – 'Génome, Populations, Interactions, Adaptation', Université Montpellier, Montpellier, France

Keywords:

duplication;
mouse;
neofunctionalization;
retroposition;
subfunctionalization.

Abstract

Inferences about the evolutionary impact of gene duplications often rely on the analysis of their long-term outcome. The fate of the majority of them must, however, be decided shortly after duplication. Here we analysed the evolutionary pattern of 10 mouse genes very recently duplicated by retrotransposition, by sequencing the retroposed copy in five to 10 closely related mouse species. In all cases the retroposed copy experienced accelerated nonsynonymous evolution whereas the divergence pattern of the source copy appeared unaffected by the duplication, consistent with the neofunctionalization model. The analysis further revealed that most retrogenes, including pseudogenes, did not experience a period of relaxed neutral evolution, but have been submitted to purifying selection ever since their retroposition. We propose that these duplicates play a biochemical role but are not indispensable. Purifying selection prevents them from acquiring a negative role until they are lost or silenced. This period of unnecessary redundancy could in rare cases give the time for new functions to evolve.

Introduction

Gene duplications are thought to be an important process of molecular evolution in eukaryotes. A gene duplication relaxes some of the constraints applying to the genome in which it occurs, opening new evolutionary pathways. One of the two copies can, of course, be quickly lost by deletion or pseudogenization, cancelling the effects of the duplication event. In many instances, however, the two copies are kept and contribute to the organism transcriptome and proteome. Ohno first proposed that gene duplications could promote adaptation. In Ohno's view, one duplicated copy can keep the original function of the gene, meanwhile the other one is free to explore the sequence space, and eventually acquire a new function (Ohno, 1970). This adaptive process was recently named 'neofunctionalization' (NF). Alternatively, Lynch and collaborators, formalizing early thoughts by Piatigorsky & Wistow (1991) and Hughes (1994), proposed that, after a duplication, the two copies could share the relaxed degrees of freedom (Force *et al.*, 1999). Accumulating deleterious mutations, the two duplicates would lose some aspects of the ancestral gene function in a complementary fashion. The two genes would therefore be required to maintain the ancestral function, a process called 'subfunctionalization' (SF, Lynch & Force, 2000). A typical example is the evolution of an ancestral ubiquitous gene (expressed, say, in two tissues) into two complementary tissue-specific genes (each one expressed in one tissue). The SF pathway does not require the occurrence of adaptive changes, that are much rarer than deleterious ones, and could thus more readily account for the relatively high retention rate of duplicated genes in some eukaryotic genomes, especially early vertebrates (Wagner, 1998; Force *et al.*, 1999).

A number of genomic data analyses have been conducted to try and assess the relative importance of NF and SF in molecular evolution. Several studies have focused on evolutionary rates, taken as an indicator of functional constraints. Under the NF model, one would expect a strongly asymmetric pattern in which the adapting duplicate evolves much faster than the one keeping the old function. The SF model, in contrast, would predict a more or less symmetrical evolution if the two copies bore the same amount of constraint. Relative rate analyses of various multigenic family data sets have been contradictory, some studies (Van de Peer *et al.*,

Correspondence: N. Galtier, CNRS UMR 5171 – CC63, Université Montpellier 2, Place E. Bataillon, 34095 Montpellier, France.
Tel.: (+33) 467 14 4684; fax: (+33) 467 14 4554; e-mail: galtier@univ-montp2.fr

[1]*Present address:* UMR 'Biologie et Génétique des Interactions Plantes-Parasites', ENSAM, CIRAD, INRA

2001), but not all (Robinson-Rechavi & Laudet, 2001), reporting significant asymmetry between the two duplicates. For relatively low sequence divergence, the level of constraint applying to a coding sequence can be more directly measured by the ratio of nonsynonymous to synonymous substitution rate (d_N/d_S). An increase of d_N/d_S posterior to duplication was observed in various eukaryotic and prokaryotic genomes (Kondrashov *et al.*, 2002). This result, however, is expected both under NF and SF. Using a large number of recently duplicated ($d_S < 0.3$) genes in human Zhang *et al.* (2003) reported a significant asymmetry of the d_N/d_S ratio for *c.* 25% of the data set. Recently, He & Zhang (2005) approached the NF vs. SF comparison more directly by making use of protein–protein interaction data in yeast. Assuming that SF should reduce the number of shared interactions between duplicates, they suggest that the fate of duplicated genes in yeast is well described by a combination of early SF followed by late NF, which contrasted with the report by Gu *et al.* (2005) of a largely asymmetric pattern of gene expression divergence between yeast duplicates. Besides these genomic studies, a large number of case studies have illustrated the role of either SF or NF for various genes and species (e.g. Yokoyama & Yokoyama, 1989; Ohta, 1994; Wang & Gu, 2001; Zhang *et al.*, 2002).

This large body of literature remains equivocal with respect to the fate of duplicated genes. One obvious reason for this is the diversity of biological contexts – there is clearly not a rule applying to every duplicate gene pair in every genome. Another difficulty is that most of these studies, with the notable exception of Zhang *et al.* (2003) and Marques *et al.* (2005), analysed relatively old duplications. The alternative models essentially differ in their predictions about the early stages of duplicate gene evolution – in the long run, the two copies are expected to evolve more or less independently, irrespective of the reasons for their initial functional divergence. For ancient events, the patterns generated immediately posterior to gene duplication are likely to be obscured by subsequent long periods of sequence divergence.

In this study, we approach the early steps of duplicate gene evolution by analysing the substitution pattern in 10 recently duplicated genes in mouse. We focus on recent retroduplications, i.e. genes originated through the insertion in the mouse genome of a retrotranscribed mRNA after the rat/mouse split, *c.* 15 million years ago (Mya). This process gives birth to an intron-less open reading frame, which, if expressed, can potentially become a new functional gene, duplicating the coding sequence of the original gene. The 10 genes were sequenced in five to 10 mouse species that diverged during the last 4–5 Mya. This sampling strategy offers the opportunity to address the issue of the level of contingency in the fate of duplicated genes. How long does it take before the decision between retention (functional divergence) or loss (pseudogenization) is made? If several species independently inherit two recently duplicated copies, do they follow a common evolutionary pathway?

