ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect

Gene 385 (2006) 89 – 95

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Does lack of recombination enhance asymmetric evolution among duplicate genes? Insights from the *Drosophila melanogaster* genome

Yves Clément, Raquel Tavares, Gabriel A.B. Marais *

Laboratoire de Biométrie et Biologie Evolutive (UMR 5558); CNRS; Univ. Lyon 1,
Bat. Gregor Mendel, 16 rue Raphaël Dubois, 69622, Villeurbanne Cedex, France

Received 31 January 2006; received in revised form 9 May 2006; accepted 19 May 2006
Available online 17 August 2006

Abstract

Gene duplication has different outcomes: pseudogenization (death of one of the two copies), gene amplification (both copies remain the same), sub-functionalization (both copies are required to perform the ancestral function) and neo-functionalization (one copy acquires a new function). Asymmetric evolution (one copy evolves faster than the other) is usually seen as a signature of neo-functionalization. However, it has been proposed that sub-functionalization could also generate asymmetric evolution among duplicate genes when they experience different local recombination rates. Indeed, the low recombination copy is expected to evolve faster because of Hill–Robertson effects. Here we tested this idea with about 100 pairs of young duplicates from the *Drosophila melanogaster* genome. Looking only at young duplicates allowed us to compare recombination rates and evolutionary rates on a similar time-scale contrary to previous work. We found that dispersed pairs tend to evolve more asymmetrically than tandem ones. Among dispersed copies, the low recombination copy tends to be the fast-evolving one. We also tested the possibility that all this was explained by a confounding factor (expression level) but found no evidence for it. In conclusion, our results do support the idea that asymmetric evolution among duplicates is enhanced by restricted recombination. However, further work is needed to clearly distinguish between sub-functionalization and neo-functionalization for the asymmetrically-evolving duplicate pairs that we found.

Keywords: Gene duplication; Recombination; Hill–Robertson effects; Sub-functionalization; Neo-functionalization

1. Introduction

One of the surprises of genomics was to realise how frequent gene and genome duplication is in eukaryotes (Wolfe and Shields, 1997; Blanc and Wolfe, 2004; Jaillon et al., 2004). This has upraised Ohno's claim that duplication is a major evolutionary force because it brings new functions (Ohno, 1970). More recent work has specified the evolutionary fates for duplicate genes (Otto and Yong, 2002; Pyne et al., 2005): (i) one copy can become a pseudogene (pseudogenization) (ii) both duplicates can remain unchanged if there is a selective pressure for increased expression (gene amplification) (iii) both copies can accumulate deleterious mutations differentially so that both copies are needed to perform the ancestral function (sub-functionalization) (iv) one copy can diverge and acquire a new function (neo-functionalization). The acquisition of a new function will be achieved by repeated fixations of advantageous mutations (positive selection). The neo-functionalized copy will therefore evolve significantly faster than the other copy that will retain the ancestral function, evolving mainly by eliminating deleterious mutations (purifying selection). In the other cases in which both copies remain active (gene amplification, sub-functionalization), no differences in evolutionary rates between duplicates are expected. This is why asymmetric evolution among duplicates (i.e. when one copy evolves significantly faster than the other) is usually seen as a signature of neo-functionalization.

Zhang and Kishino have proposed that genomic context could affect whether a pair of duplicate genes will evolve asymmetrically or not (Zhang and Kishino, 2004a,b). The key factor in their model is the local recombination rate: when two duplicates are in two different genomic contexts, the copy in the low recombination context accumulates deleterious substitutions because of Hill–Robertson effects (degeneration) and this copy will evolve faster than the copy in the high recombination

* Corresponding author.
E-mail address: marais@biomserv.univ-lyon1.fr (G.A.B. Marais).

