

Gene Conversion Among Paralogs Results in Moderate False Detection of Positive Selection Using Likelihood Methods

Claudio Casola · Matthew W. Hahn

Received: 15 December 2008 / Accepted: 15 April 2009 / Published online: 27 May 2009
© Springer Science+Business Media, LLC 2009

Abstract Previous studies have shown that recombination between allelic sequences can cause likelihood-based methods for detecting positive selection to produce many false-positive results. In this article, we use simulations to study the impact of nonallelic gene conversion on the specificity of PAML to detect positive selection among gene duplicates. Our results show that, as expected, gene conversion leads to higher rates of false-positive results, although only moderately. These rates increase with the genetic distance between sequences, the length of converted tracts, and when no outgroup sequences are included in the analysis. We also find that branch-site models will incorrectly identify unconverted sequences as the targets of positive selection when their close paralogs are converted. Bayesian prediction of sites undergoing adaptive evolution implemented in PAML is affected by conversion, albeit in a less straightforward way. Our work suggests that particular attention should be devoted to the evolutionary analysis of recent duplicates that may have experienced gene conversion because they may provide false signals of positive selection. Fortunately, these results also imply that those cases most susceptible to false-positive results—i.e., high divergence between paralogs, long conversion tracts—are also the cases where detecting gene conversion is the easiest.

Keywords Recombination · Adaptive evolution · Gene duplicates · PAML

Introduction

The identification of protein-coding genes evolving by adaptive natural selection can show the fundamental ways in which organisms adapt to their environment. One of the clearest signatures of positive selection in the coding region of genes is an excess of nonsynonymous substitutions per site (d_N) relative to synonymous substitutions per site (d_S), i.e., $d_N/d_S > 1$ (Hill and Hastie 1987; Hughes and Nei 1988). This test for positive selection can be applied either to single-copy orthologs from multiple species or to duplicated paralogs within a species. To address both the role of duplicate genes in organismal adaptation and the role of natural selection in maintaining duplicated genes, it is necessary to test for the signature of positive selection among paralogs (Bielawski and Yang 2003).

Different methods have been developed to estimate d_N/d_S , from simple counting methods (e.g., Nei and Gojobori 1986) to more complex, and more sensitive, codon-substitution models that rely on likelihood calculations (Muse and Gaut 1994; Nielsen and Yang 1998; Yang et al. 2000). CODEML, which is implemented in the PAML suite of programs, is one of the most popular likelihood tools used to estimate d_N/d_S (Yang 2007). CODEML allows pairs of nested models with and without positive selection to be tested in a likelihood ratio framework to determine if adaptive evolution has occurred. Furthermore, CODEML implements an empiric Bayes approach to identify individual codons undergoing adaptive evolution (Yang et al. 2005). A growing number of genes evolving under positive selection, including duplicated genes, have been discovered

Electronic supplementary material The online version of this article (doi:10.1007/s00239-009-9241-6) contains supplementary material, which is available to authorized users.

C. Casola (✉) · M. W. Hahn
Department of Biology and School of Informatics, Indiana University, 1001 E. 3rd Street, Bloomington, IN 47405, USA
e-mail: ccasola@indiana.edu

using these methods (Birtle et al. 2005; Des Marais and Rausher 2008; Hahn et al. 2007a, b).

The models implemented in CODEML follow the assumption that branch lengths and the topology of the phylogenetic tree do not vary across the sequences of interest. Recombination causes variation in branch lengths across a sequence, and although the models implemented in CODEML allow for variation in selective constraint across a sequence (i.e., d_N/d_S), they assume constant synonymous distances (d_S). Likewise, CODEML calculates the likelihood of the data over a single prespecified tree topology; however, recombination changes the topology from one base to the next. As a result of violating basic assumptions of the underlying model, analyses of recombining sequences show incorrect signatures of positive selection (Anisimova et al. 2003; Scheffler et al. 2006; Shriner et al. 2003). However, it is not known whether the branch-length or the topology assumption is more sensitive to violation (Anisimova et al. 2003).

Although paralogs do not recombine in the same manner as allelic sequences, “ectopic” gene conversion among paralogs can result in the exchange of sequence among duplicated genes. Gene conversion is the nonreciprocal exchange between a “donor” sequence and an “acceptor” sequence and represents one of the most common outcomes of double-stranded breaks between two homologous sequences (Chen et al. 2007; Li 1997; Slightom et al. 1980). Ectopic gene conversion has been documented in a plethora of organisms, including bacteria, plants, fungi, and metazoans (Drouin et al. 1999; Gerton et al. 2000; Mondragon-Palomino and Gaut 2005; Nielsen et al. 2003; Santoyo and Romero 2005; Semple and Wolfe 1999). Gene conversion can violate some of the same assumptions that cause PAML to incorrectly infer positive selection in the presence of recombination. However, because it is relatively common to analyze only pairs of paralogous sequences, there can be no violations of the assumption of constant tree topology in these cases. It may therefore be true that rates of false-positive inferences of natural selection are much lower when analyzing paralogs. In this study, we carried out extensive simulations to examine the rate and causes of false-positive results when considering gene conversion between paralogous sequences.

