BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 15 2006, pages 1815–1822
doi:10.1093/bioinformatics/btl280

Sequence analysis

Evolutionary simulations to detect functional lineage-specific genes

Isabelle Dupanloup[1,2] and Henrik Kaessmann[1,*]

[1]Center for Integrative Genomics, University of Lausanne, CH-1015 Lausanne, Switzerland and [2]Computational and Molecular Population Genetics Lab, Zoological Institute, University of Bern, CH-3012 Bern, Switzerland

Received on March 14, 2006; revised on May 8, 2006; accepted on June 2, 2006
Advance Access publication June 9, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Supporting the functionality of recent duplicate gene copies is usually difficult, owing to high sequence similarity between duplicate counterparts and shallow phylogenies, which hamper both the statistical and experimental inference.

Results: We developed an integrated evolutionary approach to identify functional duplicate gene copies and other lineage-specific genes. By repeatedly simulating neutral evolution, our method estimates the probability that an ORF was selectively conserved and is therefore likely to represent a bona fide coding region. In parallel, our method tests whether the accumulation of non-synonymous substitutions reveals signatures of selective constraint. We show that our approach has high power to identify functional lineage-specific genes using simulated and real data. For example, a coding region of average length (~1400 bp), restricted to hominoids, can be predicted to be functional in ~94–100% of cases. Notably, the method may support functionality for instances where classical selection tests based on the ratio of non-synonymous to synonymous substitutions fail to reveal signatures of selection. Our method is available as an automated tool, ReEVOLVER, which will also be useful to systematically detect functional lineage-specific genes of closely related species on a large scale.

Availability: ReEVOLVER is available at http://www.unil.ch/cig/page7858.html.

Contact: Henrik.Kaessmann@unil.ch

Supplementary Data: Supplementary Data are available at *Bioinformatics* online.

1 INTRODUCTION

Gene duplication is a key mechanism for the emergence of new genes and new phenotypes in different evolutionary lineages (Li, 1997; Long *et al*., 2003; Samonte and Eichler, 2002). New duplicate genes may originate through (segmental) gene duplication by intra- or interchromosomal transposition of gene-containing segments or retroposition (Long *et al*., 2003; Samonte and Eichler, 2002). Both of these mechanisms were shown to have generated a significant number of new functional genes in mammalian and invertebrate genomes (Betran *et al*., 2002; Eichler and Sankoff, 2003; Emerson *et al*., 2004; Long *et al*., 2003; Marques *et al*., 2005).

The functionality of individual gene copies is usually supported in two ways: (1) experimentally, e.g. by collecting expression evidence of the putative gene (e.g. Feral *et al*., 2001; Ting *et al*., 2004) and/or (2) by comparative genomics approaches that detect signatures of selection or sequence conservation using paralogous and orthologous genes (Boffelli *et al*., 2004a; Frazer *et al*., 2003; Li, 1997; Nekrutenko *et al*., 2003, 2002). However, inferring functionality of recently emerged gene copies using either one of these approaches is hampered by the high sequence similarity to duplicate counterparts and their limited phylogenetic distribution. This restricts, for example, the power of methods based on sequence conservation and renders the design of specific antibodies that may detect encoded proteins difficult.

Recently, a new method, phylogenetic shadowing, has been developed to predict functional elements in genomes through the analysis of sequence conservation profiles of multiple sequences from closely related species (Boffelli *et al*., 2003; Ovcharenko *et al*., 2004). This method was demonstrated to be able to successfully demarcate genes based solely on primate-specific sequences (Boffelli *et al*., 2003; Ovcharenko *et al*., 2004). However, this approach requires functional sequences to be well conserved in either a relatively large number or divergent primate species and, thus, may not be applicable for the detection of functional gene copies with a very narrow phylogenetic distribution. Also, this method requires long sequences for the generation and analysis of conservation profiles, where regions of high conservation (with 'low mutation rates') are regarded as being functionally preserved. Thus, it does not provide an explicit test that can be used to predict the functional preservation of an isolated putative coding region during evolution.

We here present an integrated evolutionary method that explicitly tests for functional preservation of a putative gene with a restricted phylogenetic distribution, such as a recently emerged duplicate gene. Our simulation approach estimates the probability that such a gene copy would have retained the integrity of its open reading frame (ORF) since the duplication event in all or most species in which it is present, if it had evolved neutrally along the lineages of the phylogeny. A similar principle (estimating the 'half-life' of a coding sequence) was previously successfully applied to test for human lineage-specific preservation of the metazoan *ASPM* gene (Zhang, 2003). However, among other differences and limitations relative to our approach, the method by Zhang is restricted to the analysis of a single sequence (see Discussion for details).

In addition to estimating pseudogenization probabilities, our method also tests for the selective accumulation of non-synonymous

*To whom correspondence should be addressed.

source: https://doi.org/10.7892/boris.115845 | downloaded: 20.4.2024

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

substitutions, thus providing two independent tests of functionality of an ORF.

We recently showed that our simulation approach can provide strong functional support for recently emerged retroduplicate gene copies (Marques *et al*., 2005). Here we systematically evaluate its power and behaviour for a wide range of parameters. Using simulated as well as real datasets, we show that it has strong power to detect functionally preserved gene copies with a limited phylogenetic distribution. We also describe the fully integrated and publicly available tool, ReEVOLVER (http://www.unil.ch/cig/page7858.html), which implements this method.