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.05.032

context, which does not suffer from degeneration. The most frequent fate of the low recombination copy will be pseudogenization after complete degeneration. Nevertheless, in some cases, sub-functionalization can prevent the low recombination copy to disappear. It is expected that most of the deleterious substitutions will be on the low recombination copy and only a few on the high recombination one. The sub-functionalized copies will therefore evolve asymmetrically. Neo-functionalization is also possible. When the low recombination copy will be accumulating deleterious substitutions, its sequence will change. Then, only a few advantageous mutations are needed to get a new function. The other copy will evolve under purifying selection and will retain the ancestral function (see Rastogi and Liberles, 2005 for how initial degeneration can help neo-functionalization). Zhang and Kishino's model may have important implications since recombination rates vary a lot along the eukaryotic genomes (Petes, 2001; McVean et al., 2004; Winckler et al., 2005). Moreover, sub-functionalization may be more frequent in the eukaryotic genomes than neo-functionalization (Lynch and Conery, 2003). In this case, Zhang and Kishino's model would predict that a substantial fraction of asymmetrically-evolving pairs, which at first sight have been considered cases of neo-functionalization, may in fact be sub-functionalized pairs with both copies differing in their genomic background.

To test their model, they used ~40 paralogous pairs from an ancient polyploidization event in yeast and found that the fast-evolving copies were preferentially located on regions of low recombination (Zhang and Kishino, 2004b). They also found some support for their model by comparing two gene clusters of amylase in *Drosophila*, which have different genomic locations: one is close to the centromere (supposedly low recombination rate) and the other at the middle of a chromosomal arm (supposedly high recombination rate) (Zhang and Kishino, 2004a). However, both studies have their problems. In the *Drosophila* study, the actual recombination rates were not available and they assumed recombination rate was low for one cluster and high for the other based on their rough chromosomal locations (peri-centromeric region, middle of a chromosomal arm). Moreover, this study was only about two gene clusters. The yeast study was conducted on a larger dataset but they compared present-day recombination rates in *Saccharomyces cerevisiae* with evolutionary rates over 100 million years. Recent work suggests that recombination is a fast-evolving trait (Ptak et al., 2004, 2005) and thus the comparison made in yeast does not seem very appropriate. Here our aim was to focus on recent duplicates in the *Drosophila melanogaster* genome in order to compare evolutionary rates with local recombination rates on a similar time-scale.

2. Materials and methods

2.1. Identification of recent duplicates in D. melanogaster

We looked for all the paralogous pairs specific to *Drosophila* using TreePattern (Dufayard et al., 2005) on the Hogenome database (version with ENSEMBL v24 data only, http://pbil.univ-lyon1.fr/). The tree motif that we looked for among all the gene family trees was groups with *Drosophila* sequences with no sequences of other eukaryotes present in Hogenome (basically all the completely sequenced eukaryotes available in ENSEMBL v24). We found 5890 paralogous pairs. Because *D. melanogaster* is the only *Drosophila* species in Hogenome, these 5890 pairs included ancient and recent duplicated genes of the *D. melanogaster* genome. In order to recover only the recent ones, we had to perform an additional step. We estimated the Ks values for each pair using JaDis and PAML (Yang, 1997; Goncalves et al., 1999). We included in the analysis only the pairs with Ks < 0.23, corresponding to the 95% percentile of the distribution of the Ks values for pairs of *D. melanogaster–D. simulans* orthologs (using a previously published dataset, Betancourt and Presgraves, 2002). This means that we chose the copies that duplicated around the *D. melanogaster–D. simulans* speciation. We did this to work on very recent duplicates in order to compare evolutionary rates with local recombination rates on a similar time-scale (see Introduction). We found 115 pairs.

2.2. Identification and annotation of the D. yakuba orthologs

To compare the evolutionary rates of our paralogous pairs, we needed an outgroup (see next section). We chose *D. yakuba* for this because *D. yakuba* was the only closely related species to *D. melanogaster* for which a complete genome was available when we conducted our study (the *D. simulans* genome was still under process). We downloaded the *D. yakuba* genome (version 1) from the WGSC ftp site (see http://genomeold.wustl.edu/projects/yakuba/). For each pair, we first made a tblastn of the CDS of the two copies from *D. melanogaster* on the complete genome of *D. yakuba*. We identified the best hits for both copies and recovered the boundaries of the *D. yakuba* genomic regions to which each *D. melanogaster* copy was most similar. In some cases the genomic region was the same for both copies, in other cases it was different. We used Fastacmd to extract the sequences of the genomic regions. We annotated these regions using the *D. melanogaster* copies with GeneWise2 (http://www.ebi.ac.uk/Wise2/). We extracted the coding sequence(s) of the *D. yakuba* ortholog(s) and made a blastp to check whether we recovered the *D. melanogaster* copies that we started with (reciprocal best hit). We found reliable orthologs for 104 pairs. Among these, 65 pairs had only one *D. yakuba* ortholog as expected (because we selected pairs that originated around the *D. melanogaster–D. simulans* speciation, see above) and 39 had more than one. These are likely to be cases of duplicated genes older than *D. melanogaster–D. yakuba* speciation but evolving under concerted evolution (so that their Ks is low). We included these pairs in the analysis because their Ks values indicate that they diverged after the last gene conversion event, which was around *D. melanogaster–D. simulans* speciation. Note that excluding these pairs does not change qualitatively the results although some tests are no longer significant due to the fact that the sample size is shortened by almost half.

