Ectopic Gene Conversions Increase the G + C Content of Duplicated Yeast and *Arabidopsis* Genes

David Benovoy, Robert T. Morris, Antoine Morin, and Guy Drouin

Département de biologie, Université d'Ottawa, Ottawa, Ontario, Canada

Allelic recombination has previously been shown to increase the GC-content of the sequences of a wide variety of eukaryotic species. Ectopic recombination between clustered tandemly repeated genes has also been shown to increase their GC-content. Here we show that gene conversions between the dispersed genes found in the duplicated regions of the yeast and *Arabidopsis* genomes also increase their GC-content when these genes are more than 88% similar.

Introduction

The nucleotide content of genes and genomes changes during evolution. Processes that increase AT-content are well known. They include the deamination of 5-methylcytosine into thymine and oxidative damage to cytosine or guanine (Lindahl 1993; Birdsell 2002). Processes that increase GC-content are not so well known (Sueoka 2002). However, many studies have shown that DNA repair mechanisms are biased toward GC nucleotides (Brown and Jiricny 1988, 1989; Bill et al. 1998). Frequent DNA repair, such as the DNA repair associated with recombination, is therefore expected to increase GC-content during evolution. These predictions have been confirmed by several studies that showed that allelic recombination does increase the GC-content of yeast, *Caenorhabditis elegans*, *Drosophila*, *Xenopus*, bird, and mammalian DNA sequences (Gerton et al. 2000; Fullerton, Bernardo Carvalho, and Clark 2001; Galtier et al. 2001; Marais, Mouchiroud, and Duret 2001; Takano-Shimizu 2001; Birdsell 2002; Duret 2002; Kong et al. 2002; Galtier 2003; Marais 2003; Jensen-Seaman et al. 2004; Meunier and Duret 2004). One would also expect that ectopic gene conversions, i.e., gene conversions between duplicated genes located at different loci, would also increase the GC-content of the genes involved. In fact, some studies have shown that ectopic recombination between clustered tandemly repeated genes also increase their GC-content (Hickey, Wang, and Magoulas 1994; Galtier 2003; Kudla, Helwak, and Lipinski 2004; Noonan et al. 2004). Here we use ohnologs, i.e., duplicated genes produced by genome duplications (Wolfe 2001), to show that gene conversions between dispersed duplicated genes also increase their GC-content.

The ohnologs found in the yeast (*Saccharomyces cerevisiae*) and *Arabidopsis thaliana* genomes are particularly well suited to test the effect of ectopic gene conversions on the GC-content of genes because they consist of pairs of duplicated genes which were all created at the same time. The yeast genome duplication occurred some 150 MYA (Langkjaer et al. 2003). As a result of this duplication, 54 duplicated gene blocks can still be found in the yeast genome and all but two of these duplicated gene blocks are found on different chromosomes (Wolfe and Shields 1997). The *A. thaliana* genome contains ohnologs derived from at least two complete genome duplications, the last of which occurred some 24–40 MYA (Blanc, Hokamp, and Wolfe 2003). Here, we only analyzed the *Arabidopsis* ohnologs from the most recent duplication in order to use genes that were duplicated at the same time. These recently duplicated genes represent 85% of the ohnologs found in the *Arabidopsis* genome, and most of them are located on different chromosomes (Blanc, Hokamp, and Wolfe 2003).

Materials and Methods

The sequences of the 750 yeast ohnologs (375 pairs of genes) and of the 4,994 *Arabidopsis* recent ohnologs (2,497 pairs of genes) were downloaded from the National Center for Biotechnology Information Web site (http://www.ncbi.nlm.nih.gov/) using the lists of duplicated genes found by the studies of Wolfe and Shields (1997) and Blanc, Hokamp, and Wolfe (2003) (http://wolfe.gen.tcd.ie/). Each pair of duplicated genes was aligned using ClustalW (Thompson, Higgins, and Gibson 1994). The average GC-content (%) at the third position of codons and the average uncorrected sequence similarity of each aligned gene pair were then computed using an in-house PERL script.