A genome-wide search has provided evidence for the continuous creation of new genes by retroposition during recent Primate evolution, and for the presence of a large number of new functional genes generated by this process in the human genome (Marques *et al.*, 2005), thus justifying the interest of studying the evolutionary dynamics of retroposed genes. Retroduplications are specific in that they are initially asymmetric, one copy keeping its genomic location (and regulatory elements), whereas the other copy experiences a new context. This specificity might influence subsequent evolutionary processes; SF appears less likely in the case of retroduplications than for standard, symmetric gene pairs, a statement we plan to test in this study. Importantly, the large physical distance between the two copies essentially ensures that the two duplicates do not exchange genetic information through ectopic recombination or gene conversion, as might occur in the case of tandem duplications. The asymmetry, finally, allows us to formally distinguish between the two copies, which we call 'source' and 'retroposed', and to compare their evolutionary pattern and fate.

Materials and methods

Selection of candidate genes

A bioinformatic search was conducted to identify recently retro-duplicated mouse genes from data banks. We first extracted 6073 intron-less annotated mouse genes from ENSEMBL (version 36, release 24) thanks to the ACNUC retrieval system (Gouy *et al.*, 1984). Among these, genes occurring in multiple copies were detected through an all vs. all BLAST search and removed. The remaining genes were compared with the coding sequences of intron-containing genes from *M. musculus* and *Rattus norvegicus* using BLASTP. Single-copy, intronless mouse genes fully matching an intron-containing mouse gene as first hit, and an intron-containing rat gene as second hit were kept, providing a list of candidate recent retroduplications. Nine candidate genes were selected based on gene length, synonymous divergence levels, and availability of functional annotations. Data from the *Zfx*/*Zfa* gene pair, analysed in detail by Tucker *et al.* (2003), were downloaded from GenBank and added to our data set. The synonymous divergence between the two *M. musculus* copies varied between 0.04 and 0.22 across the 10 analysed genes.

DNA sequencing

The complete or nearly complete coding regions of the retroposed (intron-less) genes were amplified by PCR using the primers described in Appendix S1. Nested or semi-nested PCRs were performed in difficult cases, using

combinations of the external and internal primers described in the Appendix. The source (intron-containing) copy of the *Acot9* gene was also amplified, in two fragments covering the last six exons. After treatment with ExoSap-IT (USB), the PCR products were directly sequenced in both directions using either the external or internal primers described in *Supplemental material*, and the Big Dye Terminator (Applied Biosystems, Foster City, CA, USA) protocol. Electrophoresis was carried out on a 3130 automated capillary sequencer (Applied Biosystems) after ethanol precipitation of the sequencing reactions. Various characteristics of the analysed genes are given in Table 1.

Mouse samples

The mice used here all come from wild derived, moderately inbred colonies maintained in the laboratory. Appendix S2 gives the names, taxonomic and geographical origins of the strains used.

Phylogenetic analyses

For each gene, the mouse retroposed sequences were manually aligned to the coding sequences of the mouse and rat source copy. Maximum-likelihood phylogenetic trees were reconstructed using the DNAML algorithm via PHYLO_WIN (Galtier *et al.*, 1996). A combined analysis of the whole data set was performed using the MRP algorithm (Ragan, 1992).

Pseudogenization analysis

Structural evidence for pseudogenization, namely frameshifts or premature stop codons, was searched for in the retroposed sequences. When the gene/pseudogene status of the retroposed copy was uncertain, computer simulations were conducted to assess the plausibility of the pseudogenization hypothesis, as in Zhang (2003). An ancestral intact coding sequence (identical to the source mouse copy) was neutrally evolved along the branches of the retroposed subtree according to the model described in Duret & Galtier, 2000, which accounts for unequal transition and transversion rates, equilibrium GC-content and CpG hypermutability. Maximum likelihood branch length estimates were used. The percentage of simulations in which no stop codon appeared in present-day sequences was taken as a measure of the likelihood that a retropseudogene has remained intact across mouse species. Simulations were performed using a home-made C program.

d_N/d_S analysis

Lineage-specific variations in synonymous and nonsynonymous substitution rates were analysed using PAML (Yang, 1997, 1998). Six models were compared (Fig. 2). Model M_0 assumes a common d_N/d_S for all lineages. Model M_1 discriminates the rat branch from the mouse subtree; the M_1/M_0 comparison asks whether the duplication event has modified the selective regime. Model M_2 generalizes M_1 in assigning a distinct d_N/d_S ratio to the source and retroposed mouse copies; the M_2/M_1 comparison asks whether post-duplication evolution has been symmetrical. The M_3 model adds a specific d_N/d_S ratio to the deepest branch of the retroposed subtree, thus after duplication but before speciation in *Mus*; the M_3/M_2 comparison asks whether the selective regime applying immediately posterior to duplication changed later on in the retroposed lineage. The M'_1 and M'_2 models parameterize branches in a distinct way, the rat and mouse source copies being grouped and opposed to the retroposed lineage (Fig. 2). The M'_1/M_2 and M'_2/M_3 comparisons ask whether the duplication affected the evolutionary process of the mouse source copy (when compared with rat one). In M_3, the rat, mouse source copy, mouse early retroposed and mouse late retroposed d_N/d_S ratios are called ω_0, ω_1, ω_2 and ω_3, respectively. Models M_0 to M'_2 are defined by equating some of the ω_i's as shown in Fig. 2. To detect the existence of functional constraints, or possible episodes of positive selection, the relevant d_N/d_S ratios were compared with

Table 1 Features of the 10 genes analysed in this study.