Methods

Sequence data sets were generated by Monte Carlo simulations using the EVOLVER program of the PAML 4 package (Yang 2007). All data sets were simulated without positive selection but instead with two site classes ($d_N/d_S = 0$ and $d_N/d_S = 1$), both with frequency 0.5. A uniform

codon frequency of 1/61 was applied and the transition-to-transversion rate ratio was set to $\kappa = 2$.

Two groups of data sets were built to examine the effect of the number of sequences included in analyses. The first group consists of data sets formed by simulating 3 coding sequences of 500 codons replicated 1000 times, with 5 different tree lengths. Pairwise distances between ingroup sequences, represented by d_S values, were fixed at 0.02, 0.04, 0.1, 0.2, and 0.4; these distances represent common divergence values between paralogs analyzed in the literature (e.g., Han et al. 2009). The third sequence represented the outgroup (from the same genome) and was arbitrarily set to have twice the distance from each ingroup sequence as the distance between ingroup sequences. Artificially converted data sets were built from the first group of replicates as follows: converted tracts of 50, 167, and 250 codons (i.e., 1/10, 1/3, and 1/2 of the total sequence length) were transferred from a donor to an acceptor sequence, starting from the 100th codon. These conversion tract lengths are also representative of lengths seen in nature (Benovoy and Drouin 2009; Chen et al. 2007; Gerton et al. 2000; Semple and Wolfe 1999). Four different experimental conditions were established as described later in the text.

In the second group, each data set was represented by 100 replicates of 10 coding sequences with 500 codons. For these data sets, we used the tree shown in Supplementary Fig. 1. Gene conversion was simulated between genes at different genetic distances using the second sequence as the acceptor and sequences 1, 3, or 6 as donor. Converted tracts of 50, 167, and 250 codons were transferred from each donor to the acceptor sequence, starting from the 100th codon.

Positively selected sequences and codons were detected using the CODEML program of the PAML 4 package (Yang 2007). Two different sets of models that allow d_N/d_S to vary among sites were compared: (1) the M1a and M2a models and (2) the M7 and M8 models. Model M1a allows the site classes $d_N/d_S = 1$ and $0 < d_N/d_S < 1$, whereas model M2a has the same site classes of M1a and a third class with $d_N/d_S > 1$ (Nielsen and Yang 1998; Wong et al. 2004; Yang et al. 2000, 2005). Model M7 includes several site classes with d_N/d_S ratios following the beta-distribution $B(p, q)$, whereas model M8 extends model M7 with a further class with $d_N/d_S > 1$. Likelihood ratio tests (LRTs) were carried out between models M1a/M2a and models M7/M8 as described in (Yang 2007).

For “branch-site” analysis, we used the same data sets produced by the EVOLVER program and ran CODEML with the parameters specified in the PAML 4 manual to perform “test 2” (Yang 2007; Zhang et al. 2005). In the alternative hypothesis, we fixed initial $d_N/d_S = 1.5$. As suggested in the PAML 4 documentation, we performed

the branch-site analysis with different initial values (analyses performed with an initial value of $d_N/d_S = 5$ did not diverge significantly from these outcomes).

Results and Discussion

Rate of False-Positive Results in Site Models with Gene Conversion

To determine the false-positive rate in the presence of gene conversion, we simulated protein-coding sequences evolving without positive selection and introduced conversion tracts. Sequences of length 500 codons were generated using the EVOLVER program in PAML (Yang 2007), with $d_N/d_S = 0.5$ (see Methods). Gene conversion was simulated by copying fragments of different length (50, 167, and 250 codons) from one sequence to another. Each tree was initially simulated with 3 sequences and then subject to 1 of 4 main treatments (Fig. 1): (I) conversion occurred between the two ingroup sequences, and only these two sequences were tested for positive selection; (II) conversion occurred between the two ingroup sequences, but all three sequences were included in the test for selection; (III) the outgroup sequence converted one of the ingroup sequences, but only the two ingroup sequences were tested for positive selection; or (IV) the outgroup sequence converted one of the ingroup sequences, and all of the sequences were included in the test for selection. Each treatment was simulated 1000 times for each of 5 different values of d_S and each conversion tract length. To estimate the false-positive rate for each experimental condition, we tested each simulated alignment for positive selection using likelihood ratio tests between two different sets of “site” models implemented in CODEML (M1a/M2a and M7/M8). For comparison we also estimated the false-positive rate in equivalent nonconverted data sets.

