LETTER

Selection to Maintain Paralogous Amino Acid Differences Under the Pressure of Gene Conversion in the Heat-Shock Protein Genes in Yeast

Shohei Takuno and Hideki Innan
Graduate University for Advanced Studies, Hayama, Kanagawa, Japan

A genome scan for the signatures of selection for paralogous functional amino acid differences was performed with yeast genomes. This recently developed method makes it possible to localize the target sites of selection under the pressure of gene conversion. We found that two gene pairs have strong signatures of selection. The two pairs of duplicated genes happened to be heat shock genes (Ssa1/ Ssa2 and Ssb1/ Ssb2), which have similar protein structures to each other, although the amino acid sequence identity between Ssa and Ssb is not high (~60%). Interestingly, the two gene pairs exhibit signature of selection at almost identical positions within the substrate-binding domain β. Because this domain specifies the substrate polypeptides, it is presumed that functional divergence may be advantageous in this domain. Evolutionary analysis demonstrated that the observed divergence in the two gene pairs has been maintained in many yeast species independently, suggesting long-term operation of strong selection.

It occasionally occurs that a duplicated gene evolves a novel function before it is silenced by null mutations (Ohno 1970; Walsh 2003). This phenomenon is known as neofunctionalization, and the probability that neofunctionalization occurs is given by a relatively simple equation, \( p_S \), where \( p \) is the ratio of the neofunctionalizing mutation rate to the null mutation rate (Walsh 1995) and \( S \) is the population selection parameter, \( 4N_e s \), where \( N_e \) is the effective population size and \( s \) is the selection intensity for the neofunctionalized allele. Because the rate of neofunctionalizing mutation is much smaller than that of null mutation, neofunctionalization is not a likely fate of duplicated genes.

Neofunctionalization is even less likely when interlocus (or nonallelic) gene conversion between duplicated genes is active (Innan 2003a; Teshima and Innan 2008). It is easy to imagine that gene conversion works against neofunctionalization. Suppose in a simple two-locus model, a neofunctionalized allele \( B \) is fixed in one locus, whereas the original allele \( A \) is fixed in the other. Without gene conversion, this can be considered as a terminal fate (i.e., neofunctionalization) because the back mutation rate from \( B \) to \( A \) should be very low. However, this logic does not hold with gene conversion because the neofunctionalized allele \( B \) can be easily mutated back to \( A \) with a reasonably high rate of gene conversion. In this situation, although the process is complicated, a key factor is the ratio of \( s \) to \( c \), where \( c \) is the gene conversion rate per site per generation (Teshima and Innan 2008). Note that this is the rate that a site is involved in a gene conversion event per generation, so that the rate is identical to the product of the rate of gene conversion event and the average conversion tract length (Innan 2002). \( A \) and \( B \) can be stably maintained when selection is much stronger than the gene conversion rate (i.e., \( s \gg c \) ) so that deleterious gene conversion is immediately selected against and disappears from the population (Teshima and Innan 2008). Therefore, neofunctionalization could be a very unlikely fate in gene conversion–rich species.

Yeast (Saccharomyces cerevisiae) might be one of the gene conversion–rich species and has long been a model species of gene conversion (both allelic and interlocus gene conversion) (Petes and Hill 1988; Mancera et al. 2008). Evolutionary analyses have shown that duplicated genes in yeast are indeed subject to a high rate of interlocus gene conversion, which causes a long-term coevolution of duplicated genes (i.e., concerted evolution; Ohta 1980; Zimmer et al. 1980; Dover 1982). A striking observation is that a number of duplicates originated from a whole-genome duplication (WGD) about 100–150 Ma still have almost identical DNA sequences (Wolfe and Shields 1997; Kellis et al. 2003; Gao and Innan 2004; Sugino and Innan 2006). A high rate of interlocus gene conversion is also demonstrated by polymorphism data, which make it possible to estimate the population conversion rate (Innan 2003b). Here, by using whole-genome polymorphism data in S. cerevisiae (Liti et al. 2009), we estimated the gene conversion rates for 17 pairs of duplicated genes undergoing concerted evolution. It was found that the estimated ratio of gene conversion rate to mutation rate ranges from 7 to 100 in most gene pairs and it exceeds 100 in four pairs (see Supplementary Material online). Given these observations, we might predict that neofunctionalization should be very unlikely in yeast unless some kind of physical interference of gene conversion (e.g., large insertions and deletions [indels] or integration of transposable elements) is involved.