2.3. Sequence analysis

For each set of sequences (two *D. melanogaster* copies + one or more *D. yakuba* orthologs), we made a multiple alignment

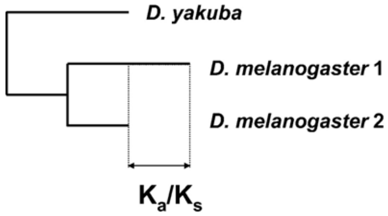

Fig. 1. Definition of ΔKa/Ks. ΔKa/Ks is the absolute value of the difference between the evolutionary rates of both copies (ΔKa/Ks=|(Ka/Ks)$_1$ − (Ka/Ks)$_2$|). Ka/Ks ratios have been obtained with RRTree, PAML and LTT (see Materials and methods).

using ClustalW on protein sequences and we back-translated it into DNA sequences (we did this to respect the CDS reading frame). We obtained the cytogenetic positions for all the *D. melanogaster* paralogs using Flybase (http://flybase.bio.indiana.edu/). We grouped the pairs in dispersed duplicates (duplicates of the same pair have different cytogenetic positions) and tandem duplicates (duplicates of the same pair have the same cytogenetic position). We had 32 dispersed pairs and 70 tandem pairs. Using the cytogenetic positions of paralogs, we got the local recombination rates for each copy using a compilation of estimates of recombination rates in *D. melanogaster* from Marais et al. (2003). The use of cytogenetic positions (and not physical positions or gene names) allowed us both to circumvent the problems of consistency between different releases of the *D. melanogaster* genome and to gather recombination data for 100 pairs (30 dispersed pairs, 70 tandem pairs). The data shown were obtained with HK02-w estimates, which capture fine-scale variations better. Some duplicates were at the boundary of two cytogenetic sections (Flybase indicated two sections for these genes and not just one). Their recombination rates were the average of those of the two cytogenetic sections indicated. This explains why some tandem pairs (those with one copy at the boundary of a section and the other copy at the other boundary of the same section) can have different recombination rates. We used RRTree (Robinson-Rechavi and Huchon, 2000) to compute the Ka, Ks and the Ka/Ks ratio for each copy and the outgroup (using *D. yakuba* ortholog(s) as outgroup) and to make relative-rates tests. In our analysis, we used Ka/Ks ratio in order to control for mutation rate. We also used PAML (Yang, 1997) and Like_tri_test (LTT, Conant and Wagner, 2003) to estimate Ka/Ks with the maximum likelihood approach. We ran PAML (codeml) on each triplet of sequences (paralogous pair + one outgroup sequence) using a non-constrained (model = 1, all branches have their own Ka/Ks ratios) and constrained model (model = 2, both copies have the same Ka/Ks ratio). Similarly, we ran LTT on each triplet using a non-constrained (-m:OFF, all branches have their own Ka/Ks ratios) and a constrained model (-m:kaks, both copies have the same Ka/Ks ratio). Gaps were excluded in both PAML and LTT analyses. For each triplet, we compared the constrained and non-constrained models with a likelihood ratio test (LRT) in both PAML and LTT analyses. We obtained expression patterns for *D. melanogaster* duplicated genes using EST data (as in Duret and Mouchiroud, 1999). We classified EST libraries in Adult (Head, Testis, Ovary), Larva, Embryo and Mixed using annotations of these libraries. These libraries did not have the same number of EST sequences. We therefore normalized them *in silico* (so that we could compare them). For each gene, we computed the maximum value among all libraries except Mixed (exp$_{max}$), the average value among all libraries (Av-exp) and the sum of libraries except Mixed (tissue-nb).