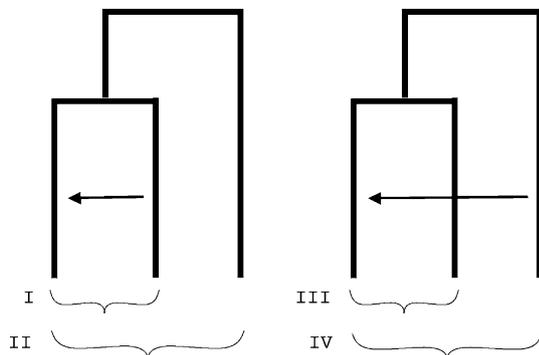


Fig. 1 Scheme of experimental conditions used in this study. All four conditions are shown, with the arrow indicating the direction of gene conversion

Our analysis shows that gene conversion can lead to a moderate increase in the proportion of genes erroneously identified as undergoing adaptive evolution (Fig. 2). Generally, the number of false-positive results is directly proportional to the genetic distance (d_S) between sequences and the length of converted tracts, whereas different sets of models (M1a/M2a or M7/M8) seem to produce similar outcomes (Fig. 2 and Supplementary Fig. 2). Larger conversion tracts and larger distances between paralogs (before conversion) result in higher numbers of false-positive rates, possibly because there is greater disparity in branch length among sites when these two values grow larger. Large conversion tracts between distant paralogs are often the easiest to identify (Sawyer 1989), which may make it easier to avoid these false-positive results (see later text).

We found that differences among the experimental conditions (i.e., conditions I through IV) were a major factor in determining false-positive rates. The outgroup-to-ingroup-converted data sets (conditions III and IV) showed at most the expected proportion of false-positive results at $p < 0.05$ (approximately 5%; Fig. 2c and d). However, ingroup-to-ingroup conversions (conditions I and II) had rates of type I error up to almost 33% (Fig. 2a and b; Supplementary Fig. 2A and B). This result was unexpected because conversion among ingroup sequences will not change the inferred tree topology. Based on previous results (Anisimova et al. 2003), we expected outgroup-to-ingroup conversion to have higher false-positive rates because they produce contrasting relations across different parts of the acceptor (ingroup) sequence.

In addition, experimental conditions in which only the two ingroup sequences were included in tests for positive selection (conditions I and III; Fig. 2a and c) had higher type I error rates compared with conditions that included an outgroup sequence (conditions II and IV; Fig. 2b and d). These outcomes were also unexpected because there is no possible way to change the topology of a tree that includes only two sequences.

One possible explanation for the increased rates of false-positive results mentioned previously is that the accuracy of the likelihood ratio test tends to be low for data sets with few sequences (Anisimova et al. 2001, 2002). Therefore, we performed a similar analysis on a data set with 10 sequences, with 3 possible simulated gene-conversion events between sequences at increasing genetic distance (see Methods and Supplementary Fig. 1). We then compared type I error rates between data sets with 10 sequences and the previously described data sets with 3 sequences using replicates with equal or similar pairwise genetic distances between acceptor and donor sequences. Conversion between the two close paralogs 1 and 2 ($1 \rightarrow 2$) in the tree with 10 sequences generated $<5\%$ false-positive