Nevertheless, by applying a simple genome scan algorithm to the yeast genomes, we found strong evidence for some form of selection to maintain paralogous amino acid divergence (e.g., neofunctionalization) in two pairs of duplicated genes that are undergoing long-term concerted evolution (for details, see Supplementary Material online). The idea of this analysis is based on the theoretical prediction that a peak of the divergence between duplicates arises when a neofunctionalized allele is fixed in one of the duplicated genes if they are currently undergoing concerted evolution (Innan 2003a; Teshima and Innan 2008). Frequent gene conversion usually keeps the nucleotide identity between duplicates high over the duplicated region, but...
Fig. 1.—Summary of the evolutionary analyses of (A) Ssb1/Ssb2 and (B) Ssa1/Ssa2. For each, the left panel shows the spatial distributions of nucleotide divergence and the numbers of types N and C sites. Sliding window analyses for divergence were performed with a 100-bp window which is moved with 20-bp increment. The distributions of the numbers of type N and type C sites are shown such that the entire coding region is divided into 50-bp subregions. The putative target sites of selection are shown with arrows. The right panel illustrates the evolutionary history of the two duplicated genes and putative targets of selection on the species tree of seven yeast species. Estimated genes trees of Ssb1/Ssb2 and Ssa1/Ssa2 are shown in supplementary fig. S3 (Supplementary Material online), which shows strong evidence for gene conversion in all lineages.

The genome scan algorithm is specified to detect signature of selection to maintain paralogous divergence by eliminating gene conversion. In other words, the algorithm can be applied to duplicated genes undergoing gene conversion. Therefore, we first identified duplicated genes in S. cerevisiae with evidence for gene conversion using the genome of a close relative, Saccharomyces paradoxus (see Supplementary Material online). The two species split roughly 10 Ma, and the synonymous divergence is ≈ 0.4. We found that in total 20 duplicated genes are shared by the two species and exhibit evidence for concerted evolution (supplementary tables S1 and S2, Supplementary Material online). In brief, we confirmed the action of gene conversion not only by synonymous paralogous divergence but also by the shape of gene tree, site configuration in the aligned sequences, and polymorphism (for details, see Supplementary Material online). These genes with highly reliable evidence for gene conversion were used for the genome scan, and we found strong evidence for selection in two gene pairs (see Supplementary Material online). Furthermore, we successfully specified the target sites of selection.

Both of the two gene pairs (Ssb1/Ssb2 and Ssa1/Ssa2) with strong evidence for selection encode heat-shock proteins. Figure 1 shows the distribution of paralogous divergences in the two gene pairs. In each pair, there is a peak of divergence in the coding region, exhibiting strong evidence for selection. We also found elevated divergence around the boundary of the coding region for several gene pairs (fig. 1). Although the possibility of selection in those regions cannot be ruled out, we did not consider them as footprints of selection because it is also likely
that gene conversion rate is reduced around the boundary region, where gene conversion can be less successful (for details, see Osada and Innan 2008). Elevated paralogous divergence around boundary is also frequently observed in Drosophila (Osada and Innan 2008). Our simulation showed that the probability that the observed peaks appear by chance (without selection) is extremely low ($P < 10^{-4}$; see supplementary table S3, Supplementary Material online).