Gene	ENSMUSG	chr	Biological process	Retrogene	chr	ENSMUSG	Expression of retrogene
Acot9	25287	X	Acyl-CoA metabolism	*Acot10*	15	47565	Yes
Cs	05683	10	Carbohydrate metabolism	*Csl*	10	46934	Testis
G6pd	31400	X	Glucose and carbohydrate metabolism	*G6pd2*	5	45120	Testis
Prdx6	26701	1	Lipid catabolism	*Prdx6-rs1*	2	50114	Testis
Zfx	00103	X	Regulation of transcription	*Zfa*	10	49576	Testis
Arpc1b	29622	5	Nucleation of actin filaments	XP_486686.1 (RefSeq)	X	46993	Unknown
Gstk1	29864	6	Glutathione metabolism	–	19	47168	Unknown
Nck2	43001	1	Actin filament organization, cell migration	Q78U92_MOUSE (UniPro)	1	63627	Thymus
Ndufa11	02379	17	Protein transport, NADH dehydrogenase activity	–	16	68487	Unknown
Psmc1	21178	12	Protein catabolism	RP23-406A14.1 (Vega)	11	05889 (vega)	Unknown

the threshold value of 1, again using likelihood ratio tests.

When evidence for pseudogenization was detected in some but not all lineages, an additional model, M_ψ, was fitted to the data. M_ψ generalizes M_1' in assigning a distinct d_N/d_S ratio to lineages carrying a nonsense or frameshift substitution (reconstructed by maximum parsimony). The M_ψ/M_1' comparison asks whether the evolution of retroposed copies prior to observable pseudogenization was already neutral. Models were compared through likelihood ratio tests, separately for each gene, or by summing the log-likelihoods over all genes. Additional analyses of branch-specific substitution rates were performed by applying the relative rate test (Robinson-Rechavi & Huchon, 2000) to coding sequences using rat as outgroup to the two mouse copies. Models assuming a variation of d_N/d_S across codons were also applied (to the retrogene subsets), but no such variation was detected, presumably because of a lack of overall variability – most codons have undergone either zero or one change.

Results

In addition to the published *M. musculus* sequence of the source copy of *Acot9*, we were able to obtain those of *M. spretus*, *M. spicilegus*, *M. macedonicus* and *M. caroli* (accession number: DQ975466–DQ975469). We also used the published source sequences of *Zfx* (*M. m. musculus*, *M. m. castaneus*, *M. spretus*, *M. spicilegus*, *M. macedonicus*, *M. caroli*, *M. cookii*, Tucker *et al.*, 2003) to show that in these two cases, the pattern of evolution of the source copy is apparently unaffected by the duplication in the mouse lineage (i.e. the mouse source lineage is not distinguishable from the rat lineage, see below). Therefore for the other genes, we concentrated our sequencing efforts on the retroposed copy. We were able to amplify and sequence the retrogene in five to 10 of the 11 sampled mouse taxa, depending on the gene considered, as indicated in Fig. 1 (accession numbers: DQ987484–DQ987490, DQ992397–DQ992415, EF014731–EF014757).

The analysed retrogenes were chosen from a survey of ENSEMBL coding sequence annotations and because of their similarity to the source gene, so that the *M. musculus* C57BL6 retrocopies have the potential to code for a similar protein. Their coding sequence might however have been altered in other lineages. Among the 10 retrogenes analysed, five showed evidence for pseudogenization in one species or more, namely *Arpc1b*, *Gstk1*, *Nck2*, *Psmc1* and *Ndufa11* (see Fig. 1). The *Prdx6* retroposed copy revealed no direct proof of pseudogenization, but the d_N/d_S analysis did not reveal any selective constraint (see below). Computer simulations indicated that the probability of observing no stop codon in present-day *Prdx6* retroposed sequences assuming it has been a pseudogene as the ancestral mouse species is around 0.1, i.e. relatively low but plausible. The *M. macedonicus Prdx6* retroposed sequence contains a deletion of three bases, weakly suggesting some constraint for preserving the frame. Finally, the *Prdx6* coding sequence exhibits two short tandem repeats whose numbers of repetitions (five and six) were conserved, preserving the reading frame, despite their potentially high mutation rate. The status of the *Prdx6* retroposed copy therefore remains unclear, but it will be considered as a gene in the following.

Phylogenetic trees confirmed the orthology of the retroposed copies from various mouse species, and the relatively young age of duplication events (after the divergence of mouse from rat) as was suggested by the BLAST analyses used to select them. The *Cs* gene, however, showed a slightly different pattern: a homologous retroposed copy branching as the sister group to the mouse retrogenes (with low bootstrap support, though) was found in the *R. norvegicus* genome, suggest-

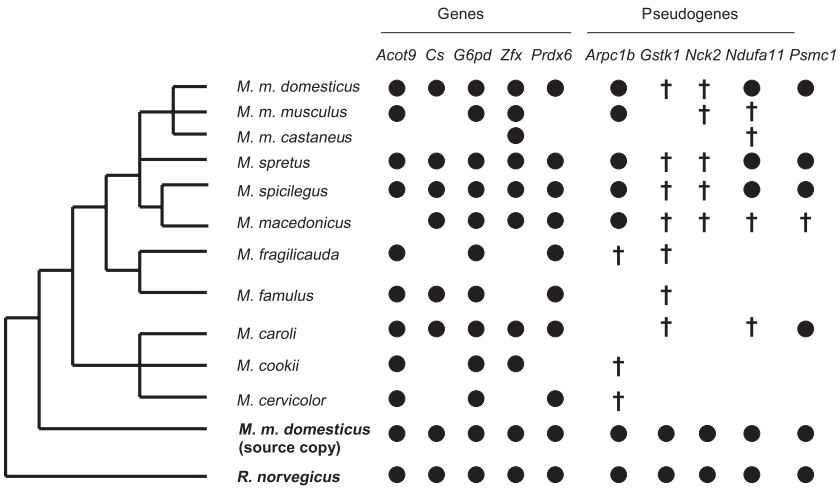


Fig. 1 Gene and species sampling. The taxonomic sampling is given for each gene. Sequences showing evidence for pseudogenization (frameshift or missense mutation) are indicated by crosses. Sequences showing no evidence for pseudogenization are indicated by dots. The source copy was available in additional mouse species for *Acot9* (*M. m. musculus*, *M. spretus*, *M. spicilegus*, *M. macedonicus*, *M. caroli*, this study) and *Zfx* (*M. m. musculus*, *M. m. castaneus*, *M. spretus*, *M. spicilegus*, *M. macedonicus*, *M. caroli*, *M. cookii*, Tucker *et al.*, 2003).

ing that the duplication might have slightly predated the mouse/rat split. This does not alter the rationale of the study, and it can be said that the duplications we analyse are typically 5–15 My old. The combined phylogenetic analysis was in agreement with the well-supported framework described in Fig. 1 (Suzuki *et al.*, 2004), but did not further resolve the uncertainties in branching order. When required, gene-specific trees were slightly modified by hand to conform to this framework in the following analyses.