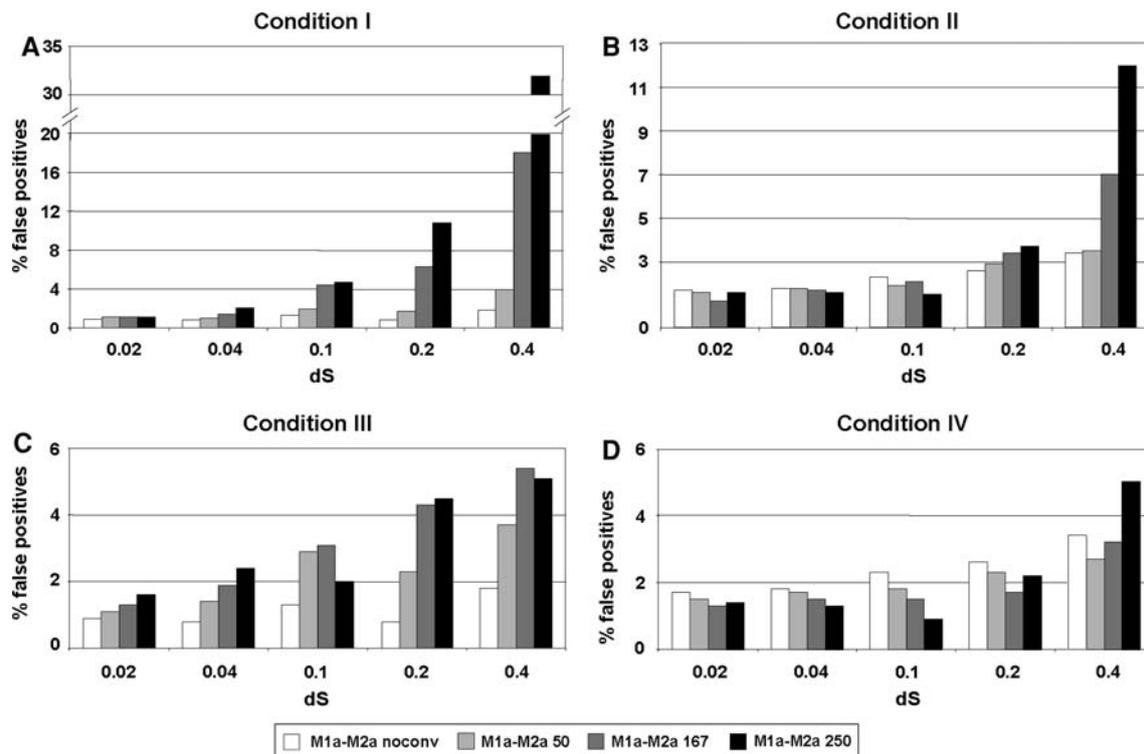


Fig. 2 Percentage of false-positive results in site models versus the pairwise genetic distance of ingroup sequences. Different experimental conditions using models M1a-M2a are compared (see text for

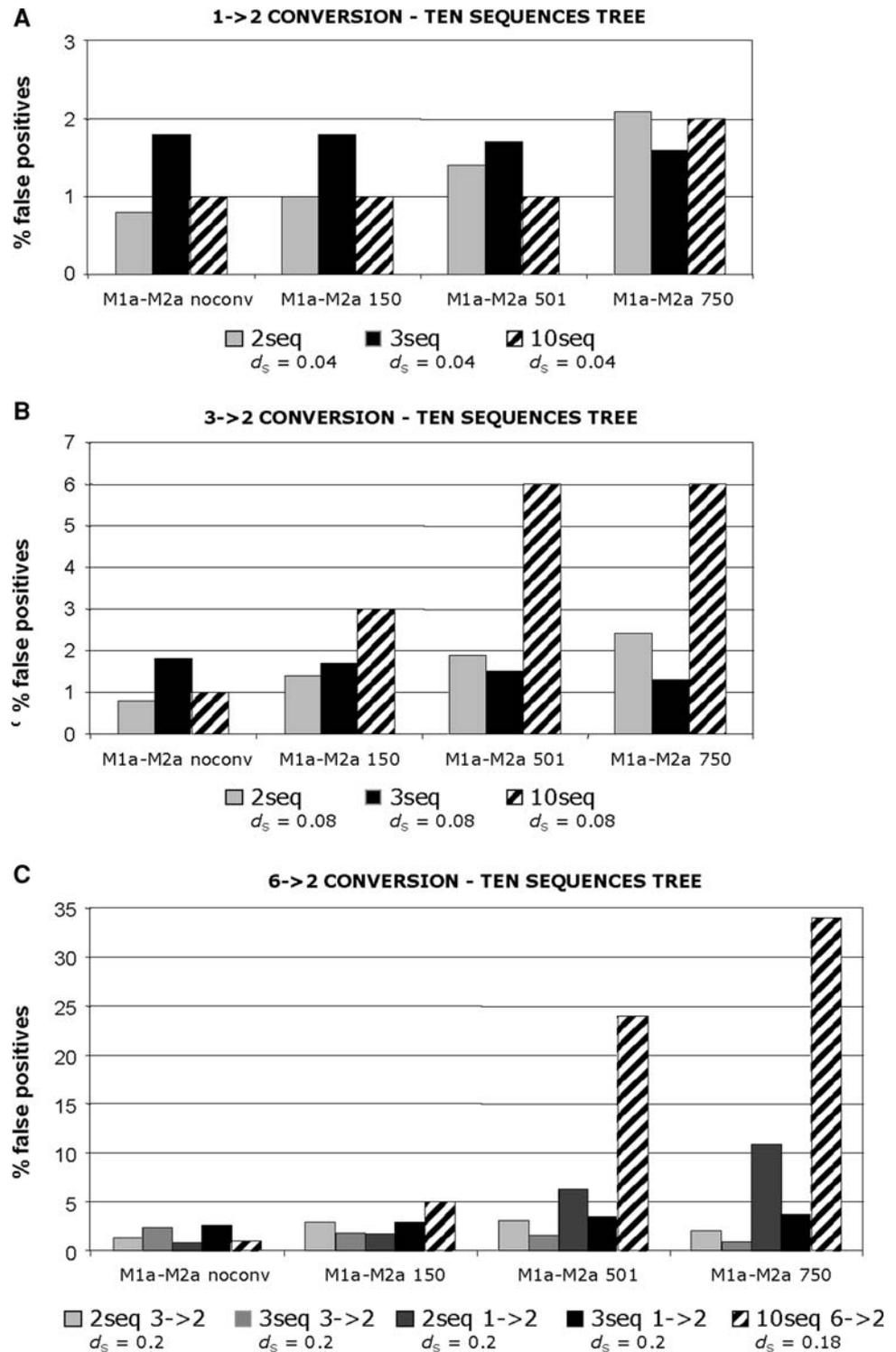
details). Noconv = data sets with no conversion. Note that the y-axis in the four panels is not on the same scale