2 METHODS

Generation of simulated datasets. For each of the different simulated datasets (Fig. 1), we first reconstructed artificial sequences at the time-point of the duplication event using codon frequencies estimated from α and β globin codon sequences of five mammalian species (Yang *et al*., 2000). Synonymous substitutions as well as indels (for the neutral scenario) were then allowed to accumulate throughout the species phylogeny using specific synonymous substitution and indel rates. We used a synonymous substitution rate of 1.0×10^{-9} per site per year as suggested previously for hominoids and Old World monkeys (Yi *et al*., 2002), and an indel rate (for the neutral scenario) of 1.0×10^{-10} per site per year (Zhang and Webb, 2003). Indels with a multiple of three nucleotides (16%) were assumed to be non-deleterious, as they do not disrupt the ORF (Britten, 2002; Silva and Kondrashov, 2002). Non-synonymous substitutions were added on the different branches of the phylogeny according to the K_A/K_S ratio specified for the simulations. The number of non-synonymous and synonymous sites in the sequences, needed to estimate the number of substitutions from their corresponding rate, were obtained using the reconstructed ancestral sequence and the software GESTIMATOR from the libsequence library (Thornton, 2003).

Ancestral sequence reconstruction in ReEVOLVER. The ancestral state of the gene copy at the time point of the oldest speciation event following gene duplication is reconstructed using a maximum parsimony [using the DNAPARS software available in the PHYLIP package, (Felsenstein, 1996; Yang, 1997)] or a maximum likelihood approach [using the codeml program of PAML (Felsenstein, 1996; Yang, 1997)]. For the benchmarking analysis in this article, ancestral sequences were reconstructed using codeml.

Real datasets. For the four genes (*GLUD*2, *ECP*, *GMCL*2 and *OdsH*) for which our method was tested individually, codon sequences were aligned on the basis of the translated sequence alignments using the EMBOSS package (Rice *et al*., 2000). Phylogenetic trees and divergence times between species were used as proposed in Goodman (1999) for primates and in Kliman *et al*. (2000) for *Drosophila*. For our simulations of the primate genes, we used a substitution rate of 1.0×10^{-9} per site per year (Yi *et al*., 2002) and an indel of 1.0×10^{-10} per site per year (Zhang and Webb, 2003). For the *Drosophila Odysseus* locus, we used a substitution rate of 3.0×10^{-8} per site per year, which is slightly higher than common estimates (Li, 1997; Li *et al*., 1985) but better fits the data (i.e. the observed number of synonymous substitutions fall within the corresponding distribution generated by simulations), and an indel rate of 4.5×10^{-9} per site per year, which corresponds to 15% of the substitution rate (Parsch, 2003).

PAML analyses. For the duplicate genes considered in this study, models of variable K_A/K_S ratios among primate or *Drosophila* lineages were fitted by maximum likelihood to the sequence alignments using codeml from the PAML package version 3.14 (Yang, 1997). The significance of difference between different models was assessed using likelihood ratio tests as previously described (Yang, 1998). First, the one-ratio (M0) and two-ratio models implemented in codeml were compared to examine whether the duplicate lineages show different K_A/K_S ratios than parental/paralogous lineages. We further tested whether the K_A/K_S ratio on the duplicate gene lineage after the first speciation event was significantly different from 1 by comparing two-ratio models where K_A/K_S is estimated from the data or fixed to 1. This test was also used to detect purifying selection (K_A/K_S significantly below 1) in the simulated datasets. To test for the presence of sites under diversifying selection ($K_A/K_S > 1$) on the duplicate lineages, model M1 and model A were compared. Model M1 assumes two classes of sites for the sequences in the whole phylogeny: conserved sites ($K_A/K_S < 1$) and neutral sites ($K_A/K_S = 1$). Model A adds a third class of sites in the duplicates lineages, with K_A/K_S as a free parameter, allowing for sites with $K_A/K_S > 1$. In an additional test, this model was then also used in a comparison with a similar model, where the third site class is fixed at one. Sites under positive selection in the duplicate gene lineages were identified using a recently developed Bayesian approach (Yang *et al*., 2005).

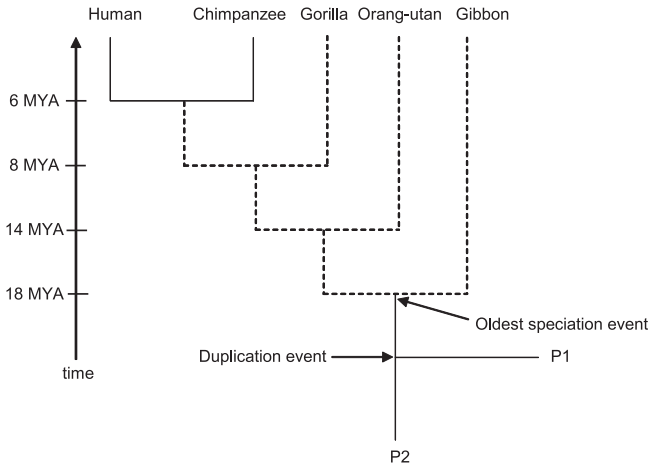

Fig. 1. Tree topology and speciation times used in the simulations. We considered four types of phylogenies, with the gene copy present and typed in two (human, chimpanzee), three (human, chimpanzee, gorilla), four (human, chimpanzee, gorilla, orang-utan) and five (human, chimpanzee, gorilla, orang-utan, gibbon) hominoid species, and the sequences of the progenitor locus present in 2 species (P1 and P2). Speciation times were taken from Goodman (1999).

3 IMPLEMENTATION

We implemented our simulation approach, which includes the selection tests ($test_{dis}$ and $test_{NaNs}$) described in Results, in the publicly available tool ReEVOLVER (http://www.unil.ch/cig/page7858.html). ReEVOLVER was written in the standard C++ programming language (we provide the source code on the webpage), which ensures short runtimes (see below). In addition to the ReEVOLVER tool, the codeml program of PAML (Yang, 1997, http://abacus.gene.ucl.ac.uk/software/paml.html) and dnapars of the PHYLIP package (Felsenstein, 1996, http://evolution.genetics.washington.edu/phylip.html) need to be downloaded and installed; ReEVOLVER calls these programs for the initial step of ancestral sequence reconstruction. Linux, Windows and Mac (OS $\times$ 10.4) versions of the ReEVOLVER software are available.