3. Results

The idea of our test was to compare evolutionary rates and recombination rates for young duplicates. We focused on 100 pairs of young duplicates of the *D. melanogaster* genome (see Materials and methods), of which 30 are dispersed duplicates and 70 are tandem duplicates. For each pair, we computed the Ka/Ks ratio between each copy and the *D. yakuba* ortholog(s) with different methods RRTree, PAML, LTT (see Materials and methods). We obtained the figures using RRTree estimates but we also mention what we found with other methods in the text. Tables show results obtained with the three methods. Following previous work (e.g. Conant and Wagner, 2003), we defined ΔKa/Ks as the absolute value of the difference between the Ka/Ks of both copies, which tells us whether one copy has been evolving faster than the other (asymmetric evolution) or not (symmetric evolution) (see Fig. 1). Note that other definitions of asymmetry exist (e.g. G+C asymmetry, see Rodin and Parkhomchuk, 2004; Jabbari et al., 2003), but hereafter we are only referring to rate asymmetry.

First, we compared ΔKa/Ks for tandem and dispersed duplicates (see Fig. 2). A significant source of ΔKa/Ks variation can be attributed to gene function varying from pair to pair. To make the ΔKa/Ks as comparable as possible between pairs having different functions, we standardized ΔKa/Ks by dividing it by the sum of the Ka/Ks ratios of both copies (ΔKa/Ks* = ΔKa/Ks / [(Ka/Ks)$_1$ + (Ka/Ks)$_2$], as in Conant and Wagner, 2003). We found that ΔKa/Ks* was almost 30% higher in dispersed duplicates than in tandem duplicates although a Mann–Whitney test was statistically non-significant (ΔKa/Ks* estimates from PAML or LTT gave similar results). This tend to suggest that copies that experienced different genomic contexts tend to evolve more asymmetrically than others, which is in agreement with Zhang and Kishino's model. The question then is: Are differences in recombination rates really explaining this as Zhang and Kishino's model would predict?

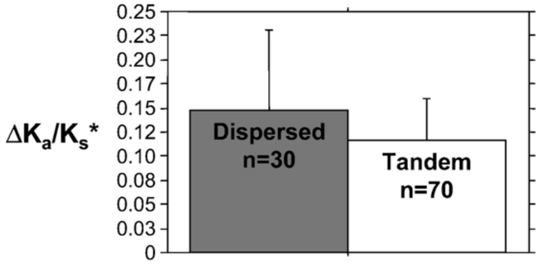

Fig. 2. Level of asymmetric evolution for dispersed and tandem duplicates. Asymmetry is measured by ΔKa/Ks*, which is a standardized version of ΔKa/Ks (see Fig. 1 and text). Dispersed and tandem duplicates are defined in Materials and methods. Error bars indicate 95% confidence interval.

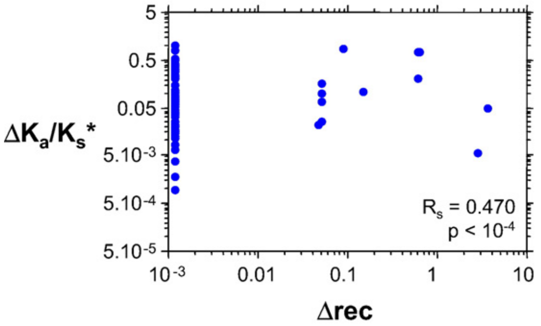

Fig. 3. Relationship between the level of asymmetric evolution (ΔKa/Ks*) and differences of recombination rates (Δrec) for duplicates. Asymmetry is measured by ΔKa/Ks*, which is a standardized version of ΔKa/Ks (see Fig. 1 and text). See Materials and methods to know how recombination rates are estimated. Both axes are in log scale.