results, similarly to the rate recovered from conditions I and II (pairwise $d_S = 0.04$ between ingroup sequences) in data sets with only 3 sequences (Fig. 3a; Supplementary Fig. 3B). In the second scenario involving the larger tree, we recreated a transfer from sequence 3 (the closest outgroup to paralogs 1 and 2) to sequence 2. The number of false-positive results in this case is slightly higher than in the corresponding replicates of conditions III and IV (pairwise $d_S = 0.04$ between ingroup sequences) described previously, but it is still not much greater than the expected 5% for either model comparison (Fig. 3b; Supplementary Fig. 3B). The third simulated conversion involved a transfer from sequence 6 to sequence 2 (6→2). Given the pairwise $d_S = 0.18$ between donor and acceptor sequence, we compared these replicates with data sets with pairwise ingroup-to-ingroup distance of $d_S = 0.1$ (conditions III and IV) and $d_S = 0.2$ (conditions I and II; Fig. 3c). In the latter scenario, the number of false-positive results is at least three times higher in replicates with 6→2 conversion than in data sets with only 2 or 3 sequences for conversion tracts of 167 and 250 codons (Fig. 3c; Supplementary Fig. 3C). This is in agreement with the reported results from LRTs between codon models using trees of different size with recombination among sequences (Anisimova et al. 2003).

Gene conversion between paralogs may occur repeatedly and at different times, producing an acceptor gene that

is a mosaic of sequences with different genetic distances from the donor gene(s). Because our simulations thus far have only considered extremely recent conversion events, and only one event per paralog, we further investigated the effects of these processes on the accuracy of CODEML. We again generated data sets with three sequences and simulated either one or two ingroup-to-ingroup conversion events occurring at different times since their split (see Methods and Supplementary Table 1). As observed with the other data sets (Figs. 2 and 3), the number of false-positive results was >5% only for the largest genetic distance between the sequences of the tree (pairwise $d_S = 0.4$ between ingroup sequences) and was higher when the outgroup sequence was removed from the analysis of positive selection, whether there was one (Fig. 4a and b) or multiple conversion events (Fig. 4c and d). Both sets of results also show that there are a larger number of false-positive results the more recently the conversion event occurred, regardless of the models being compared (Fig. 4; Supplementary Fig. 4). Finally, we found little difference in the number of false-positive results between data sets simulated with one or two conversion events, except for the highest divergence between ingroup paralogs, where data sets with two events showed approximately 5% more false-positive results than replicates with only one event (Fig. 4; Supplementary Fig. 4).

Fig. 3 Percentage of false-positives results in site models in data sets with 2, 3, and 10 sequences using models M1a-M2a. Sequence 2 in the tree is the fixed acceptor sequence, and donor sequences are sequence 1 (1→2), 3 (3→2), and 6 (6→2). Noconv = data sets with no conversion. Results obtained using different codon models and conversion tract lengths are shown (see text for further details). The pairwise d_s value between donor and acceptor sequences in each data set is shown. 2seq = condition I; 3seq = condition II; 10seq = data sets with 10 sequences; 2seq 3→2 = condition III; 3seq 3→2 = condition IV; 2seq 1→2 = condition I; 3seq 1→2 = condition II. Note that the y-axis in the three panels is not on the same scale



Rate of False-Positive Results in Branch-Site Models with Gene Conversion

Together with “site” models such M1a/M2a and M7/M8, CODEML also implements methods to look for positive selection on individual codons along specific branches of a

phylogenetic tree (Yang and Nielsen 2002; Zhang et al. 2005). These methods, also referred to as “branch-site” models, require subdivision of the tree into foreground and background lineages. A model allowing positive selection on foreground branches is compared by a likelihood ratio test with a second model that assumes no positive selection