To run the software, the user provides the coding sequence alignment (without ORF disruptions) of the putative gene of interest, point substitution rate, indel rate, CpG rate (optional), species phylogeny of the sequences, observed number of stop codons/indels per branch, individual branch lengths in millions of years and number of simulations to be performed. The branch lengths can also be used to adjust lineage-specific evolutionary

rates, as the number of substitutions that accumulate on a given branch is the result of the evolutionary rate and branch length. Batch runs of multiple sequence alignments can also be performed.

The output of the program consists of four files: (1) a logfile containing all information for each round of the simulations such as the inferred ancestral sequence, the number of observed substitutions along the different lineages of the species phylogeny and details about the simulations procedure; (2) the number of deleterious mutations (stop codons and indels) for each replicate; (3) the number of non-synonymous and synonymous substitutions per replicate run and (4) the test probabilities, P_{dis} and P_{NaNs} (including associated information).

An average run, during which neutral evolution is repeatedly simulated 100 000 times for an ORF of average length (1401 bp) that is present in all hominoid species analyzed (five sequences), requires around 1 h of computational time on a single CPU (2.2 GHz, Intel Pentium processor). A similar simulation of a sequence present in humans and chimpanzees is completed after 10 min. Thus, a large-scale genomic scan of, for example, 1000 hominoid-specific genes can be completed in <5 days on a small computer cluster of 10 nodes.

4 RESULTS

4.1 Principle of the method

A non-functional gene copy will accumulate mutations—stop codons or frameshifts—over time that disrupt its ORF and may preclude gene function, whereas under functional constraint, natural selection will prevent the accumulation of such deleterious mutations (Li, 1997). On the basis of this assumption and observed sequences from different (closely related) species, our method tests whether a duplicate gene copy (or any other putative gene for which functionality is ascertained in a limited number of lineages, see Discussion) would have retained its ORF in all or most species considered, if it had evolved neutrally along the respective lineages of the species tree. In parallel, our approach tests whether the number of non-synonymous substitutions that accumulated since the duplication event along the different branches of the species phylogeny is consistent with neutral evolution or, alternatively, selective constraint (Li, 1997).

Specifically, we first reconstruct the ancestral state of the ORF at the time-point of the oldest speciation event following gene duplication (Fig. 1) using a maximum parsimony or maximum likelihood approach (Felsenstein, 1996; Yang, 1997, Methods). To this end, we use the sequences of the gene copy in the species analyzed and an outgroup sequence (e.g. the immediate duplicate counterpart from one or more species). Next, we repeatedly simulate the evolution of this ancestral sequence throughout the species phylogeny assuming neutral evolution; point mutations as well as insertions and deletions (indels) accumulate according to a neutral model of sequence evolution.

We use a Kimura-2-parameters model of sequence evolution (Kimura, 1980), which allows for different transition/transversion rates in the simulations. Thus, the generally higher transition rate observed in eukaryotic genomes can be taken into account (Li, 1997). Optionally, a different substitution rate at CpG sites—which usually evolve rapidly in mammals (Cooper, 1993)—can be specified. Speciation times as well as neutral substitution and indel rates, specified *a priori* before the simulations, are used to define the number of substitutions and indels that accumulate along the different lineages of the species tree. After each simulation, the number of stop codons and indels that cause a shift in the reading frame (i.e. which are not a multiple of three) are counted. In addition, the number of non-synonymous substitutions (N_A) and synonymous substitutions (N_S) that are inserted throughout the phylogeny are recorded. The distribution of the number of synonymous substitutions after multiple rounds of simulations can also be used to adjust the specified point substitution rate. For example, if the distribution of synonymous substitutions during the simulations is consistently lower than the observed number in the phylogeny, the substitution rate may be increased for new rounds of simulations.

Two tests are used to assess the functionality of an ORF in our method. $Test_{dis}$ estimates the probability (P_{dis}) of preserving the integrity of the putative coding region under study. P_{dis} corresponds to the proportion of simulated datasets that show a number of deleterious (frame-disrupting) mutations (stop codons and frameshifts) smaller or equal to the observed number. $Test_{NaNs}$ compares the observed N_A/N_S ratio with its null distribution as estimated by the simulations. The probability associated to this test—which measures the deviation of the number of non-synonymous substitutions from the neutral expectation (P_{NaNs})—is estimated by the proportion of simulated datasets that show an N_A/N_S ratio smaller or equal to the observation.

4.2 Simulated datasets

To assess the power and accuracy of our approach, we generated simulated datasets using different combinations of parameters that may affect its behaviour: (1) the length of the ORF, (2) the age of the gene copy (i.e. the number of species in which the gene duplicate is present and/or has been sequenced) and (3) the extent of selection on the gene copy in the phylogeny (Supplementary Tables 1, 2 and 4).

With respect to the latter, we considered two major evolutionary fates of the simulated ancestral sequences: (1) Selective preservation of the duplicate gene, i.e. no indels are added and non-synonymous over synonymous rates (K_A/K_S) are kept low (nucleotide substitutions under purifying selection) and (2) the ORF is evolving neutrally, i.e. indels and substitutions accumulate according to a neutral model of sequence evolution. The first series of simulated data were generated to evaluate the power of the method in detecting functional genes, whereas the second series was used for assessing the accuracy (false positive rate) of the approach (see below).