The yeast recombination data of the study of Gerton et al. (2000) were obtained from http://derisilab.ucsf.edu/hotspots/. The median recombination rate was computed from the seven replicates of red:green ratios for each of the 750 yeast ohnologs. Our yeast recombination values are therefore median recombination rates. Because of the low density of genetic markers, the recombination map of *Arabidopsis* still does not allow to measure local recombination rates (Wright, Agrawal, and Bureau 2003; Marais, Charlesworth, and Wright 2004). We therefore did not attempt to measure the effect of recombination on the GC3-content of *Arabidopsis* ohnologs.

All statistical analyses (Kolmogorov-Smirnov tests of normality, linear and nonlinear regression analyses, etc.) were performed using S-plus v6.2 (Insightful Corporation, Seattle, Wash.) and Excel (Microsoft Corporation, Redmond, Wash.).

Results

Figure 1 clearly shows that the genes found in the duplicated regions of the yeast genome are divided into two groups. The first group is composed of sequences less than 87.7% similar, and there is no correlation between sequence similarity and GC-content at third positions of

Key words: ectopic, gene conversion, GC-content, recombination, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*.

E-mail: gdrouin@science.uottawa.ca.

Mol. Biol. Evol. 22(9):1865–1868. 2005
doi:10.1093/molbev/msi176
Advance Access publication May 25, 2005

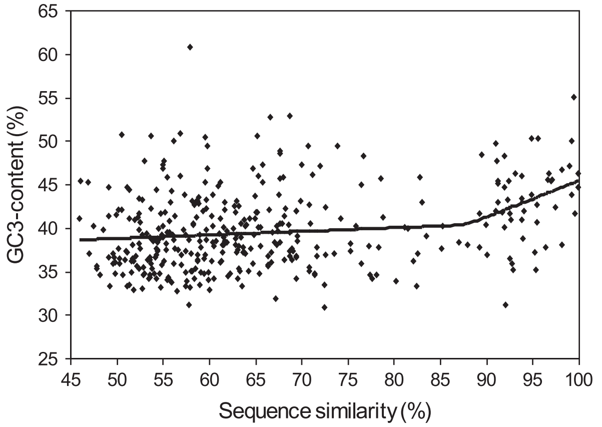

FIG. 1.—Relationship between the average GC-content of third codon positions (GC3) and the average sequence similarity of the 375 pairs of ohnologs in the yeast genome.

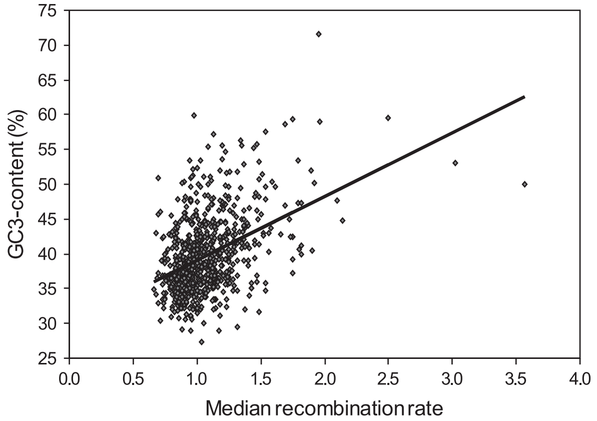

FIG. 2.—Relationship between the GC-content of third codon positions (GC3) and the median recombination rate of the 750 ohnologs found in the yeast genome.

codons ($r^2 = 1 \times 10^{-6}$, $P = 0.98$). The second group is composed of sequences more than 87.7% similar, and there is a significant correlation between sequence similarity and GC-content at third positions of codons ($r^2 = 0.085$, $P = 0.036$). This division into two groups (i.e., with two regressions) is significantly better than a less complex model with a single regression ($F = 2.93$, $P = 0$), and the inflection point is at 87.7% similarity (95% confidence interval [CI] of 79.1%–94.3%). The mean GC3-content of sequences less than 87.7% similar is 39.3% and is significantly lower (Wilcoxon rank-sum test, $Z = 5.42$, $P = 0$) than that of sequences more than 87.7% similar (with a mean GC3-content of 43.0%). In contrast, the mean median recombination rate (and standard error) of sequences less than 87.7% similar (1.07 ± 0.01) is not significantly different ($Z = 0.07$, $P = 0.95$) from that of sequences more than 87.7% similar (1.09 ± 0.02).