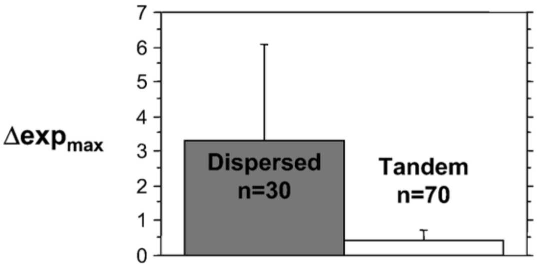

Fig. 4. Differences in expression level for dispersed and tandem duplicates. Expression level is measured by $\exp_{\max}$ (see Materials and methods). Av-exp gave very similar results. Values of tissue-nb are identical between copies of all the pairs that we looked at so that Δtissue-nb is always 0. Dispersed and tandem duplicates are defined in Materials and methods. Error bars indicate 95% confidence interval.

To address this question, we compared ΔKa/Ks* to Δrec, which we defined as the absolute value of the difference between the recombination rates of the copies for each pair (see Materials and methods for how recombination rates are estimated). We found that both parameters are positively correlated and this correlation is significant ($R_s = 0.470$ with $p < 10^{-4}$, see Fig. 3). Similar results were found with ΔKa/Ks* from PAML ($R_s = 0.333$ with $p = 0.0009$), LTT ($R_s = 0.280$ with $p = 0.0053$) and also using another measure of asymmetry — the standardized difference between Ka among duplicates (ΔKa*) – as in Conant and Wagner (2003) – ($R_s = 0.444$ with $p < 10^{-4}$). Note that the correlation between ΔKa/Ks* and Δrec is rather weak and the plot between these two variables is quite scattered probably because (i) the differential accumulation of substitutions among paralogs is a stochastic process, (ii) the rate of conversion (decreasing divergence between copies) may vary among tandem pairs (explaining why pairs with Δrec = 0 can have quite different ΔKa/Ks*) and (iii) recombination rates are roughly estimated (see for example Marais et al., 2003 for a discussion on this issue). Anyway, the above result suggests that differences in recombination among duplicated genes increase the chance that they evolve in an asymmetric manner, in agreement with Zhang and Kishino's model.

However, Zhang and Kishino's predictions are more specific. In their model, the copy experiencing the lowest recombination rate should be the one evolving fast, because of degeneration and possibly positive selection (if neo-functionalization is occurring, see Introduction). To test whether this was true in our dataset, we counted the number of pairs with the low recombination copy evolving the fastest (i.e. high Ka/Ks) (in agreement with Zhang and Kishino's model), the number of pairs with the high recombination copy evolving the fastest (not in agreement with Zhang and Kishino's model) and the number of pairs for which we cannot conclude (no differences in evolutionary rates or/and no differences in recombination rates) (see Table 1). We focused on dispersed genes because they are the only ones that can give relevant information about the test of Zhang and Kishino's model. Indeed, for tandem pairs, we only had two pairs with differences in recombination rates (see Table 1, see also Materials and methods). We found that for 68 to 83% (depending on the method) of the dispersed pairs the fastest-evolving copy was actually the low recombination copy, in agreement with Zhang and Kishino's model. Table 1 also shows that there are more pairs with significant asymmetry (detected by various methods) that support Zhang and Kishino's model than the contrary.

4. Discussion

In their model, Zhang and Kishino suggested that duplicates will evolve at different rates when they are in different genomic contexts mainly through differences in recombination rates. Our results tend to support their prediction. However, genomic regions can differ not only for recombination but also for other aspects. In particular, they can differ in global transcriptional activity. It is now known that there are chromosomal domains with low transcriptional activity and others with high

Table 1
Number of duplicate pairs in agreement or not with Zhang and Kishino's (ZK) model

	?			ZK+			ZK−		
	RRTree	PAML	LTT	RRTree	PAML	LTT	RRTree	PAML	LTT
Dispersed duplicates	15 (3)	16 (2)	18 (2)	11 (1)	11 (5)	10 (4)	3 (0)	3 (0)	2 (1)
Tandem duplicates	68 (6)	68 (17)	68 (17)	1 (1)	1 (1)	1 (1)	1 (1)	1 (0)	1 (0)

?: no conclusion can be reached either because both copies evolve at the same rate (ΔKa/Ks = 0) or because they have the same recombination rate (Δrec = 0).
ZK+: the low recombination copy is fast-evolving (in agreement with ZK model).
ZK−: the low recombination copy is slow-evolving (not in agreement with ZK model).
() : number of pairs with significant asymmetric evolution detected with RRTree (for Ka only), PAML or LTT (see Materials and methods).
Fisher exact test $p < 10^{-3}$.