The specific parameters for simulating the datasets were chosen to be representative of data that can be sampled in primate species, including humans. We considered four different phylogenetic ages of gene copies in the human genome, where the copy is present in different hominoid species (divergence time estimates from Goodman, 1999): humans-chimpanzees (~6 million years, MY), humans-African apes (~8 MY), humans-great apes (~14 MY) and hominoids (~18 MY). In addition, we used six different ORF lengths (600, 900, 1200, 1401, 2100 and 3000 bp) and three different K_A/K_S values (0.1, 0.5 and 1, respectively), reflecting different extents of purifying selection or neutral evolution. Other simulation conditions, such as substitution rates (and indel rates for the neutral scenario), speciation times and codon frequencies, were fixed to current estimates from primates (Methods). For each

of the resulting parameter combinations, we generated 1000 simulated datasets (Supplementary Tables 1–4).

4.3 Power and accuracy

To assess the power and accuracy of both tests ($test_{dis}$ and $test_{NaNs}$), we counted the number of times (out of 1000 different simulated datasets) for which the tests were significant at the 5, 1 and 0.1% levels (P_{dis} and P_{NaNs} below 0.05, 0.01, and 0.001, respectively).

Our results show that the power of $test_{dis}$ is strongly correlated with the length and phylogenetic distribution of the gene copy (Supplementary Tables 1 and 2). This is expected, since the probability to accumulate deleterious mutations in a sequence should increase with these parameters. For example, the test is powerful (~87–95% success in detecting functionality of the gene copy) when all four hominid sequences are available ($P_{dis} < 0.05$), in spite of a short (600 bp) coding sequence (100% success when coding sequence length ≥900 bp). When sequences from humans and the African apes are sampled, a coding region similar to the mean/median coding sequence length of human genes (~1400 bp; Lander *et al*., 2001) is required to achieve a similarly high success rate. To reliably detect functionality of ORFs in human–chimpanzee comparisons using $test_{dis}$, the coding region should be larger; a coding region of 2100 bp leads to a high success rate of 100%, whereas significantly shorter sequences result in few significant tests. Importantly, the power of $test_{dis}$ is independent of the extent of selective constraint (strong or weaker purifying selection), since very similar results are obtained under both selective scenarios (Supplementary Tables 1 and 2).

In addition, we were interested to examine the power of the approach in detecting selective preservation of a gene copy specifically on the human lineage (after the human–chimpanzee split). To this end, we performed $test_{dis}$ on the human lineage using the simulated data of gene copies present in humans and chimpanzee. The results show that a coding sequence of 3000 bp or more is required to identify functional genes with this test for such a scenario (Supplementary Table 3).

To assess the false positive rate of $test_{dis}$, we applied our simulation program to the datasets generated under neutral evolution. Overall, the type I error rate of this test varies between 0 and 11% at the 0.05 significance level (Supplementary Table 4). The number of false positives tends to increase with the number of species sampled as well as the length of sequences. However, this is more than compensated by the increase in power of deeper phylogenies/longer sequences (see above). Nevertheless, this result suggests the use of a lower significance level (P-values < 0.1%) when deeper phylogenies and/or longer sequences are tested.

To assess the performance of $test_{NaNs}$, we evaluated the accumulation of non-synonymous and synonymous substitutions that are monitored throughout the simulations. At low K_A/K_S (0.1), the test is very powerful even when the sequence is short and the phylogenetic distribution of the sequence is restricted (Supplementary Table 1). For example, for a sequence present only in humans and chimpanzees the success rate of the test is ~75% for an ORF of 600 bp ($P < 0.05$) and increases to over 95% when the ORF is 900 bp or longer. A similar (somewhat lower) success rate is obtained when testing specifically the human branch (Supplementary Table 3). However, to achieve reasonable power (>70% success rate) at $K_A/K_S = 0.5$ with this test (Supplementary Table 2), a sequence needs to be phylogenetically more widely distributed (present in all hominoids or in humans/great apes, depending on the length of the sequence) or have a length of at least 3000 bp. Notably, $test_{NaNs}$ virtually never rejects the null hypothesis of neutrality when applying the test to datasets generated under neutral evolution (Supplementary Table 4). Thus, this test appears to be conservative.

To further evaluate the behaviour of this test, we compared it with the widely used maximum likelihood selection tests (Yang and Bielawski, 2000) implemented in the program codeml of the PAML package (Yang, 1997). For each simulated dataset, we used a likelihood ratio test to examine whether the K_A/K_S ratio on all branches after the oldest speciation event (i.e. the first common node after duplication) is significantly smaller than one (see Methods).

$Test_{NaNs}$ and the likelihood ratio test performed with codeml behaved overall similarly, showing good power for sequences under strong purifying selection ($K_A/K_S = 0.1$). However, $test_{NaNs}$ shows higher success rates in all comparisons, in particular for less constrained sequences. For example, for an intact ORF of 1200 bp present in humans and great apes ($K_A/K_S = 0.5$), $test_{NaNs}$ detects selective constraint in 95.3% tests ($P < 0.05$), whereas codeml shows a significant result in <50% of the cases (Supplementary Table 2). However, it should be noted that a fair comparison of the two methods is difficult, because they use different input data (e.g. our method also uses divergence time information, contrary to codeml) and codeml estimates more parameters. Nevertheless, the results suggest that $test_{NaNs}$ has good power to detect purifying selection.