The selective forces applying before and after duplication were approached by modelling d_N/d_S variations across lineages, calculating the log-likelihood of various hypotheses, and summing over genes to extract the global pattern. The parameters estimated and hypotheses tested are summarized in Figs 2 and 3. Log-likelihoods for genes and pseudogenes are given in Tables 2 and 3, together with d_N/d_S ratio estimates. Likelihood ratio tests were performed to compare alternative models (Table 4). The M_1 model was largely preferred over M_0, indicating a different selection regime before and after the duplication. M_2 also strongly rejected M_1, which means that the effect of the duplication has been asymmetric. The estimated values for ω_1 (mouse source gene) and ω_2 (mouse retroposed copies) indicate a strong acceleration of the nonsynonymous rate in the retroposed lineage (Tables 2 and 3). The asymmetry was confirmed by relative rate tests: faster evolution of the retroposed copy, when compared with the source one, was detected for all the analysed genes (Table 5). The M_2/M_1' and M_3/M_2' comparisons suggest that the mouse source copies were not affected by the duplication: their d_N/d_S ratios are not different from those in the rat branch. Finally, our attempt to distinguish between early and late evolution post-duplication was not successful: the M_3 model did not improve M_2, nor did M_2' improve M_1'. Overall, the most appropriate model appeared to be M_1', the model assigning one d_N/d_S to the rat and mouse source genes, and one d_N/d_S ratio to the retroposed lineage. M_1' is the model selected by Akaike's Information Criterion (result not shown).

The gene-specific patterns were remarkably homogeneous, and conform to the general scenario outlined above. The M_1/M_0 and M_2/M_1 tests were significant for eight and six retroduplications respectively (Table 4). The d_N/d_S ratio of the mouse source copy never differed from that of the rat, and no significant difference was detected between early and late post-duplication evolution. Nonsynonymous/synonymous ratios, in contrast, varied widely among genes. The d_N/d_S estimate for the source copy varied from 0 (*Psmc1*) to nearly 0.5 (*Gstk1*, Table 3). The retroposed copy appeared moderately constrained (d_N/d_S between 0.2 and 0.65, Tables 2 and 3, model M_1') in eight retroduplications, and unconstrained in *Prdx6* (but see above) and *Gstk1* (d_N/d_S around 1). No instance of a d_N/d_S ratio significantly higher than 1 was detected for any of the analysed genes and lineages under any model.

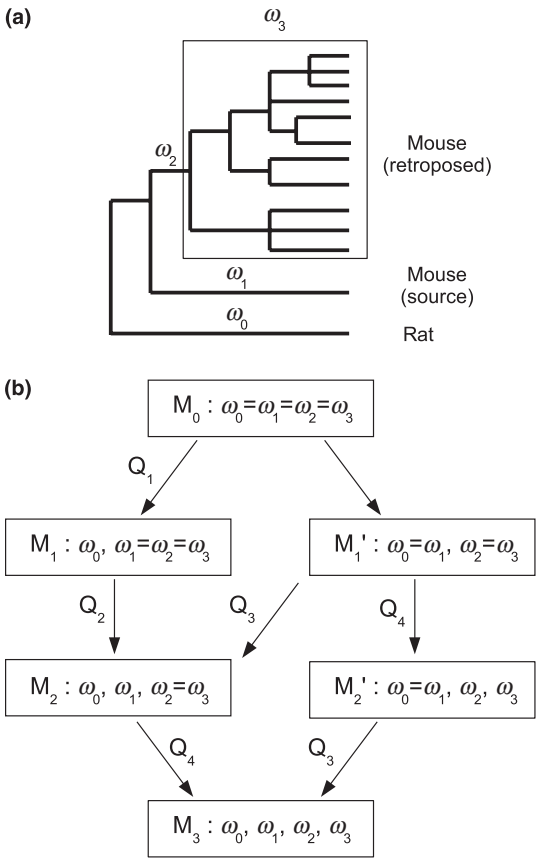


Fig. 2 Modelling d_N/d_S variations across lineages. (a) Schematic definition of ω_i's; (b) models used in this study and their nested relationships (arrows). Q1–Q4 are the biological questions addressed by the various model comparisons. Q1: Did the duplication have any effect on the selective regime? Q2: Was the effect asymmetric? Q3: Was the source copy affected at all by the duplication? Q4: Are the early and late selective regimes posterior to duplication distinct?

Remarkably, the *Nck2* retrogene appears to have evolved under significant functional constraint ($d_N/d_S = 0.37$) although there is evidence of pseudogenization in all mouse species studied here (Fig. 1). The retrocopies of *Arpc1b*, *Psmc1* and *Ndufa11* also show significant evolutionary constraint, but are clearly pseudogenes in some mouse species. For the latter, we fitted the M_ψ model which assigns one d_N/d_S ratio to the source copy, and two to the retroposed one, thus contrasting lineages showing vs. not showing direct evidence for pseudogenization. In all cases, the estimated d_N/d_S ratio was lower in the lineages that lacked evidence of pseudogenization in their sequences than in those displaying such evidence (Table 4). The effect was highly significant for *Arpc1b*, with a d_N/d_S ratio in nonpseudogenized lineages significantly lower than 1. This is evidence for the existence of functional constraints in lineages carrying an intact coding sequence, indicating that these sequences are not functionally irrelevant.