Figure 2 does not show a clear division of yeast ohnologs into two groups based on their median recombination rates. However, it shows that lower recombination rates are more frequent than higher recombination rates and that recombination rates are positively correlated with GC3-content ($r^2 = 0.16$, $P = 0$). We also performed a multiple nonlinear regression analysis of the effect of similarity and recombination on GC3-content. We found that recombination rate has no effect on GC3-content. In fact, for recombination rate, both the slopes before and after the inflection point are not significantly different from zero ($P = 0.24$ and 0.16, respectively).

Figure 3 shows that the ohnologs found in the *Arabidopsis* genome are also divided into two groups. The first group is composed of sequences less than 86.6% similar, and there is no correlation between sequence similarity and GC-content at third positions of codons ($r^2 = 0.001$, $P = 0.08$). The second group is composed of sequences more than 86.6% similar, and there is a significant correlation between sequence similarity and GC-content at third positions of codons ($r^2 = 0.10$, $P = 2 \times 10^{-5}$). This division into two groups (i.e., with two regressions) is significantly better than a less complex model with a single regression ($F = 20.70$, $P = 0$), and the inflection point is at 86.6% similarity (95% CI of 85.6%–87.6%). The mean GC3-content of sequences less than 86.6% similar is 43.49% and is significantly lower (Wilcoxon rank-sum test, $Z = 3.40$, $P = 0.0007$) than that of sequences more than 86.6% similar (45.65%).

Discussion

In both yeast and *Arabidopsis*, the GC-content of the third codon positions of sequences less than 88% similar shows no correlation with sequence similarity, whereas that of sequences more than 88% similar shows a significant correlation with sequence similarity (figs. 1 and 3). Because this division into two groups is not due to differences in recombination (fig. 2), our results suggest that ectopic gene conversions increase the CG-content of dispersed duplicated yeast and *Arabidopsis* genes. Some of the genes which were duplicated 150 MYA in the yeast genome and 24–40 MYA in the *Arabidopsis* genome have retained a high level of similarity through gene conversions, and these conversions have also increased their GC-content.

Both experimental and sequence analyses studies have shown that gene conversion is more frequent between more similar sequences (Borts and Haber 1987; Modrich and

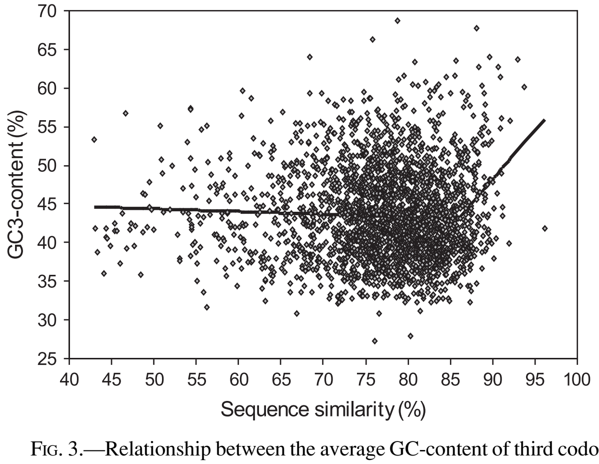

FIG. 3.—Relationship between the average GC-content of third codon positions (GC3) and the average sequence similarity of the 2,497 pairs of recent ohnologs in the *Arabidopsis* genome.

Lahue 1996; Drouin 2002), and the study of Gao and Innan (2004) has shown that many yeast ohnologs have been subject to numerous gene conversions. One therefore expects similar sequences to become even more similar due to gene conversions, whereas less similar sequences will gradually diverge from one another and thus escape gene conversions. In fact, the clear division of the yeast ohnologs into two groups (below and above 87.7% similarity; fig. 1) likely represents genes that escaped and genes still undergoing gene conversions, respectively. Furthermore, this division into two groups is not due to different recombination rates of the genes found in the two groups because the average recombination rate is the same in both groups. The absence of such visually obvious groups in *Arabidopsis* (fig. 3) could be the result of the lower level of recombination in *Arabidopsis* and the fact that the ohnologs of this species diverged more recently. In fact, the shape of the distribution observed in *Arabidopsis* and the excess of data points between 70% and 85% similarity are what would be expected under the hypothesis of recently duplicated genes undergoing a continuous rate of escape from gene conversion (fig. 3). The fact that a similarity of at least 88% is necessary to observe an effect in both species suggests that the mechanisms responsible for ectopic gene conversions are similar in fungi and plants.