When comparing $test_{dis}$ and $test_{NaNs}$, several differences stand out. For sequences under strong purifying selection ($K_A/K_S = 0.1$), $test_{NaNs}$ provides on average better power than $test_{dis}$, being less dependent on the length and the phylogenetic distribution of the sequence. However, under a scenario of less intense constraint ($K_A/K_S = 0.5$), $test_{dis}$ is overall superior, if coding sequences are at least of approximately average length (1200 bp or more), present in at least humans/African apes, or both. This is probably because $test_{dis}$ is largely independent of substitutional patterns as reflected by K_A/K_S. This advantage of $test_{dis}$ relative to $test_{NaNs}$ is expected to often be even more pronounced for real data, where new gene copies may show average K_A/K_S values >0.5, e.g. because of positive selection at certain sites and/or a neutral phase of evolution upon emergence (Long *et al*., 1999).

It is also important to point out that $test_{dis}$ generally yields lower P-values than $test_{NaNs}$—irrespective of the extent of purifying selection—if the sequence length and/or phylogenetic distribution exceeds certain thresholds (Supplementary Tables 1 and 2), which allows for extensive multiple test corrections (Marques *et al*., 2005). For example, a gene with a coding sequence of length 1200 bp carried by humans and great apes yields $P_{dis} < 10^{-3}$ in ~71% of replicates ($K_A/K_S = 0.5$), whereas $test_{NaNs}$ shows $P_{NaNs} < 10^{-3}$ in only 28/1000 replicates (Supplementary Table 2). Performing a larger number of simulations (100 000) for individual (real) genes reveals that P_{dis} values are lower than 10^{-5} for genes of approximately average length (1400–1500 bp) present in hominoids, whereas P_{NaNs} is generally much higher, thus being less likely to show significance after extensive multiple test corrections [see below and Marques *et al*. (2005)]. Thus, $test_{dis}$ is expected to be powerful in large-scale comparative genome analysis of lineage-specific genes, as was exemplified in our systematic screen

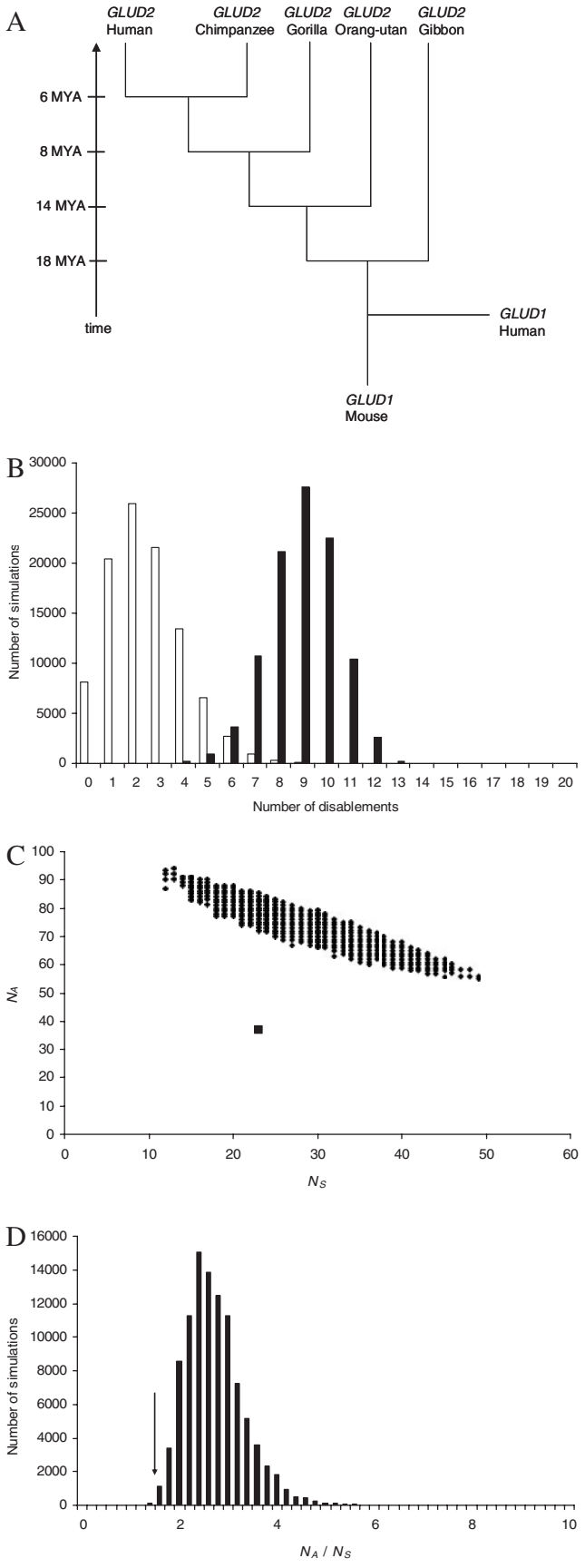

Fig. 2. Illustration of the simulation results for *GLUD2*. (**A**) Tree topology and sequences used. (**B**) Distribution of the number of disruptions observed in 10^5 simulations of *GLUD2* evolution under neutrality. The frequency distribution of stop codons is shown in white and that of deleterious indels in black. (**C**) Non-synonymous (N_A) and synonymous (N_S) substitutions observed in 105 simulations of neutral *GLUD2* evolution (diamonds). The black square indicates the observed N_A and N_S substitutions in the *GLUD2* primate phylogeny. (**D**) N_A/N_S ratio in 105 simulations of *GLUD2* evolution in the primate lineages. The arrow indicates the observed N_A/N_S ratio in the *GLUD2* primate phylogeny.

for functional retrocopies in the human genome (Marques *et al*., 2005).