Each of the observed peaks represents a local region in which gene conversion is strongly restricted, which is obvious in the alignment of the four sequences, two duplicates in S. cerevisiae (denoted by Xc and Yc) and their orthologs in S. paradoxus (denoted by Xp and Yp). Informative sites in the alignment were classified into two categories, type C and type N sites (“C” and “N” represent “conversion” and “nonconversion”, respectively), following Osada and Innan (2008). The allelic state of the four sequences at a type C site is given by \{$Xc,Yc,Xp,Yp\} = \{0,0,1,1\}$, which is a typical pattern at sites that experienced interlocus gene conversion, whereas \{$Xc,Yc,Xp,Yp\} = \{0,1,0,1\}$ at a type N site, a typical pattern when there is no gene conversion after speciation. Most of the informative sites belong to either of the two categories (but see below). As theory predicts, the peak of divergence is represented by a cluster of type N sites (table 1), whereas in the surrounding region, there are many type C sites, indicating that gene conversion is active (fig. 1). Thus, the two species share the peak of divergence at the same location, indicating that the peak appeared before the speciation event of the two species and has been independently preserved until present (i.e., $\sim 10$ My). The peak of the Ssb1/Ssb2 gene pair consists of five nonsynonymous type N sites characterizing three amino acid differences between Ssb1 and Ssb2, whereas the peak of the Ssa1/Ssa2 gene pair includes only one nonsynonymous site (i.e., one amino acid difference, I vs. V). Those amino acid differences are presumably protected by selection against gene conversion for a long time. Even though the amino acids I and V are biochemically similar, this difference for the Ssa1/Ssa2 gene pair should play a role to characterize the products of the two genes because there is no other nonsynonymous difference.

It is difficult to find other mechanisms to explain the observed clusters of type N sites than selection, for example, relaxation of negative selection or local elevation of mutation rate. They would explain a peak of divergence alone, but it hardly predicts the observation that identical sequences are kept in the two species. Furthermore, we did not find elevated divergence around the putative target sites of selection when the orthologous genes were compared (not shown). Thus, a local reduction in the gene conversion rate seems to be a very likely explanation. Gene conversion may be locally restricted if there is a drastic change in DNA structure (e.g., large indel or transposon), but this is unlikely in coding region. For the two pairs of duplicated genes, we found no structural difference.

A close look at the DNA sequence variation around the divergence peaks in the two gene pairs reveals an interesting type of informative sites, which do not belong to either type C or N (table 1). At the third positions at the 416th amino acid of Ssb1/Ssb2 and at 426th amino acid of Ssa1/Ssa2, the configuration can be written as \{$Xc,Yc,Xp,Yp\} = \{0,0,1,1\}$, which is denoted by type N$. Multiple mutations at a single site can create this pattern, but a more likely explanation here may be a flip of nucleotide states between paralogs, which occurred in one of the lineages to the two species. Suppose the ancestral state is \{$X,Y\} = \{0,1\}$. Gene conversion creates \{0,0\} and \{0,1\} and their recombination results in \{1,0\}. If such a \{1,0\} site is fixed in one lineage, we observe a type N$ site. This scenario may be more likely when the gene conversion rate is much higher than the mutation rate, which holds here. In other words, type N$ sites indicate the action of gene conversion around the putative target sites of selection, even though gene conversion should be eliminated by selection in a short time.

Phylogenetic analysis of other yeast species (Cliften et al. 2003; Kellis et al. 2003) makes it possible to further trace back the origin of the peak (fig. 1). For the Ssb1/Ssb2 pair, we found that they were produced by

---

**Table 1**

**Summary of Informative Sites around the Putative Target Sites of Selection**

<table>
<thead>
<tr>
<th>Species</th>
<th>Gene</th>
<th>Amino Acid Position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type</td>
<td>378</td>
</tr>
<tr>
<td>R or S</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>Ssb1</td>
<td>gtT</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>Saccharomyces paradoxus</td>
<td>Ssb1</td>
<td>gtT</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>Ssa1</td>
<td>gaT</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>Ssa1</td>
<td>gaT</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>S</td>
</tr>
</tbody>
</table>

Note.—Variable sites are shown in capitals. Putative target sites of selection are shown in bold.

a Type of site configuration. See text for details.