Another hypothesis which could explain our results would be that they reflect differences in codon usage. Under this hypothesis, more conserved genes would use more codons containing guanine or cytosine in third codon positions. However, this hypothesis would not explain why the correlation between GC-content and similarity is limited to gene having more than 88% similarity, why this correlation is limited to gene having more than 88% similarity in two very different species, and why this correlation is of the same magnitude in both species (r^2 of 0.085 and 0.10 for yeast and *Arabidopsis*, respectively). Because optimal codons are known to be species specific and there is strong selection for optimal codons in yeast but not in *Arabidopsis*, one would not expect selection for optimal codons to lead to similar increases in GC3 in these two very different species (Sharp et al. 1988; Duret and Mouchiroud 1999). In contrast, the GC-biased gene conversion hypothesis explains both the fact that conversions are limited to very similar sequences and the fact that GC-content increases with similarity.

The fact that the correlation between GC-content and similarity is relatively small is consistent with the previous yeast study of Gerton et al. (2000), where frequent allelic recombination only resulted in GC-content increases of a few percent. Similarly, the correlation between the GC-content of the third codon positions of 6,143 yeast open reading frames and their mean allelic recombination rate is also relatively low ($\rho^2 = 0.156$) but is highly significant ($P = 3.7 \times 10^{-211}$, Birdsell 2002). Because gene conversions between unlinked repeated sequences are less frequent than between alleles (Petes and Hill 1988; Haber et al. 1991; Goldman and Lichten 1996), one expects ectopic gene conversions to have a smaller effect than allelic gene conversions. Interestingly, the correlation we observed between the recombination rate and GC3-content of yeast ohnologs ($r^2 = 0.16$; fig. 2) is very similar to that of Birdsell (2002). This suggests that yeast ohnologs are a representative sample of yeast genes.

The effect of biased gene conversion on GC-content requires that the gene being converted and its template be different (Galtier et al. 2001). Because highly inbreed species would be homozygous for most of their genes, one would not expect biased gene conversion to affect the GC-content of their genes. This prediction is supported by the absence of correlation between the rate of crossing over and the GC-content of the genes found in *Arabidopsis*, a species with a selfing rate of about 99% (Marais, Charlesworth, and Wright 2004). The presence of a positive correlation between recombination rate and GC-content in yeast (see above), another species with a selfing rate of about 99% (Johnson et al. 2004), might be due to the very high level of recombination of this species. The fact that we observed significant correlations between the similarity of ohnologs more than 88% similar and the GC-content in both *Arabidopsis* and yeast is therefore likely due to the relatively high level of mismatches between these duplicated genes relative to those of alleles and the fact that ectopic conversions occur even in self-fertilizing species (Haubold et al. 2002).

Acknowledgments

We would like to thank the two anonymous reviewers for their constructive comments on previous versions of the manuscript. This research was supported by the National Sciences and Engineering Research Council Discovery grants to A.M. and G.D.

Literature Cited

Bill, C. A., W. A. Duran, N. R. Miselis, and J. A. Nickoloff. 1998. Efficient repair of all types of single-base mismatches in recombination intermediates in Chinese hamster ovary cells. Competition between long-patch and G-T glycosylase-mediated repair of G-T mismatches. Genetics **149**:1935–1943.

Birdsell, J. A. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol. Biol. Evol. **19**:1181–1197.

Blanc, G., K. Hokamp, and K. H. Wolfe. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res. **13**:137–144.

Borts, R. H., and J. E. Haber. 1987. Meiotic recombination in yeast: alteration by multiple heterozygosities. Science **237**:1459–1465.

Brown, T. C., and J. Jiricny. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. Cell **54**:705–711.

———. 1989. Repair of base-base mismatches in simian and human cells. Genome **31**:578–583.

Drouin, G. 2002. Characterization of the gene conversions between the multigene family members of the yeast genome. J. Mol. Evol. **55**:14–23.

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. **12**:640–649.

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

Fullerton, S. M., A. Bernardo Carvalho, and A. G. Clark. 2001. Local rates of recombination are positively correlated with

GC content in the human genome. Mol. Biol. Evol. **18**: 1139–1142.

Galtier, N. 2003. Gene conversion drives GC content evolution in mammalian histones. Trends Genet. **19**:65–68.