4.4 Application to real data

To further test the behaviour of our simulation approach, we applied it to four genes (*GLUD*2, *ECP*, *GMCL*2 and *OdsH*) that recently emerged in primates or *Drosophila* and were shown or predicted to be functional by both evolutionary and experimental analyses (Burki and Kaessmann, 2004; Domachowske *et al*., 1998; Marques *et al*., 2005; Plaitakis *et al*., 2003; Rosenberg and Dyer, 1995; Ting *et al*., 2004; Zhang *et al*., 1998). These genes also represent the two major duplication mechanisms, segmental gene duplication (*ECP* and *OdsH*) and retroposition (*GMCL*2 and *GLUD*2). We performed 100 000 simulations of neutral evolution for each of these datasets using our method.

We applied our method to a gene, *GLUD*2, which emerged after a retroduplication event of the transcript of its parent (*GLUD*1) in the hominoid ancestor, less than 23 million years ago (Burki and Kaessmann, 2004, Fig. 2A), and is probably involved in glutamate flux regulation in the nervous system and potentially other tissues (Plaitakis *et al*. 2003). Although this retrogene has experienced positive selection at a subset of sites (Burki and Kaessmann, 2004, Supplementary Table 5), $\text{test}_{\text{NaNs}}$ reveals that it was overall shaped by purifying selection since the duplication event to sustain functionality ($P_{\text{NaNs}} = 0.013$, Fig. 2C and D). Test_{dis} also strongly supports the integrity and hence functionality of *GLUD*2, since each of the 10^5 simulations of a neutral process for this gene copy showed at least one disruption ($P_{\text{dis}} < 10^{-5}$, Fig. 2B).

The eosinophil-derived neurotoxin (*EDN*) and eosinophil cationic protein (*ECP*) genes belong to the ribonuclease gene family and are probably involved in antiviral and antibacterial defense (Domachowske *et al*., 1998; Rosenberg and Dyer, 1995). The *ECP* gene was generated by duplication of the *EDN* gene in the ancestor of hominoids and Old World monkeys ~25–35 million years ago (Zhang *et al*., 1998). The simulation of a neutral process for this gene copy using our method (Supplementary Figure 1A) indicates that the integrity of its ORF has been selectively maintained since the duplication event ($P_{\text{dis}} = 0.003$, Supplementary Figure 1B). It also shows that the observed number of non-synonymous substitutions is higher than those simulated under neutrality (Supplementary Figure 1C). This observation is consistent with the hypothesis that, after duplication, positive Darwinian selection operated in the early stage of evolution of the *ECP* gene to enhance its novel anti-pathogen function (Zhang *et al*., 1998). However, the ratio of non-synonymous to synonymous substitutions is found within its null distribution (Supplementary Figure 1D), rendering $\text{test}_{\text{NaNs}}$ non-significant ($P_{\text{NaNs}} = 0.919$), perhaps because only a subset of sites evolve under positive selection. The use of

a sensitive site-specific maximum likelihood analysis implemented in codeml (see Methods for details) indeed shows that whereas purifying selection has shaped many *ECP* codons (43.90%) during primate evolution, 10.61% of the codons evolve neutrally, and 45.49% of the codons were the target of positive selection (Supplementary Table 5), consistent with the previous findings by Zhang *et al*. (1998).

During a reanalysis of the *GMCL2* gene (a retrogene probably important for sperm formation, Marques *et al*., 2005), the comparison between the number of synonymous substitutions that occurred in the observed phylogeny and the N_S distribution generated by the simulations revealed that taking into account the hypermutability of CpG sites (10 times higher than the specified substitution rate, Cooper *et al*., 2001) provides a better fit to the data [compare Figure S5 in Marques *et al*. (2005) with Supplementary Figure 2C). Our new analysis confirms the functionality of the *GMCL2* retrocopy ($P_{NaNs} < 10^{-3}$, $P_{dis} < 10^{-5}$; Supplementary Figure 2B–D). In general, we advise to use our method both with and without an additional CpG rate and then analyze which model corresponds better to the observed data, since the rates and abundance of CpG sites may vary between different genomic regions (Cooper *et al*., 2001).

To test our method using data from other species than primates, we chose the *Odysseus* locus (*OdsH*), a potential 'speciation gene' that arose by gene duplication in the common ancestor of the *Drosophila* genome ~40 million years ago (Ting *et al*., 2004). We applied our approach using the available *OdsH* sequences from *Drosophila yakuba* (outgroup) and sequences from the clade encompassing *Drosophila melanogaster* and its three sibling species *Drosophila sechellia*, *Drosophila simulans* and *Drosophila mauritiana* (Supplementary Figure 3A). The deepest split in this clade corresponds to only ~2 million years. In spite of this shallow phylogeny, $test_{dis}$ strongly supports functionality of this locus ($P_{dis} < 10^{-5}$, Supplementary Figure 3B). This is probably owing to the fact that substitution and indel rates in *Drosophila* (Methods, Li, 1997; Parsch, 2003) are significantly higher than those in primates, rendering the test more powerful in this genus. $Test_{NaNs}$ is not significant ($P_{NaNs} = 0.3$, Supplementary Figure 3C and D), consistent with the notion that *OdsH* has accumulated many non-synonymous substitutions during its evolution in *D. melanogaster* and its sibling species (Ting *et al*., 2004).

DISCUSSION

A number of tools are available to assess the conservation and/or the functionality of duplicate gene copies with a wide phylogenetic distribution, such as those shared among different mammalian lineages (Boffelli *et al*., 2004a; Frazer *et al*., 2003; Nekrutenko *et al*., 2003, 2002). However, supporting functionality of gene duplicates with a limited phylogenetic distribution—such as genes that recently emerged on the primate lineage leading to humans—is difficult.

Here we evaluated in detail the behaviour and applicability of a new evolutionary method to predict the functionality of young gene copies, which was recently successfully applied to the detection of primate-specific retrogenes (Marques *et al*., 2005).