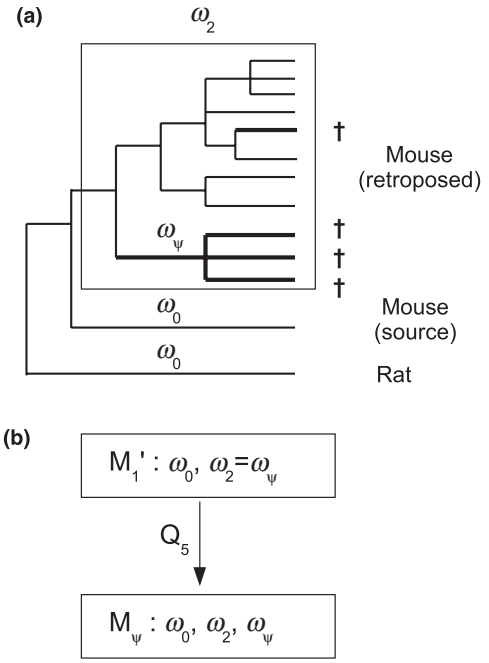


Fig. 3 The M_ψ model (pseudogenes only). M_ψ generalizes M_1' in assigning a distinct d_N/d_S ratio, ω_ψ, to retrogene lineages showing evidence for pseudogenization (thick lines). The question asked by the M_ψ/M_1' comparison, Q5, is: Did the retroposed copies evolve distinctly whether they were structurally pseudogenized or not?

Discussion

Focusing on very recent duplicates in several closely related species, we were able to dissect the early stage of post-duplication evolution in 10 genes. Mouse is an ideal model for this purpose thanks to the many close relatives to the fully sequenced *M. musculus*. By using retroduplications we ensured that gene conversion between paralogues has not occurred, and that we could compare the independent fates of the source and retroposed copies.

The pattern we observed is dominated by the asymmetry between the two copies, as could have been expected given the specificity of the retroduplication process. In all of the 10 cases, the source copy showed unchanged evolutionary rate and d_N/d_S ratio, whereas the retroposed copy experienced a significant acceleration. This is true of pseudogenes, as expected, but also of genes. This asymmetric pattern is not compatible with the SF model of duplicated genes evolution: there is no evidence that the source copies, apparently unaffected by the retroduplication events, have lost any of their ancestral functions. The retroposed copies were either functionally lost by pseudogenization (five or six genes), or kept (i.e. not pseudogenized, four or five genes), consistent with the NF model of duplicated genes evolution.

No evidence for positive selection, however, could be detected from the d_N/d_S analysis. There are several possible explanations for this negative result. First, the nonsynonymous substitution pattern probably reflects a combination of advantageous, neutral and deleterious changes, so that some of the amino-acid changes having occurred in the retroposed lineages can be adaptive without resulting in a d_N/d_S ratio higher than 1. We lack power with this dataset to detect individual codons under positive selection. Secondly, adaptive changes might have occurred in the regulatory regions controlling the expression of the retrogenes. One could even imagine that the retogenes reached a new function upon birth, e.g. by retroinserting into a genomic context favouring their expression in a new tissue. NF could occur without any adaptive change under this scenario. In all cases where such information is available, the retrogenes appear to have a tissue-specific expression pattern (Table 1), mostly in the testis, as appears to be often the case of retrocopied genes (Marques *et al.*, 2005).

The asymmetric evolutionary pattern is most probably because of the asymmetric nature of retroduplications, one copy keeping its location and regulatory elements whereas the other one explores a new genomic context. Retroduplications obviously do not fit the paradigm of initially indistinguishable duplicates, assumed in most theoretical models for duplicate gene evolution, and this specificity has a strong impact on the fate of duplicates. Retroduplication is not a negligible mechanism of gene duplication, at least in mammals. About 6000 ENSEMBL-annotated mouse genes out of 30 000 are intron-less, suggesting a retroposed origin. Dozens of human retrogenes originated during primate evolution are functionally active (Marques *et al.*, 2005), generally in testis.

It should be noted that in the long run the initial asymmetry can be overcome. Comparing genes having a recent retrogene paralogue to genes without such a retroduplicate, Huminiecki & Wolfe (2004) reported evidence for a more frequent shift in gene expression pattern for the former category in human and mouse, suggesting that late SF can follow early NF. This scenario contrasts with He & Zhang's (2005) report of an early SF followed by late NF in yeast duplicates. But He & Zhang (2005) analysed regular genes, whereas Huminiecki & Wolfe (2004) and the present study focused on retroduplications. This suggests that both NF and SF are parts of the long-term evolutionary process of most duplicated genes irrespective of their origin, although early duplicate evolution is strongly influenced by the symmetric/asymmetric status of the duplication event. Applying our approach to a set of nonretroposed recent duplications would provide further assessment to this statement. The *e(y)2* gene of *Drosophila melanogaster* is an extreme example of how the long-term fate of duplicates can depart from the original situation: *e(y)2* is an ubiquitously expressed retrogene, whereas its intron-containing paralogue is expressed only in male germ cells (Krasnov *et al.*, 2005).

Table 2 Log-likelihood and d_N/d_S ratio estimates for five genes.