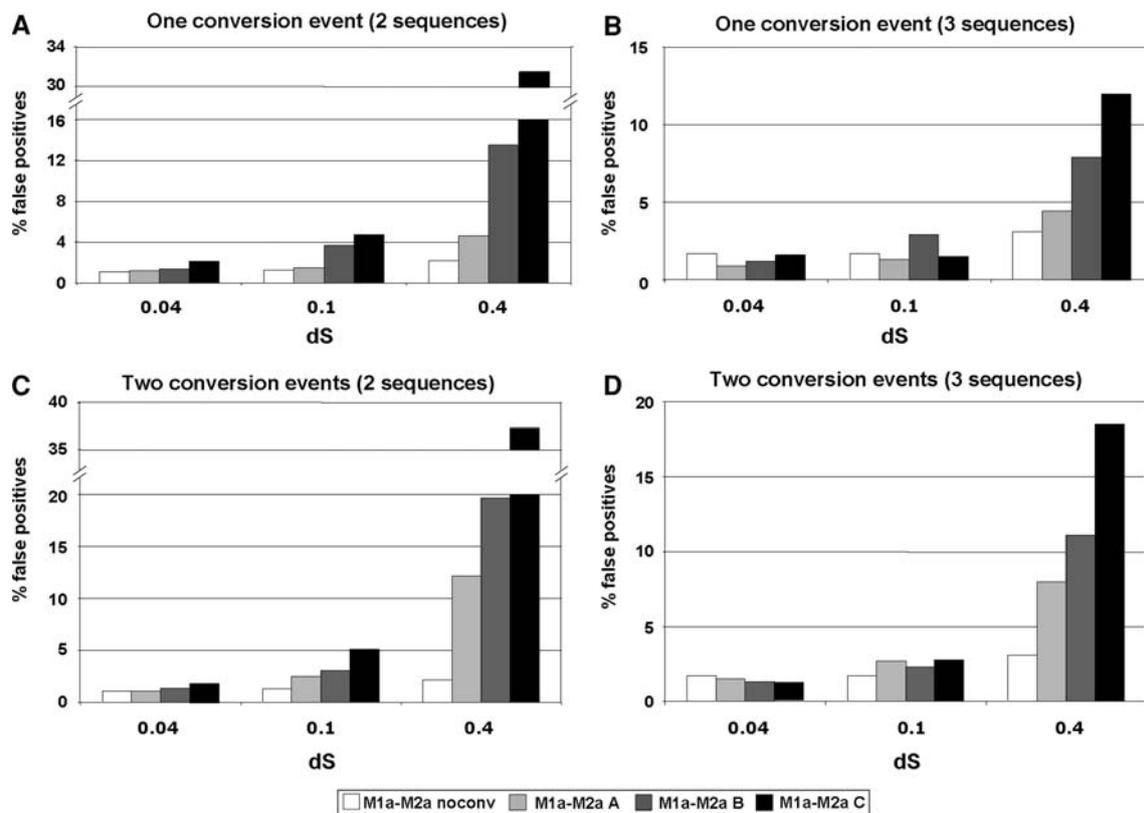


Fig. 4 Percentage of false-positive results in site models depending on the age and number of conversion events. Different times of conversion, experimental conditions, and models are compared (see text and Supplementary Table 1 for details). The pairwise d_S value

between ingroup sequences is shown on the x-axis. **a** = old conversion; **b** = recent conversion; **c** = new conversion; noconv = data sets with no conversion. Note that the y-axis in the two panels is not on the same scale

on these branches (see Methods). Given this a priori requirement, we could only examine replicates containing the outgroup sequence (conditions II and IV), testing one or the other ingroup sequence as the foreground branch in two independent analyses.

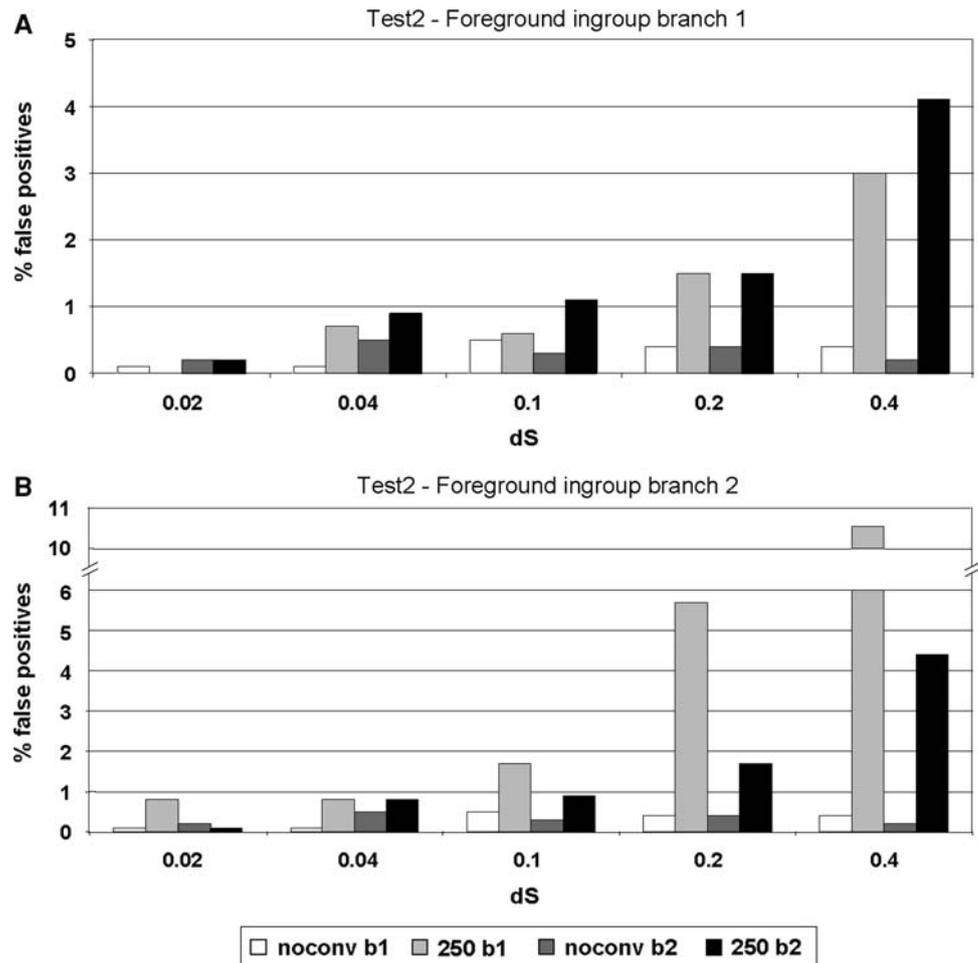
Type I error rates obtained from the LRT of the two branch-site models range from 0% to 13% (Fig. 5; Supplementary Fig. 5). As observed for the site models, these rates increase with the genetic distance and length of converted tracts. The largest difference in the proportion of false-positive results is seen between condition II (ingroup-to-ingroup conversion) and condition IV (outgroup-to-ingroup conversion). We used both ingroup sequences as foreground branches, with the acceptor sequence of the simulated gene conversion always denoted as ingroup branch 2 (“b2”). In condition II, there is little difference in the rate of false-positive results when using either branch 1 (“b1”) or branch 2 as the foreground lineage (Fig. 5a). This makes sense because the two lineages have been at least partly homogenized by gene conversion. In contrast, in condition V there is a significantly higher rate of false-positive results when using branch 1 as the foreground lineage (Fig. 5b). Because the outgroup branch and branch

