Letter to the Editor

Does the Recombination Rate Affect the Efficiency of Purifying Selection? The Yeast Genome Provides a Partial Answer

Csaba Pál,∗† Balázs Papp,† and Laurence D. Hurst‡

∗Collegium Budapest, Institute for Advanced Study, Budapest, Hungary; †Department of Plant Taxonomy and Ecology, Eötvös Loránd University, Budapest, Hungary; and ‡Department of Biology and Biochemistry, University of Bath, Bath, England

Much attention has been focused on the effect of recombination on preservation of favorable variants and on reducing the mutational load (Muller 1964; Kon drashov 1993; Barton 1995; Hurst and Peck 1996; West, Lively, and Read 1999). As most amino acid changes are slightly deleterious (Li 1997), their accumulation is expected to depend on the efficiency of purifying selection. Theoretical models have revealed that recombination might prevent the accumulation of such slightly deleterious mutations. This is potentially of importance because it suggests that purifying selection is less effective in genomic regions of low recombination (Hill and Robertson 1966; Carvalho and Clark 1999; Comeron, Kreitman, and Aguade 1999; Comeron and Kreitman 2000; Williams and Hurst 2000).

If purifying selection were more effective in regions of high recombination, accounting for within-genome variation, the rate of protein evolution should be lower in order to minimize the harmful effect of mutations on protein function. Furthermore, codon usage bias, which is at least partly influenced by selection, should be more pronounced in these genomic regions. Previous tests have employed this logic and shown that codon usage bias is slightly more biased under high recombination rates in Drosophila melanogaster (Comeron, Kreitman, and Aguade 1999). Two other studies (Carvalho and Clark 1999; Comeron and Kreitman 2000) have confirmed a negative relationship between intron size and recombination rate in the same organism (although the two groups give different interpretations of their results). One possible shortcoming of these works is that they did not control for variation in gene expression level. Some experimental works on yeast suggest that genomic regions with high transcription rates are also more prone to the initiation of meiotic double-strand break (DSB) events (Gerton et al. 2000), and there are also some hints that this feature might be a general feature of eukaryotic organisms (e.g., Jones et al. 1997). As codon usage bias (e.g., Duret and Mouchiroud 1999) and intron size (Vinogradov 2001) are affected by differences in gene expression level in numerous model organisms, the conclusions drawn from the association of these parameters with recombination rate should be treated with caution.

Yeast is currently a unique tool with which to address all of these issues owing to the availability of recombination (Gerton et al. 2000) and large-scale microarray expression data (Holstege et al. 1998) from Saccharomyces cerevisiae, along with extensive sequence data from S. cerevisiae (Ball et al. 2001) and the closely related Candida albicans. Here we show that recombination frequency has only weak, albeit significant, effects on protein evolution, on codon usage, and on protein size. Whether these relationships are owing to variation in the efficiency of purifying selection is unclear. We show that gene expression rates are correlated with recombination rates, and when we control for expression rates, the effect of recombination on the above parameters becomes only marginally significant.

Complete Saccharomyces protein and DNA sequences were obtained from Saccharomyces Genome Database (Ball et al. 2001), while sequence data for C. albicans were obtained from the Stanford DNA Sequencing and Technology Center website at http://www-sequence.stanford.edu/group/candida (contig version 6 was used for this purpose). We carried out BLASTP 2.0 amino acid sequence similarity searches of the S. cerevisiae protein against all available complete C. albicans proteins and vice versa using BLOSUM62 substitution matrices and the SEG filter for low-complexity regions (Altschul et al. 1997). The criteria used to define orthologs were (1) a BLASTP significance (E) value of <10⁻²⁰ and (2) the occurrence of the highest blast score when the putative orthologs were searched against each other. It is also worth noting that numerous sequences in S. cerevisiae belong to large, diverged protein families, but the duplication events most likely occurred before the divergence of the Candida and Saccharomyces lineages. As these sequences did not match our orthology criteria (see above), they were omitted. Some sequence alignments were ambiguous, and therefore they were omitted from the analysis. With the resulting 3,347 gene pairs, protein sequence alignments were carried out with CLUSTAL W, version 1.8, using the default settings (Thompson, Higgins, and Gibson 1994). Protein distances (d₅) were calculated using PHYLIP (Felsenstein 1989), employing PAM substitution matrices and default settings. Only gene pairs with moderate protein distances (d₅ < 1.5) were retained. To estimate the recombination rate adjacent to each open reading frame (ORF), a published data set was used (Gerton et al. 2000). Gerton et al. (2000) used seven DNA microarrays to estimate variation in the level of nearby meiotic DSBs along the whole yeast genome. The median value for each ORF was used as an approximation of local recombination rates, leading to a final data set of 2,758 gene pairs. In figure 1, three groups of genes are defined: those with low, medium, and high recombination rates.
The data set was divided into three subsets of equal sizes. Whole-genome transcription data from Holstege et al. (1998) were used as estimates for gene expression levels. Using microarray technology, Holstege et al. (1998) estimated the number of mRNA molecules per cell and the mRNA half-life for a large collection of genes in the *Saccharomyces cerevisiae* under conditions of growth to mid-log phase in YPD media.