Gene	lg†	d_S‡	M_0	M_1	M_2	M'_1§	M'_2	M_3
Acot9	168	0.076	−1100.07 $\omega_0 = 0.133$	−1100.06 $\omega_0 = 0.124$ $\omega_1 = 0.136$	−1098.36 $\omega_0 = 0.124$ $\omega_1 = 0.057$ $\omega_2 = 0.202$	−1098.85 $\omega_0 = 0.088$ $\omega_2 = 0.202^{**}$	−1098.04 $\omega_0 = 0.088$ $\omega_2 = \text{inf}$ $\omega_3 = 0.187$	−1097.54 $\omega_0 = 0.124$ $\omega_1 = 0.057$ $\omega_2 = \text{inf}$ $\omega_3 = 0.187$
Cs	379	0.220	−2346.60 $\omega_0 = 0.170$	−2340.93 $\omega_0 = 0.041$ $\omega_1 = 0.232$	−2338.38 $\omega_0 = 0.043$ $\omega_1 = 0.095$ $\omega_2 = 0.296$	−2338.96 $\omega_0 = 0.064$ $\omega_2 = 0.296^{**}$	−2338.27 $\omega_0 = 0.064$ $\omega_2 = 0.381$ $\omega_3 = 0.223$	−2337.69 $\omega_0 = 0.043$ $\omega_1 = 0.094$ $\omega_2 = 0.380$ $\omega_3 = 0.223$
G6pd	487	0.126	−3369.58 $\omega_0 = 0.284$	−3353.77 $\omega_0 = 0.039$ $\omega_1 = 0.402$	−3346.95 $\omega_0 = 0.041$ $\omega_1 = 0.043$ $\omega_2 = 0.479$	−3346.95 $\omega_0 = 0.041$ $\omega_2 = 0.479^{**}$	−3346.40 $\omega_0 = 0.042$ $\omega_2 = 0.343$ $\omega_3 = 0.520$	−3346.40 $\omega_0 = 0.041$ $\omega_1 = 0.044$ $\omega_2 = 0.343$ $\omega_3 = 0.520$
Prdx6	219	0.095	−1683.26 $\omega_0 = 0.559$	−1674.12 $\omega_0 = 0.147$ $\omega_1 = 0.898$	−1673.25 $\omega_0 = 0.151$ $\omega_1 = 0.392$ $\omega_2 = 1.008$	−1674.06 $\omega_0 = 0.194$ $\omega_2 = 1.020$	−1673.69 $\omega_0 = 0.193$ $\omega_2 = 1.518$ $\omega_3 = 0.908$	−1672.94 $\omega_0 = 0.151$ $\omega_1 = 0.373$ $\omega_2 = 1.450$ $\omega_3 = 0.907$
Zfx	323	0.043	−1733.88 $\omega_0 = 0.178$	−1729.50 $\omega_0 = 0.0$ $\omega_1 = 0.241$	−1725.31 $\omega_0 = 0.0$ $\omega_1 = 0.031$ $\omega_2 = 0.357$	−1726.00 $\omega_0 = 0.015$ $\omega_2 = 0.357^{**}$	−1726.00 $\omega_0 = 0.015$ $\omega_2 = 0.159$ $\omega_3 = 0.357$	−1725.31 $\omega_0 = 0.0$ $\omega_1 = 0.031$ $\omega_2 = 0.362$ $\omega_3 = 0.357$

*Significantly lower than 1 at the 5 % level.
**Significantly lower than 1 at the 1 % level.
†Number of codons analysed.
‡Synonymous divergence between the two copies of *M. musculus*.
§Parameter ω_2 in M'_1 was compared to the threshold value of 1.

Our sampling of several closely related species and several genes potentially allows us to address the issue of the contingency/reproductibility in the fate of duplicated genes. When a pair of recent duplicates is independently inherited by various species, does it evolve along the same lines in distinct lineages (Scannell *et al.*, 2006)? The asymmetry between the two copies was, again, the dominant pattern. The source copy was essentially unaffected by the duplication, and this was true in all mouse species for *Acot9* and *Zfx*, for which we had multispecies data for the source gene. For this reason, we concentrated our effort on sequencing the retroposed copy of the other genes, the major uncertainty upon duplication being whether the retroposed copy will be functionally kept or lost, and how long it takes before such a decision is made.

Besides the enigmatic *Prdx6* retrogene, which appears unconstrained but shows no sign of pseudogenization, four genes (*Acot9*, *Cs*, *G6pd*, *Zfx*) probably took a rapid NF decision: they are intact and under functional constraint in all the examined lineages, and there is good evidence of expression (Ashworth *et al.*, 1990; Hendriksen *et al.*, 1997; Poupon *et al.*, 1999; Mammalian Gene Collection Program Team, 2002) that appears to be specific, mostly in the testis. The *Gstk1* retrogene apparently took a rapid gene loss decision: it is a structural pseudogene in all the surveyed species, and shows a d_N/d_S ratio not significantly different from 1. Three or four genes, in contrast, are obvious pseudogenes in some species, but probably functional in other lineages, as indicated by a d_N/d_S ratio significantly lower than 1 under model M'_1 (*Arpc1b*, *Nck2*, *Ndufa11*; *Psmc1* *P*-value: 0.06). These genes might therefore illustrate the role of random processes during the early stages of duplicated gene evolution, when few crucial mutations determine the ultimate fate of duplicated genes. This would imply a variable duration of this critical stage across retrogenes, some coming very rapidly to a decision, whereas others are still undetermined millions of years after the duplication event.

Some aspects of our results, however, suggest that this interpretation is not sufficient. The d_N/d_S analysis reveals that at least two (*Arpc1b* and *Nck2*) retrogenes have been evolving under significant functional constraint before they became pseudogenes. This might also be true of *Psmc1* and *Ndufa11*, as suggested by estimated d_N/d_S ratios under M_ψ. This observation is not consistent with a pure NF model, which predicts a neutral stage posterior to duplication, followed by either gene loss of functional

Table 3 Log-likelihood and d_N/d_S ratio estimates for five pseudogenes.