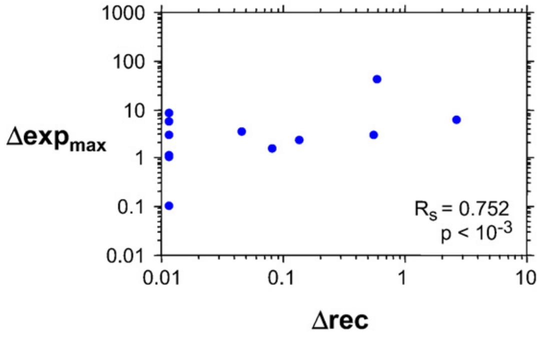

Fig. 5. Relationship between differences in expression level (Δexp$_{max}$) and differences of recombination rates (Δrec) for duplicates. See Materials and methods to know how exp$_{max}$ and recombination rates are estimated. Both axes are in log scale.

transcriptional activity in many organisms including *Drosophila* (Boutanaev et al., 2002; Cremer and Cremer, 2001). Gene expression is also known to be a major (if not the most important) determinant of evolutionary rate (see Rocha, 2006 for review). Highly and broadly expressed genes evolve slowly whereas lowly and tissue-specific genes evolve fast. If two duplicated genes are located in a transcriptionally active domain and the other in the transcriptionally inactive one, we expect them to evolve at different rates.

The problem here is that the transcriptionally inactive domains and the low recombination regions may be the same. Thus, gene expression may be a confounding factor in our study. To address this question, we estimated the level and breadth of gene expression for both copies of our 100 pairs (see Materials and methods). We did not find any differences in expression breadth among our 100 young duplicate pairs. We only found differences in expression level. We defined Δexp$_{max}$ as the absolute value of the difference in expression level (measured with exp$_{max}$, see Materials and methods) between the two copies of each pair. We found that Δexp$_{max}$ was higher in dispersed duplicates than tandem duplicates (see Fig. 4, significant Mann–Whitney test $p<10^{-3}$). We also found that Δexp$_{max}$ was positively correlated to Δrec (see Fig. 5). All this suggests that gene expression can potentially explain our results. We tried to remove the effect of gene expression on asymmetric evolution by computing the residuals of the correlation between ΔKa/Ks* and Δexp$_{max}$ and to correlate them to Δrec. We still found an effect of recombination (the correlations were marginally or fully significant depending on the ΔKa/Ks* estimates: RRTree: $R_s=0.403$ with $p<10^{-4}$, PAML: $R_s=0.306$ with $p=0.0023$, LTT: $R_s=0.269$ with $p=0.0075$, the result was the same with ΔKa*: $R_s=0.323$ with $p=0.0014$). Furthermore, we examined pairs with the lowly expressed copy evolving the fastest (i.e. high Ka/Ks) (in agreement with the expression domains hypothesis), pairs with the highly expressed copy evolving the fastest (not in agreement with the expression domains hypothesis) and pairs for which we cannot conclude (no differences in evolutionary rates or/and no differences in expression level) (see Table 2). We found more dispersed pairs not in agreement with the expression domains hypothesis than dispersed pairs in agreement with it. We did not have thus evidence that our previous results could be explained by gene expression alone.

Another possible caveat of our study is the sample size, which is very small indeed. We had only 30 pairs with dispersed copies and only 15 of them had different recombination rates. However, we must stress that we observed differences between dispersed and tandem pairs (see Fig. 2) and that we found a statistically significant excess of dispersed pairs in agreement with Zhang and Kishino's expectations (see Table 1). Moreover, we looked over all the *D. melanogaster* genome for young duplicates and the 100 pairs are all that we found. There is not much opportunity for increasing our sample size.

Zhang and Kishino's model relies on the idea that copies located on regions of low recombination will degenerate (see Introduction). Previous work suggests that such degeneration should only occur at very low recombination rates (i.e.>1 cM/Mb) in *D. melanogaster* (Marais and Piganeau, 2002). Interestingly, in many of the pairs in agreement with Zhang and Kishino's model (see Table 1) the low recombination copy has a recombination rate close to 1 (9/11). However, in some of the pairs not in agreement with Zhang and Kishino's model (see Table 1) this is also the case (2/3).