$Test_{dis}$ of our method uses the same basic evolutionary principle—the probability of preserving the integrity of an ORF under neutrality—as that of Zhang (2003). However, $test_{dis}$ represents an integrated method that differs fundamentally in several ways from the method of Zhang. First and foremost, our method makes use of the information inherent in the entire phylogeny of an ORF and is hence not limited to a single sequence/lineage. As a consequence, it is expected to be much more powerful for the functional prediction of genes present in at least two species. It may thus also test scenarios with different evolutionary rates and also fates of an ORF in different lineages (e.g. if a gene evolves neutrally along a certain lineage and therefore carries ORF disruptions in some sequences, while it is preserved in others). Second, contrary to the method of Zhang that provides as an output only the half-life of the single input sequence, our method compares the distribution of stop codon/indels in the phylogeny (as exemplified in Fig. 2 and Supplementary Figures 1–3) generated under the neutral model with the observed data and provides the associated *P*-value of the test. Third, our method automatically reconstructs the common ancestral node of the ORF, from which forward simulations of sequence evolution proceed. Fourth, sequence evolution is simulated using a more sophisticated model of sequence evolution that takes into account the transition/transversion bias and specific rates at CpG sites. Finally, our method includes the additional $test_{NaNs}$, which is carried out in parallel to $test_{dis}$ and assesses the selective accumulation of non-synonymous substitutions over time. Therefore, in conclusion, our method represents the first comprehensive and integrated method to test for the functional preservation of an ORF with a restricted phylogenetic distribution.

Our benchmarking showed that $test_{dis}$ has strong power in supporting functionality of recently emerged ORFs. This is especially true for ORFs of at least average length (~1400 bp) that are carried by at least a few species (e.g. humans and African apes). Our analysis of human lineage-specific simulations suggests that long ORFs (> 3000 bp) even allow to support functionality for human-unique sequences, but this possibility will be investigated further (see also below). Our second approach, based on the accumulation of non-synonymous/synonymous substitutions ($test_{NaNs}$), appears to have high sensitivity in detecting purifying selection, even with limited data.

Importantly, we show that $test_{dis}$ is independent of the substitutional pattern, which is usually used to infer selective preservation in comparative genomics approaches (Nekrutenko *et al*., 2002). Thus, $test_{dis}$ may often be superior relative to methods based on K_A/K_S selective signatures, in particular when screening for new genes, because such genes are often differentially affected by selection in certain lineages or at specific sites, thus complicating K_A/K_S analyses. $Test_{dis}$ generally also shows lower *P*-values than non-synonymous/synonymous approaches ($test_{NaNs}$ and codeml), especially when purifying selection is not strong. Thus, $test_{dis}$ may be very powerful in functionality predictions when applied on a larger scale. In support of this, we previously showed that our method has good power to pinpoint recent functional retrogenes, even though some ORFs were not intact in all evolutionary lineages where the copy is present (Marques *et al*., 2005).

Our method should be of value for evolutionary biologists to test their gene of interest. It should also be useful for large-scale genome annotation and evolutionary analyses, also because $test_{dis}$ generally provides low *P*-values that allow for extensive multiple test corrections. The chimpanzee genome has recently become available (Mikkelsen *et al*., 2005) and most hominoid genomes will be

available soon (Dennis, 2005). Thus, it will be very interesting to systematically test the functionality of genes recently acquired on the primate lineage leading to humans or other primate-specific genes (see also below). To this end, it may also be useful to combine ReEVOLVER, which implements our method, with conservation-profile based programs such as eShadow (Boffelli *et al*., 2003; Ovcharenko *et al*., 2004), in particular for poorly annotated genomic regions such as recent segmental duplications; eShadow may identify potentially conserved ORFs within such regions, while ReEVOLVER can be used to support the functionality of these ORFs.

We note that although our method has been presented mostly in the context of recently emerged duplicate gene copies, it is equally well suited to assess functionality of intact ORFs that degenerated or were lost in other lineages, such as olfactory and vomeronasal pheromone receptor genes (Gilad *et al*., 2003; Zhang and Webb, 2003), or other putative genes in primates (Newman *et al*., 2005). Our method is also suitable to distinguish between loss of function and positive selection (both leading to elevated K_A/K_S) in a specific lineage of an 'old', phylogenetically widespread gene as was previously demonstrated with respect to the *ASPM* gene using a similar methodological principle (Zhang, 2003).

To enhance the power for very recently emerged gene copies or single evolutionary lineages (such as human-unique sequences), we plan to extend our method so that it can make use of multiple sequences within a species. This will complement a recent variant of eShadow, which was able to predict functional genomic elements using a large number of population sequences from a highly polymorphic urochordate (Boffelli *et al*., 2004b). Analogous to the between-species simulations, the program will then simulate neutral evolution since the most recent common ancestor of the sequences from different individuals based on the coalescent (Hartl and Clark, 1997). Because recent advances in large-scale sequencing make it feasible to sequence an ORF from a large number of individuals, this will greatly increase the power of the method for very recent duplicate genes or species-unique genes. Thus, this modification will be particularly relevant for the study of the functionality of lineage-specific ORFs—such as human-unique (duplicate) genes—that are emerging in the course of the ongoing hominoid sequencing projects (Cheng *et al*., 2005; Newman *et al*., 2005).

ACKNOWLEDGEMENTS

The authors thank Victor Jongeneel and the Vital-IT unit for computational support; Laurent Excoffier for comments on the manuscript; and the lab members of H.K. for helpful discussions. This research was supported by funds to H.K. from the Center for Integrative Genomics (University of Lausanne), the Swiss National Science Foundation (grant 3100A0-104181) and the European Union (grant PKB140404).