Gene	lg†	d_S‡	M_0	M_1	M_2	M'_1§	M'_2	M_3	M_ψ§
Arpc1b	359	0.077	−2272.18 $\omega_0 = 0.202$	−2262.69 $\omega_0 = 0.049$ $\omega_1 = 0.313$	−2253.06 $\omega_0 = 0.047$ $\omega_1 = 0.0$ $\omega_2 = 0.426$	−2254.88 $\omega_0 = 0.035$ $\omega_2 = 0.426$**	−2253.55 $\omega_0 = 0.039$ $\omega_2 = 0.0$ $\omega_3 = 0.426$	−2251.94 $\omega_0 = 0.050$ $\omega_1 = 0.0$ $\omega_2 = 0.0$ $\omega_3 = 0.426$	−2252.06 $\omega_0 = 0.037$ $\omega_2 = 0.222$** $\omega_\psi = 0.668$
Gstk1	200	0.170	−1630.27 $\omega_0 = 0.766$	−1629.63 $\omega_0 = 0.498$ $\omega_1 = 0.845$	−1628.26 $\omega_0 = 0.496$ $\omega_1 = 0.470$ $\omega_2 = 1.027$	−1628.26 $\omega_0 = 0.482$ $\omega_2 = 1.027$	−1628.05 $\omega_0 = 0.488$ $\omega_2 = 0.787$ $\omega_3 = 1.119$	−1628.05 $\omega_0 = 0.502$ $\omega_1 = 0.475$ $\omega_2 = 0.787$ $\omega_3 = 1.119$	NA
Nck2	290	0.065	−1722.58 $\omega_0 = 0.150$	−1706.48 $\omega_0 = 0.007$ $\omega_1 = 0.322$	−1703.75 $\omega_0 = 0.008$ $\omega_1 = 0.0$ $\omega_2 = 0.374$	−1703.85 $\omega_0 = 0.007$ $\omega_2 = 0.373$**	−1702.42 $\omega_0 = 0.007$ $\omega_2 = 0.086$ $\omega_3 = 0.440$	−1702.33 $\omega_0 = 0.008$ $\omega_1 = 0.0$ $\omega_2 = 0.086$ $\omega_3 = 0.441$	NA
Ndufa11	121	0.125	−923.45 $\omega_0 = 0.233$	−921.48 $\omega_0 = 0.129$ $\omega_1 = 0.331$	−920.01 $\omega_0 = 0.124$ $\omega_1 = 0.083$ $\omega_2 = 0.432$	−920.11 $\omega_0 = 0.116$ $\omega_2 = 0.427$*	−919.33 $\omega_0 = 0.132$ $\omega_2 = 0.072$ $\omega_3 = 0.452$	−919.33 $\omega_0 = 0.132$ $\omega_1 = 0.128$ $\omega_2 = 0.073$ $\omega_3 = 0.452$	−920.10 $\omega_0 = 0.117$ $\omega_2 = 0.407$ $\omega_\psi = 0.451$
Psmc1	304	0.077	−1828.53 $\omega_0 = 0.225$	−1805.67 $\omega_0 = 0.0$ $\omega_1 = 0.508$	−1801.30 $\omega_0 = 0.0$ $\omega_1 = 0.0$ $\omega_2 = 0.632$	−1801.30 $\omega_0 = 0.0$ $\omega_2 = 0.632$	−1800.08 $\omega_0 = 0.0$ $\omega_2 = 0.0$ $\omega_3 = 0.645$	−1800.08 $\omega_0 = 0.0$ $\omega_1 = 0.0$ $\omega_2 = 0.0$ $\omega_3 = 0.645$	−1801.05 $\omega_0 = 0.$ $\omega_2 = 0.603$ $\omega_\psi = 1.340$

NA, not applicable.
*Significantly lower than 1 at the 5 % level.
**Significantly lower than 1 at the 1 % level.
†Number of codons analysed.
‡Synonymous divergence between the two copies of *M. musculus*.
§Parameter ω_2 in M'_1 and in M_ψ was compared with the threshold value of 1.

Table 4 Likelihood ratio tests.

LRT†	*Acot9*	*Cs*	*G6pd*	*Prdx6*	*Zfx*	*Arpc1b*	*Gstk1*	*Nck2*	*Ndufa11*	*Psmc1*	All
M_1/M_0 (Q1)	NS	**	**	**	**	**	NS	**	*	**	**
M_2/M_1 (Q2)	NS	*	**	NS	**	**	NS	*	NS	**	**
M_2/M'_1 (Q3)	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
M_3/M'_2 (Q3)	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
M'_2/M_1 (Q4)	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
M_3/M_2 (Q4)	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
M_ψ/M'_1 (Q5)	NA	NA	NA	NA	NA	*	NA	NA	NS	NS	NS

NS, not significant.
*Significant at the 5 % level.
**Significant at the 1 % level.
†Model comparisons and biological questions addressed (see Figs 2 and 3).

divergence. It is difficult to understand how a gene undergoing purifying selection can become a pseudogene: how could natural selection act to remove deleterious nonsynonymous changes, but later on allow the loss of the whole gene function?

One possible explanation to this paradox would be that these retrogenes, as they are expressed and code for a functional protein, have an effective but dispensable functional impact. This could occur if, for instance, the protein encoded by the retrogene was competing with that encoded by the source gene. Mutations in the retrogene could produce a deficient protein with a deleterious impact, for instance if the mutated protein linked the substrate but did not process it. Such duplicates would therefore generate an undesired functional redundancy, increasing the genetic load without conferring any advantage, until a fully silencing mutation (e.g. an early stop codon, a frameshift in the coding sequence

Table 5 Relative rate tests.

Gene	d_N source†	d_N retroposed‡	*P*-value
Acot9	0.019	0.027	*
Cs	0.013	0.039	**
G6pd	0.007	0.036	**
Prdx6	0.038	0.088	**
Zfx	0.001	0.010	**
Arpc1b	0.009	0.029	**
Gstk1	0.077	0.116	**
Nck2	0.002	0.018	**
Ndufa11	0.055	0.072	NS
Psmc1	0	0.013	**

NS, not significant.
*Significant at the 5 % level.
**Significant at the 1 % level.
†Average non-synonymous substitution rate between rat and source copies.
‡Average non-synonymous substitution rate between rat and retroposed copies.

or a transcription arrest) eventually leads to pseudogenization *sensu stricto*. We note that among the retrogenes we studied a majority of those with an autosomal source copy show signs of pseudogenization, whereas none of the three with an X source copy does. Functional redundancy is less likely to occur for the latter category in the testis (the major expression site of retrogenes) as most X genes are inactivated during spermatogenesis. This could in turn reduce, or even suppress the selection pressure for pseudogenization on these retrogenes of X origin and contribute to explain why autosomal retrogenes more often derive from X than autosomal source copies (Emerson *et al*., 2004). Purifying selection against dominant mutations on redundant genes was invoked by Hughes (1994) as a possible explanation for the lower than 1 d_N/d_S ratio observed between recent duplicates in *Xenopus laevis* (Hughes & Hughes, 1993). If confirmed, this process would imply a longer time period before complete loss of function of the retrogenes, when compared with the standard model, leaving more time for potential adaptive changes of expression pattern. If the redundancy load is strong, it might even be the case that mutations resulting in complete silencing of retrogenes are favoured by positive selection, which should be testable by comparative analyses of recently pseudogenized retrogenes and by polymorphism studies around such retrogenes.