Some of the pairs not in agreement with Zhang and Kishino's model (see Table 1) have high Ka/Ks values and at least one of them is involved in reproductive function, which suggests that those gene families have been evolving under positive selection, probably even before duplication. For this kind of genes, the Zhang and Kishino predictions may not apply. The copy in the low recombination environment should be less able to fix advantageous mutations and may evolve somewhat slower than the copy in the high recombination context, contrary to the Zhang and Kishino prediction (see Introduction). This may explain why not all of our pairs are in agreement with the Zhang

Table 2
Number of duplicate pairs in agreement or not with the expression domains (ED) hypothesis

	?			ED+			ED−		
	RRTree	PAML	LTT	RRTree	PAML	LTT	RRTree	PAML	LTT
Dispersed duplicates	10 (0)	13 (1)	16 (1)	7 (3)	4 (2)	7 (3)	12 (1)	5 (4)	9 (3)
Tandem duplicates	60 (4)	60 (16)	61 (16)	6 (2)	15 (2)	5 (2)	4 (2)	5 (0)	4 (0)

?: no conclusion can be reached either because both copies evolve at the same rate (ΔKa/Ks=0) or because they have the same expression level (Δexp$_{max}$=0).
ED+: the lowly expressed copy is fast-evolving (in agreement with ED hypothesis).
ED−: the lowly expressed copy is slow-evolving (not in agreement with ED hypothesis).
() : number of pairs with significant asymmetric evolution detected with RRTree (for Ka only), PAML or LTT (see Materials and methods).
Fisher exact test $p<10^{-3}$.

and Kishino model although it is difficult to draw strong conclusions with so few pairs (see Table 1).

5. Concluding remarks

In conclusion, we can say that our test has the advantage of comparing evolutionary rates and recombination rates on a similar time-scale. However, it also has some disadvantages: we only had 100 pairs for our analysis, many of which are tandem pairs (70%), and the differences between young duplicates are small and the relative-rates test is often non-significant. Our results show that the genomic context does have an influence on asymmetric evolution of paralogs and supports Zhang and Kishino's model in that respect. However, it is difficult to say whether recombination is the main factor at work. Our data tend to support this view but we have a very small sample size. Our work does not suggest that gene expression contributes to the asymmetric evolution of paralogs that we observed but again this may be due to small sample size. Indeed, expression levels (and even recombination rates) are roughly measured by our methods and larger sample size may be needed to detect clear patterns. Alternatively, expression changes may be the consequences and not the causes of divergence between duplicates as it has been suggested previously (see Li et al., 2005 for review). We have mentioned the possibility that the genomic background of a duplicate gene affects its expression (because of the existence of transcriptionally active domains in the genome) and thus its evolutionary rate, because both parameters are strongly correlated (see Discussion). However, recent work has suggested that evolutionary rates at coding and non-coding regions of a same gene are coupled (Castillo-Davis et al., 2004; Marais et al., 2005). Differences in genomic contexts (e.g. recombination rates) among duplicates could generate asymmetric evolution at regulatory elements located on non-coding DNA, and this would make gene expression of duplicates diverge. Finally, we think our approach may yield to stronger conclusions with larger genomes (possibly with more young duplicates) than that of the *Drosophila* species. Moreover, our approach did not allow us to distinguish between sub-functionalization and neo-functionalization when we found asymmetric evolution among duplicates (this was also a problem in Zhang and Kishino's previous papers). The longer branch of the low recombination copy (see Fig. 1) can indeed only reflect degeneration (sub-functionalization) or degeneration plus positive selection (neo-functionalization, see Introduction). This is certainly the next thing to investigate if we want to know whether a significant fraction of asymmetric-evolving pairs found in eukaryotic genomes are actually sub-functionalized and not neo-functionalized as usually admitted. Combining divergence data and polymorphism data could help in doing this by directly checking for positive selection on fast-evolving copies.

Acknowledgements

We thank Dominique Mouchiroud for support in this project and advice with the statistics, Laurent Duret for helpful suggestions about gene annotation, and Gavin Conant for help with the Like_tri_test. We thank WGSC for making the *D. yakuba* genome available. This project was financially supported by the C.N.R.S.