To analyze the effects of recombination on rates of evolution, a reasonably large data set was compiled including the protein distances of *Candida-Saccharomyces* orthologs (see above). Although recombination data come only from *S. cerevisiae*, whereas protein evolution depends also on what has happened during evolution in the *Candida* lineage, we still found a highly significant association between the absolute rate of protein evolution ($d_A$) and recombination frequency ($r = -0.117, P < 10^{-9}$; see fig. 1). There are some further clues suggesting a relationship between the efficiency of selection and recombination. In agreement with previous findings for *Drosophila* (Comeron, Kreitman, and Aguade 1999), we found a positive association between codon adaptation index (which measures departure from random codon usage in favor of optimal codons) and recombination frequency ($r = 0.100, P < 10^{-6}$).

Protein size distribution is also influenced by recombination pattern: genes coding for long proteins also tend to occur in regions of low recombination ($r = -0.161, P < 10^{-6}$). This result can be explained if we assume that selection acts on gene length in order to reduce the expense of transcription or translation. Because large proteins are energetically expensive to make, such proteins are disadvantageous if shorter proteins can perform the same function (Moriyama and Powell 1998). Hence, longer proteins are allowed only in genomic regions in which the strength of purifying selection is sufficiently weak. However, we cannot exclude the possibility that recombination may induce deletions, and hence we examined a mutation bias rather than a result of selection (see also Comeron and Kreitman 2000).

The analyses above did not exclude the possibility that recombination per se was not the key variable. It has recently become apparent that transcriptionally active genomic regions are more prone to the initiation of meiotic recombination events (Nicolas 1998; Gerton et al. 2000). This we can confirm. Using a public database (Holstege et al. 1998), we found a significant association between transcriptional frequency and recombination rate ($r = 0.104, P < 10^{-6}$). It has also previously been shown that highly expressed genes in yeast evolve slowly (Pål, Papp, and Hurst 2001).

Although the interpretation of this latter result is uncertain, one might argue that the negative association between recombination frequency and rate of protein evolution may simply reflect the covariation of gene expression level and recombination rate. In a similar vein, codon bias is especially pronounced in highly expressed genes in yeast, most likely reflecting selection for enhanced efficiency of translation (e.g., Coghlan and Wolfe 2000). Hence, the association between codon adaptation and recombination might be explained by the covariation of recombination rate and gene expression level. A similar explanation may hold for protein size distribution. Highly expressed proteins are expected to pose a higher energetic cost for the organism. Hence, it is conceivable that selection pressure for reduced protein size is especially pronounced in highly expressed genes (Moriyama and Powell 1998; Coghlan and Wolfe 2000).

We found that the nonindependence of recombination and gene expression level can explain in large part the patterns found above. Using multiple-regression analysis (Sokal and Rohlf 1995) to control for the covariation of gene expression and recombination rate, we found that the correlation between recombination rate and protein evolution ($d_A$) became only marginally significant (table 1). Thus, only a very small fraction of