Our results therefore confirm that, as could have been expected, retrocopied duplicates most frequently evolve through NF and/or gene loss, but not SF. We furthermore outline the fact that purifying selection is the predominant mode of sequence evolution very quickly posterior to retroduplication, when the ultimate fate of the duplicates is not yet determined. The 'redundancy load' hypothesis, if confirmed, illustrates how ambiguous the definition of the functional status of a duplicated gene can be.

Acknowledgments

Mice from the 'Conservatoire Génétique de souris sauvages' (http://www.univ-montp2.fr/~genetix/souris.htm) were kindly provided by Jean-Jacques Duquesne, Annie Orth and François Bonhomme. We thank E. Desmarais and Institut Fédératif de Recherche 119 for providing the DNA sequencing equipment and protocols. This work was funded by institutional support from the CNRS and Université Montpellier II.

References

Ashworth, A., Skene, B., Swift, S. & Lovell-Badge, R. 1990. Zfa is an expressed retroposon derived from an alternative transcript of the Zfx gene. *EMBO J.* **9**: 1529–1534.

Duret, L. & Galtier, N. 2000. The covariation between TpA deficiency, CpG deficiency and the G + C-content of human isochores is a mathematical artefact. *Mol. Biol. Evol.* **17**: 1620–1625.

Emerson, J.J., Kaessmann, H., Betran, E. & Long, M. 2004. Extensive gene traffic on the mammalian X chromosome. *Science* **303**: 537–540.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L. & Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545.

Galtier, N., Gouy, M. & Gautier, C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comp. Appl. Biosci.* **12**: 543–548.

Gouy, M., Milleret, F., Mugnier, C., Jacobzone, M. & Gautier, C. 1984. ACNUC: a nucleic acid sequence data base and analysis system. *Nucleic Acids Res* **12**: 121–127.

Gu, X., Zhang, Z. & Huang, W. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc. Natl. Acad. Sci. USA* **102**: 707–712.

He, X. & Zhang, J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**: 1157–1164.

Hendriksen, P.J., Hoogerbrugge, J.W., Baarends, W.M., de Boer, P., Vreeburg, J.T., Vos, E.A., van der Lende, T. & Grootegoed, J.A. 1997. Testis-specific expression of a functional retroposon encoding glucose-6-phosphate dehydrogenase in the mouse. *Genomics* **41**: 350–359.

Hughes, A.L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc. Biol. Sci.* **256**: 119–124.

Hughes, M.K. & Hughes, A.L. 1993. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol. Biol. Evol.* **10**: 1360–1369.

Huminiecki, L. & Wolfe, K.H. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res* **14**: 1870–1879.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I. & Koonin, E.V. 2002. Selection in the evolution of gene duplications. *Genome Biol.* **3**: research 0008.

Krasnov, A.N., Kurshakova, M.M., Ramensky, V.E., Mardanov, P.V., Nabirochkina, E.N. & Georgieva, S.G. 2005. A retrocopy of a gene can functionally displace the source gene in evolution. *Nucleic Acids Res* **33**: 6654–6661.

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

Mammalian Gene Collection Program Team 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci. USA* **99**: 16899–16903.

Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A. & Kaessmann, H. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.* **3**: e357.

Ohno, S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, Berlin.

Ohta, T. 1994. Further examples of evolution by gene duplication revealed through DNA sequence comparisons. *Genetics* **138**: 1331–1337.

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

Poupon, V., Begue, B., Gagnon, J., Dautry-Varsat, A., Cerf-Bensussan, N. & Benmerah, A. 1999. Molecular cloning and characterization of MT-ACT48, a novel mitochondrial acyl-CoA thioesterase. *J. Biol. Chem.* **274**: 19188–19194.

Ragan, M.A. 1992. Phylogenetic inference based on matrix representation of trees. *Mol. Phylogenet. Evol.* **1**: 53–58.

Robinson-Rechavi, M. & Huchon, D. 2000. RRTree: relative-rate tests between groups of sequences on a phylogenetic tree. *Bioinformatics* **16**: 296–297.

Robinson-Rechavi, M. & Laudet, V. 2001. Evolutionary rates of duplicate genes in fish and mammals. *Mol. Biol. Evol.* **18**: 681–683.

Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S. & Wolfe, K.H. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**: 340–345.

Suzuki, H., Shimada, T., Terashima, M., Tsuchiya, K. & Aplin, K. 2004. Temporal, spatial, and ecological modes of evolution of Eurasian *Mus* based on mitochondrial and nuclear gene sequences. *Mol. Phylogenet. Evol.* **33**: 626–646.

Tucker, P.K., Adkins, R. & Rest, J.S. 2003. Differential rates of evolution for the ZFY-related zinc finger genes, *Zfy*, *Zfx*, and *Zfa* in the mouse genus *Mus*. *Mol. Biol. Evol.* **20**: 999–1005.

Van de Peer, Y., Taylor, J.S., Braasch, I. & Meyer, A. 2001, The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J. Mol. Evol.* **53**: 436–446.

Wagner, A. 1998. The fate of duplicated genes: loss or new function? *Bioessays* **20**: 785–788.

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

Yang, Z. 1997. PAML: a programme package for phylogenetic analysis by maximum likelihood. *Comp. 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.

Yokoyama, S. & Yokoyama, R. 1989. Molecular evolution of human visual pigment genes. *Mol. Biol. Evol.* **62**: 186–197.

Zhang, J. 2003. Evolution of the human ASPM gene, a major determinant of brain size. *Genetics* **165**: 2063–2070.

Zhang, J., Zhang, Y.P. & Rosenberg, H.F. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat. Genet.* **30**: 411–415.

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

Supplementary Material

The following supplementary material is available for this article online:

Appendix S1. Data.

Appendix S2. Primers.

Appendix S3. Strains.

This material is available as part of the online article from http://www.blackwell-synergy.com

Received 5 July 2006; revised 25 August 2006; accepted 28 August 2006