2 are homogenized in condition IV, branch 1 (which is unaffected by conversion) will appear to be evolving at a much higher rate. This heterogeneity in branch lengths may cause the higher rate of false-positive results. Overall, changes in the tree topology seem to affect the specificity of branch-site methods in PAML when recombination occurs between background lineages. In general, branch-site models lead to lower type I error rates compared with site models, possibly because of their decreased sensitivity (Zhang et al. 2005).

Proportion of False-Positive Sites in Paralogs with Gene Conversion

CODEML site models that include parameters allowing positive selection (M2a and M8) also include two Bayesian estimations of codons evolving under positive selection using either the naïve empiric Bayes (NEB) or the Bayes empiric Bayes (BEB) algorithms. NEB does not account for sampling errors and is rather inaccurate, especially for small data sets with highly similar sequences (Yang 2007; Yang et al. 2005); therefore, we used only the results from the BEB method to infer the extent of type I error in

Fig. 5 Percentage of false-positive results in branch-site models versus the pairwise genetic distance of ingroup sequences for experimental condition II (a) and IV (b). “b1” and “b2” represent foreground ingroup branches 1 and 2, respectively. Results from replicates with simulated conversion tracts of 250 codons are shown. Noconv = data sets with no conversion. Note that the y-axis in the two panels is not on the same scale



detecting sites undergoing adaptive evolution, considering only sites with Bayesian confidence levels $\geq 95\%$ (Yang et al. 2005).

Overall, the BEB method produces few false-positive results, not exceeding 0.084% of all codons. However, given that the BEB method is conservative (Yang et al. 2005) and that only a few “positive” sites are identified even in the presence of adaptive evolution, we addressed which factors affect more significantly the distribution of such false-positive results. In this analysis, experimental condition I is not examined because converted regions between ingroup sequences are perfectly identical and show no false-positive results.

Experimental conditions are one of the most prominent factors shaping the BEB type I error, especially when converted and nonconverted regions are compared. In replicates including the outgroup sequence (conditions II and IV), BEB false-positive results in nonconverted regions increase with d_S , but only when model M8 is used, whereas the length of converted tracts seems to have only a minor effect (Supplementary Fig. 6A and 6C). Converted regions show a few BEB false-positive results regardless of

genetic distance, conversion tract length, and codon models. In experimental condition III (outgroup-to-ingroup conversion; only ingroup sequences analyzed), higher type I error rates are associated with converted regions, especially for longer converted tracts at $d_S = 0.02$, and using model M8 (Supplementary Fig. 6B).