References

Betancourt, A.J., Presgraves, D.C., 2002. Linkage limits the power of natural selection in *Drosophila*. Proc. Natl. Acad. Sci. U. S. A. 99, 13616–13620.

Blanc, G., Wolfe, K.H., 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16, 1667–1678.

Boutanaev, A.M., Kalmykova, A.I., Shevelyov, Y.Y., Nurminsky, D.I., 2002. Large clusters of co-expressed genes in the *Drosophila* genome. Nature 420, 666–669.

Castillo-Davis, C.I., Hartl, D.L., Achaz, G., 2004. cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res. 14, 1530–1536.

Conant, G.C., Wagner, A., 2003. Asymmetric sequence divergence of duplicate genes. Genome Res. 13, 2052–2058.

Cremer, T., Cremer, C., 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat. Rev., Genet. 2, 292–301.

Dufayard, J.F., Duret, L., Penel, S., Gouy, M., Rechenmann, F., Perriere, G., 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics 21, 2596–2603.

Duret, L., Mouchiroud, D., 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc. Natl. Acad. Sci. U. S. A. 96, 4482–4487.

Goncalves, I., Robinson, M., Perriere, G., Mouchiroud, D., 1999. JaDis: computing distances between nucleic acid sequences. Bioinformatics 15, 424–425.

Jabbari, K., Rayko, E., Bernardi, G., 2003. The major shifts of human duplicated genes. Gene 317, 203–208.

Jaillon, O., et al., 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431, 946–957.

Li, W.H., Yang, J., Gu, X., 2005. Expression divergence between duplicate genes. Trends Genet. 21, 602–607.

Lynch, M., Conery, J.S., 2003. The origins of genome complexity. Science 302, 1401–1404.

Marais, G., Piganeau, G., 2002. Hill–Robertson interference is a minor determinant of variations in codon bias across *Drosophila melanogaster* and *Caenorhabditis elegans* genomes. Mol. Biol. Evol. 19, 1399–1406.

Marais, G., Mouchiroud, D., Duret, L., 2003. Neutral effect of recombination on base composition in *Drosophila*. Genet. Res. 81, 79–87.

Marais, G., Nouvellet, P., Keightley, P.D., Charlesworth, B., 2005. Intron size and exon evolution in *Drosophila*. Genetics 170, 481–485.

McVean, G.A., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R., Donnelly, P., 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304, 581–584.

Ohno, S., 1970. Evolution by Gene Duplication. Springer-Verlag, New York.

Otto, S.P., Yong, P., 2002. The evolution of gene duplicates. Adv. Genet. 46, 451–483.

Petes, T.D., 2001. Meiotic recombination hot spots and cold spots. Nat. Rev., Genet. 2, 360–369.

Ptak, S.E., Roeder, A.D., Stephens, M., Gilad, Y., Paabo, S., Przeworski, M., 2004. Absence of the TAP2 human recombination hotspot in chimpanzees. PLoS Biol. 2, e155.

Ptak, S.E., Hinds, D.A., Koehler, K., Nickel, B., Patil, N., Ballinger, D.G., Przeworski, M., Frazer, K.A., Paabo, S., 2005. Fine-scale recombination patterns differ between chimpanzees and humans. Nat. Genet. 37, 429–434.

Pyne, S., Skiena, S., Futcher, B., 2005. Copy correction and concerted evolution in the conservation of yeast genes. Genetics 170, 1501–1513.

Rastogi, S., Liberles, D.A., 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol. Biol. 5, 28.

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

Rocha, E.P., 2006. The universals of protein evolution. Trends Genet. 22 (8), 412–416 (Epub 2006 Jun 30).

Rodin, S.N., Parkhomchuk, D.V., 2004. Position-associated GC asymmetry of gene duplicates. J. Mol. Evol. 59, 372–384.

Winckler, W., et al., 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308, 107–111.

Wolfe, K.H., Shields, D.C., 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387, 708–713.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.

Zhang, Z., Kishino, H., 2004a. Genomic background drives the divergence of duplicated amylase genes at synonymous sites in *Drosophila*. Mol. Biol. Evol. 21, 222–227.

Zhang, Z., Kishino, H., 2004b. Genomic background predicts the fate of duplicated genes: evidence from the yeast genome. Genetics 166, 1995–1999.