b Amino acid replacement (R) or synonymous (S). When a replacement change is involved, two alternative amino acids are shown in parentheses.
the WGD event and that the origin of two amino acid differences (MC/IV) predates the speciation event of *S. cerevisiae* and *Saccharomyces bayanus*, which occurred ∼20 Ma. For the Ssa1/Ssa2 gene pair, the duplication event may be older than the speciation of *S. cerevisiae* and *S. bayanus*, and the origin of the IV amino acid difference should predate the speciation of *S. cerevisiae* and *Saccharomyces mikatae*. These results indicate very old origins of those mutations. Strong selection should have worked to maintain them independently in multiple lineages under the pressure of homogenization by gene conversion.

Polymorphism data in *S. cerevisiae* (Liti et al. 2009) were also useful to estimate the selection intensity. For the Ssb1/Ssb2 pair, we assumed equal intensity of selection for the three amino acid differences because we cannot determine the true target of selection. We found that the two types (MCA and IVS) are completely fixed in each gene, indicating extremely strong selection (*s* = ∞ according to equation (18) in Innan 2003a). At the Ssa1/Ssa2 gene pair, only one nonsynonymous type N site is involved in the peak, which characterizes the two types, I and V. We found that almost all individuals (35/37) have I in Ssa1 and V in Ssa2, so that the two amino acids are nearly fixed at each locus. One of the remaining two individuals has I and the other has V at both loci due to recent gene conversion. This observation gives an estimate of the selection intensity (*s*) to be ∼40 times greater than the gene conversion rate. Thus, the two gene pairs well satisfy the condition of *s* ≫ *c*.

It is intriguing that the two gene pairs with strong signatures of selection happened to be chaperone genes that belong to the Hsp70 protein superfamily (70 kDa heat-shock proteins). Obviously, they have similar functions: impeding protein aggregation and assisting in the refolding of misfolded proteins under stressed environments (Hartl and Hayer-Hartl 2002; Mayer and Bukau 2005). The four intronless genes encode ∼600 amino acids of chaperone proteins, but the amino acid sequence identity between Ssa and Ssb is not so high (∼60%). Ssb1 and Ssb2 are associated with translating ribosomes and bind to nascent polypeptide, whereas Ssa1 and Ssa2 play a role in protein folding in cytoplasm (Craig et al. 1993). Thus, although the two gene pairs have some functional difference with quite much amino acid divergence, they still keep similar domain structures, and the signatures of selection are found in a similar location within the substrate-binding domain β (Hartl and Hayer-Hartl 2002; Mayer and Bukau 2005). This domain might determine the specificity of substrate polypeptides, so that it may be possible to presume that amino acid variation in this domain can be beneficial to increase the specificity (i.e., hot spot of functional divergence).

It should also be pointed out that the Ssa1/Ssa2 gene pair in *S. bayanus* has the I and T amino acids, respectively, at the position where the gene pair in *S. cerevisiae*, *S. paradoxus*, and *S. mikatae* has I and V (fig. 1). As the ancestral state would be I, it may be possible that in the two lineages independently, selection maintains functional differences with independent origins. This would also support our finding; this region would be a hot spot of functional divergence.

Thus, our genome scan detected strong signatures of selection for paralogous amino acid divergence in two chaperone gene pairs: Ssb1/Ssb2 and Ssa1/Ssa2. The patterns of divergence and polymorphism in those genes well agree with theoretical predictions. It is very likely that diversification of their functional specificity would be the target of long-term strong selection. Our successful application of the genome scan for the signatures of selection indicates that this can be a promising method to detect selection for functional divergence in early stages of duplicated genes.

**Supplementary Material**

Supplementary methods, tables S1–S3, and figs. S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

We are grateful to K.H. Wolfe, two anonymous reviewers, and the members of the Innan Lab for various comments and discussion. S. T. is a research fellow of the Japan Society for the Promotion of Science (JSPS). H.I. is in part supported by grants from JSPS, National Science Foundation (USA), and the Graduate University for Advanced Studies.

**Literature Cited**