Compared with the LRT results across whole sequences (i.e., M1a/M2a and M7/M8 comparisons), BEB predictions are based on single codon estimates of the numbers of synonymous and nonsynonymous substitutions; therefore, they are not significantly affected by changes in the phylogenetic tree topology introduced by recombination. In agreement with this, we noticed that the number of false-positive results, considering all BEB sites, is influenced in different ways than are LRTs by the length of converted tracts, codon model, and d_S values. LRT type I error rates increase with d_S and the length of conversion tracts for each experimental condition (Fig. 2; Supplementary Fig. 2). BEB false-positive results tend to be higher at extreme d_S values (0.02 and 0.4; see also Arbiza et al. 2006) with model M8 and when the outgroup sequence is included (Supplementary Fig. 6). Although this analysis

showed a highly variable number of false-positive results predicted by the BEB method, these numbers are always rather low, as noted by Yang et al. (2005).

Conclusion

Our results demonstrate that inferences of adaptive evolution in duplicate genes by the models implemented in CODEML can have moderately high type I error rates (up to approximately 33%) when conversion occurs between duplicated genes. Our results also suggest that using an outgroup sequence can increase specificity of the analysis when site methods are used, whereas this approach may produce the opposite effect with branch-site methods. In addition, larger gene trees negatively affect the accuracy of site models to predict adaptive selection in the presence of conversion, especially when donor and acceptor sequences are more distantly related and when conversion tracts are long. Overall, such results imply that erroneous between-paralogs inferences of positive selection due to gene conversion can be limited by using one outgroup sequence, even if this sequence is another paralog from the same genome. This approach is likely more effective than using large trees because large trees will also inevitably have more chances to harbor genes that have undergone conversion events. Importantly, the highest rates of false-positive results occur in exactly those conditions where gene conversion is easiest to detect (i.e., long conversion tracts and high d_S). This indicates that it will be relatively easy to exclude converted sequences from analyses of positive selection and therefore avoid an unnecessarily high proportion of false-positive results.

Acknowledgments We thank Mira Han for assistance and three reviewers for their insightful comments. This work was supported by a grant from the National Science Foundation to MWH (Grant No. DBI-0543586).

References

- Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* 18:1585–1592
- Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol Biol Evol* 19:950–958
- Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164:1229–1336
- Arbiza L, Dopazo J, Dopazo H (2006) Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput Biol* 2:288–300
- Benovoy D, Drouin G (2009) Ectopic gene conversions in the human genome. *Genomics* 93:27–32
- Bielawski JP, Yang Z (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. *J Struct Funct Genomics* 3:201–212
- Birtle Z, Goodstadt L, Ponting C (2005) Duplication and positive selection among hominin-specific PRAME genes. *BMC Genomics* 6:120
- Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP (2007) Gene conversion: Mechanisms, evolution and human disease. *Nat Rev Genet* 8:762–775
- Des Marais DL, Rausher MD (2008) Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* 454:762–765
- Drouin G, Prat F, Ell M, Clarke GD (1999) Detecting and characterizing gene conversions between multigene family members. *Mol Biol Evol* 16:1369–1390
- Gerton JL, De Risi J, Shroff R, Lichten M, Brown PO, Petes TD (2000) Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* 97:11383–11390
- Hahn MW, Demuth JP, Han SG (2007a) Accelerated rate of gene gain and loss in primates. *Genetics* 177:1941–1949
- Hahn MW, Han MV, Han SG (2007b) Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet* 3:1–12
- Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW (2009) Adaptive evolution of young gene duplicates in mammals. *Genome Res* 19:859–867
- Hill RE, Hastie ND (1987) Accelerated evolution in the reactive centre regions of serine protease inhibitors. *Nature* 326:96–99
- Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:167–170
- Li W-H (1997) *Molecular evolution*. Sinauer Associates, Sunderland, MA
- Mondragon-Palomino M, Gaut BS (2005) Gene conversion and the evolution of three leucine-rich repeat gene families in *Arabidopsis thaliana*. *Mol Biol Evol* 22:2444–2456
- Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* 11:715–724
- Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3:418–426
- Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936
- Nielsen KM, Kasper J, Choi M, Bedford T, Kristiansen K, Wirth DF, Volkman SK, Lozovsky ER, Hartl DL (2003) Gene conversion as a source of nucleotide diversity in *Plasmodium falciparum*. *Mol Biol Evol* 20:726–734
- Santoyo G, Romero D (2005) Gene conversion and concerted evolution in bacterial genomes. *FEMS Microbiol Rev* 29:169–183
- Sawyer S (1989) Statistical tests for detecting gene conversion. *Mol Biol Evol* 6:526–538
- Scheffler K, Martin DP, Seoighe C (2006) Robust inference of positive selection from recombining coding sequences. *Bioinformatics* 22:2493–2499
- Semple C, Wolfe KH (1999) Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. *J Mol Evol* 48:555–564
- Shriner D, Nickle DC, Jensen MA, Mullins JI (2003) Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet Res* 81:115–121
- Slightom JL, Blechl AE, Smithies O (1980) Human fetal G gamma- and A gamma-globin genes: Complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell* 21:627–638

- Wong WS, Yang Z, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168:1041–1051
- Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591
- Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* 19:908–917
- Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449
- Yang Z, Wong WS, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22:1107–1118
- Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* 22:2472–2479