Conflict of Interest: none declared.

REFERENCES

Betran,E. *et al*. (2002) Retroposed new genes out of the X in *Drosophila*. *Genome Res*., **12**, 1854–1859.

Boffelli,D. *et al*. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391–1394.

Boffelli,D. *et al*. (2004a) Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet*., **5**, 456–465.

Boffelli,D. *et al*. (2004b) Intraspecies sequence comparisons for annotating genomes. *Genome Res*., **14**, 2406–2411.

Britten,R.J. (2002) Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc. Natl Acad. Sci. USA*, **99**, 13633–13635.

Burki,F. and Kaessmann,H. (2004) Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nat. Genet*., **36**, 1061–1063.

Cheng,Z. *et al*. (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*, **437**, 88–93.

Cooper,D.N. (1993) *Human Gene Mutation*. Bios Scientific, Oxford.

Cooper,D.N., Antonarakis,S.E. and Krawczak,M. (2001) The nature and mechanisms of human gene mutation. In Scriver CR, Beaudet AL, Sly WS, Valle D, Childs B, Kinzler KW, Vogelstein B (eds), *The Metabolic and Molecular Bases of Inherited Disease*. McGraw-Hill Co., New York, pp. 259–291.

Dennis,C. (2005) Chimp genome: branching out. *Nature*, **437**, 17–19.

Domachowske,J.B. *et al*. (1998) Eosinophil cationic protein/RNase 3 is another RNase A-family ribonuclease with direct antiviral activity. *Nucleic Acids Res*., **26**, 3358–3363.

Eichler,E.E. and Sankoff,D. (2003) Structural dynamics of eukaryotic chromosome evolution. *Science*, **301**, 793–797.

Emerson,J.J. *et al*. (2004) Extensive gene traffic on the mammalian X chromosome. *Science*, **303**, 537–540.

Felsenstein,J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol*., **266**, 418–427.

Feral,C. *et al*. (2001) Human testis expresses a specific poly(A)-binding protein. *Nucleic Acids Res*., **29**, 1872–1883.

Frazer,K.A. *et al*. (2003) Cross-species sequence comparisons: a review of methods and available resources. *Genome Res*., **13**, 1–12.

Gilad,Y. *et al*. (2003) Human specific loss of olfactory receptor genes. *Proc. Natl Acad. Sci. USA*, **100**, 3324–3327.

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

Hartl,D.L. and Clark,A.G. (1997) *Principles of Population Genetics*. Sinauer Associates, Sunderland, MA.

Kimura,M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol*., **16**, 111–120.

Kliman,R.M. *et al*. (2000) The population genetics of the origin and divergence of the *Drosophila* simulans complex species. *Genetics*, **156**, 1913–1931.

Lander,E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Li,W.H. (1997) *Molecular Evolution*. Sinauer Associates, Sunderland MA.

Li,W.H. *et al*. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol*., **2**, 150–174.

Long,M. *et al*. (1999) Origin of new genes and source for N-terminal domain of the chimerical gene, jingwei, in *Drosophila*. *Gene*, **238**, 135–141.

Long,M. *et al*. (2003) The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet*., **4**, 865–875.

Marques,A. *et al*. (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol*., **3**, e357.

Mikkelsen,T.S. *et al*. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.

Nekrutenko,A. *et al*. (2002) The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res*., **12**, 198–202.

Nekrutenko,A. *et al*. (2003) ETOPE: Evolutionary test of predicted exons. *Nucleic Acids Res*., **31**, 3564–3567.

Newman,T.L. *et al*. (2005) A genome-wide survey of structural variation between human and chimpanzee. *Genome Res*., **15**, 1344–1356.

Ovcharenko,I. *et al*. (2004) eShadow: a tool for comparing closely related sequences. *Genome Res*., **14**, 1191–1198.

Parsch,J. (2003) Selective constraints on intron evolution in *Drosophila*. *Genetics*, **165**, 1843–1851.

Plaitakis,A. *et al*. (2003) Study of structure–function relationships in human glutamate dehydrogenases reveals novel molecular mechanisms for the regulation of the nerve tissue-specific (GLUD2) isoenzyme. *Neurochem. Int*., **43**, 401–410.

Rice,P. *et al*. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*., **16**, 276–277.

Rosenberg,H.F. and Dyer,K.D. (1995) Eosinophil cationic protein and eosinophil-derived neurotoxin. Evolution of novel function in a primate ribonuclease gene family. *J. Biol. Chem*., **270**, 21539–21544.

Samonte,R.V. and Eichler,E.E. (2002) Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.*, **3**, 65–72.

Silva,J.C. and Kondrashov,A.S. (2002) Patterns in spontaneous mutation revealed by human-baboon sequence comparison. *Trends Genet.*, **18**, 544–547.

Thornton,K. (2003) Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics*, **19**, 2325–2327.

Ting,C.T. *et al.* (2004) Gene duplication and speciation in *Drosophila*: evidence from the *Odysseus* locus. *Proc. Natl Acad. Sci. USA*, **101**, 12232–12235.

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. and Bielawski,J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol. Evo.*, **15**, 496–503.

Yang,Z. *et al.* (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.

Yang,Z. *et al.* (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.*, **22**, 1107–1118.

Yi,S. *et al.* (2002) Slow molecular clocks in Old World monkeys, apes, and humans. *Mol. Biol. Evol.*, **19**, 2191–2198.

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

Zhang,J. and Webb,D.M. (2003) Evolutionary deterioration of the vomeronasal pheromone transduction pathway in catarrhine primates. *Proc. Natl Acad. Sci. USA*, **100**, 8337–8341.

Zhang,J. *et al.* (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl Acad. Sci. USA*, **95**, 3708–3713.