\[
\text{Table 1} \\
\text{Results from Multiple-Regression Analysis of Recombination Rate and Gene Expression Level with Different} \\
\text{Dependent Variables} \\
\begin{array}{|c|c|c|c|}
\hline
\text{Dependent Variable} & \text{Pearson Correlation with} & \text{Effect of Recombination} & \text{Effect of Logarithm of} \\
& \text{Recombination Rate} & \text{Rate when Controlled for} & \text{mRNA Copies per Cell} \text{a} \\
& & \text{the Logarithm of mRNA} & \text{when Controlled for} \\
& & \text{Copies per Cell} \text{a} & \text{Recombination Rate} \text{a} \\
\hline
\text{d}_A & r = -0.117, P < 10^{-9} & \beta = -0.04, P < 0.05 & \beta = -0.51, P < 10^{-6} \\
\text{CAI (Saccharomyces cerevisiae)} & r = 0.100, P < 10^{-6} & \beta = 0.004, P = 0.76 & \beta = 0.693, P < 10^{-6} \\
\text{Protein size (S. cerevisiae)} & r = -0.161, P < 10^{-6} & \beta = -0.127, P < 10^{-6} & \beta = -0.216, P < 10^{-6} \\
\hline
\end{array}
\]

\text{NOTE.} d_A \text{ = protein distance; CAI = codon adaptation index.} \\
\text{a} \text{ Because of the limited data on mRNA copies per cell, 2,642 gene pairs were included in the multiple regression analysis.}
the variation in protein evolution is attributable to local differences in genomic recombination rates when expression level is controlled for. In contrast, the correlation between protein size and recombination rate remains largely unaffected.

Can we be certain that the high rate of evolution of genes in regions of low recombination and low gene expression is due to weaker purifying selection, or could it be that regions with a low recombination rate or a low gene expression level may also have a higher mutation rate? Experimental facts do not seem to support this hypothesis. In yeast, increased transcription rates are associated with increased mutation rates (Datta and Jinks-Robertson 1995; Morey, Greene, and Jinks-Robertson 2000). Furthermore, recombinational repair of DSBs in yeast, traditionally believed to be an error-free DNA repair pathway, was recently shown to substantially increase the frequency of point mutations in a nearby interval (Strathern, Shafer, and McGill 1995; Holbeck and Strathern 1997). These are the opposite of the patterns one would expect to explain the negative correlation between recombination and rate of protein evolution. Therefore, the association between recombination and rate of protein evolution is unlikely to result from differences in mutational patterns alone.

What about the association between recombination rate and codon usage bias? There is a general relationship between recombination rate and genomic G+C content in numerous eukaryotes, including yeast (see Gerton et al. 2000). This possibly reflects mutational bias toward G and C in regions of high recombination. If optimal codons are GC-rich, then mutational bias could alone produce a positive correlation between recombination rate and codon usage. This mechanism is most likely to work in D. melanogaster and Caenorhabditis elegans (Marais, Mouchiroud, and Duret 2001). However, this neutral scenario is unlikely to hold for yeast. First, in contrast to D. melanogaster and C. elegans, optimal codons in yeast do not preferentially end in G or C (Sharp and Cowe 1991). Second, codon usage is largely influenced by gene expression rather than by recombination rate (table 1). Third, there is no significant negative association between codon adaptation index and recombination rate when only AT-ending codons are considered (unpublished data). This is in contrast to the prediction of the neutral scenario (Marais, Mouchiroud, and Duret 2001).

However, we cannot exclude the possibility that recombination induces deletions. Hence, the association between recombination rate and protein size may be the result of mutation bias rather than selection. It is worth noting that protein size was the only parameter in our study for which the association with recombination remained largely unaffected after controlling for gene expression level.

From these results, we conclude that when gene expression level is controlled for, genomic variation in recombination rates has only a very weak influence on the efficiency of purifying selection. This conclusion is drawn from the analysis of the unicellular yeast, in which meiotic recombination events and transcription occur in the same cell. What are the implications of our finding for multicellular organisms, where meiosis is restricted to germ cells? Clearly, in this case expression level in somatic tissues does not matter: only transcription in germ cells can influence meiotic recombination rate. Indeed, in humans, male meiotic recombination is specifically associated with transcription in testis tissue (Jones et al. 1997). If generally true, this suggests that the examination of covariance with recombination is a problematic method of studying variation in the intensity of selection, and previous results from the use of this method (e.g., Carvalho and Clark 1999; Comeron, Kreitman, and Aguaude 1999) need to be treated with caution until covariance with germ cell expression level is eliminated as causative.

LITERATURE CITED


William Martin, reviewing editor

Accepted July 28, 2001