Galtier, N., G. Piganeau, D. Mouchiroud, and L. Duret. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics **159**:907–911.

Gao, L. Z., and H. Innan. 2004. Very low gene duplication rate in the yeast genome. Science **306**:1367–1370.

Gerton, J. L., J. DeRisi, R. Shroff, M. Lichten, P. O. Brown, and T. D. Petes. 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA **97**:11383–11390.

Goldman, A. S., and M. Lichten. 1996. The efficiency of meiotic recombination between dispersed sequences in *Saccharomyces cerevisiae* depends upon their chromosomal location. Genetics **144**:43–55.

Haber, J. E., W. Y. Leung, R. H. Borts, and M. Lichten. 1991. The frequency of meiotic recombination in yeast is independent of the number and position of homologous donor sequences: implications for chromosome pairing. Proc. Natl. Acad. Sci. USA **88**:1120–1124.

Haubold, B., J. Kroymann, A. Ratzka, T. Mitchell-Olds, and T. Wiehe. 2002. Recombination and gene conversion in a 170-kb genomic region of *Arabidopsis thaliana*. Genetics **161**:1269–1278.

Hickey, D. A., S. Wang, and C. Magoulas. 1994. Gene duplication, gene conversion and codon bias. Pp. 199–207 *in* G. B. Golding, ed. Nonneutral evolution: theories and molecular data. Chapman and Hall, Inc., New York.

Jensen-Seaman, M. I., T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, C. F. Chen, M. A. Thomas, D. Haussler, and H. J. Jacob. 2004. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. **14**:528–538.

Johnson, L. J., V. Koufopanou, M. R. Goddard, R. Hetherington, S. M. Schäfer, and A. Burt. 2004. Population genetics of the wild yeast *Saccharomyces paradoxus*. Genetics **166**: 43–52.

Kong, A., D. F. Gudbjartsson, J. Sainz et al. (16 co-authors). 2002. A high-resolution recombination map of the human genome. Nat. Genet. **31**:241–247.

Kudla, G., A. Helwak, and L. Lipinski. 2004. Gene conversion and GC-content evolution in mammalian Hsp70. Mol. Biol. Evol. **21**:1438–1444.

Langkjaer, R. B., P. F. Cliften, M. Johnston, and J. Piskur. 2003. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. Nature **421**:848–852.

Lindahl, T. 1993. Instability and decay of the primary structure of DNA. Nature **362**:709–715.

Marais, G. 2003. Biased gene conversion: implications for genome and sex evolution. Trends Genet. **19**:330–338.

Marais, G., B. Charlesworth, and S. I. Wright. 2004. Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. Genome Biol. **5**:R45.

Marais, G., D. Mouchiroud, and L. Duret. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc. Natl. Acad. Sci. USA **98**:5688–5692.

Meunier, J., and L. Duret. 2004. Recombination drives the evolution of GC-content in the human genome. Mol. Biol. Evol. **21**:984–990.

Modrich, P., and R. Lahue. 1996. Mismatch repair in replication fidelity, genetic recombination, and cancer biology. Annu. Rev. Biochem. **65**:101–133.

Noonan, J. P., J. Grimwood, J. Schmutz, M. Dickson, and R. M. Myers. 2004. Gene conversion and the evolution of protocadherin gene cluster diversity. Genome Res. **14**:354–366.

Petes, T. D., and C. W. Hill. 1988. Recombination between repeated genes in microorganisms. Annu. Rev. Genet. **22**:147–168.

Sharp, P. M., E. Cowe, D. G. Higgins, D. C. Shields, K. H. Wolfe, and F. Wright. 1988. Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. Nucleic Acids Res. **16**:8207–8211.

Sueoka, N. 2002. Wide intragenomic G + C heterogeneity in human and chicken is mainly due to strand-symmetric directional mutation pressures: dGTP-oxidation and symmetric cytosine-deamination hypotheses. Gene **300**:141–154.

Takano-Shimizu, T. 2001. Local changes in GC/AT substitution biases and in crossover frequencies on *Drosophila* chromosomes. Mol. Biol. Evol. **18**:606–619.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Wolfe, K. H. 2001. Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. **2**:333–341.

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

Wright, S. I., N. Agrawal, and T. E. Bureau. 2003. Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. Genome Res. **13**: 1897–1903.

Manolo Gouy, Associate Editor

Accepted May 24, 2005