Natural Selection on Gene Order in the Genome Reorganization Process After Whole-Genome Duplication of Yeast

Ryuichi P. Sugino¹ and Hideki Innan*¹,²

¹Department of Evolutionary Studies of Biosystems, Graduate University for Advanced Studies, Hayama, Kanagawa, Japan
²PRESTO, Japan Science and Technology Agency, Saitama, Japan

*Corresponding author: E-mail: innan.hideki@soken.ac.jp.

Associate editor: James McInerney

Abstract

A genome must locate its coding genes on the chromosomes in a meaningful manner with the help of natural selection, but the mechanism of gene order evolution is poorly understood. To explore the role of selection in shaping the current order of coding genes and their cis-regulatory elements, a comparative genomic approach was applied to the baker’s yeast Saccharomyces cerevisiae and its close relatives. S. cerevisiae have experienced a whole-genome duplication followed by an extensive reorganization process of gene order, during which a number of new adjacent gene pairs appeared. We found that the proportion of new adjacent gene pairs in divergent orientation is significantly reduced, suggesting that such new divergent gene pairs may be disfavored most likely because their coregulation may be deleterious. It is also found that such new divergent gene pairs have particularly long intergenic regions. These observations suggest that selection specifically worked against deletions in intergenic regions of new divergent gene pairs, perhaps because they should be physically kept away so that they are not coregulated. It is indicated that gene regulation would be one of the major factors to determine the order of coding genes.

Key words: gene order, whole genome duplication, budding yeast, gene regulation.

Introduction

It is widely accepted that the order of coding genes is not random (Hurst et al. 2004), most likely because of complicated relationships between the locations of coding genes and their cis-regulatory and promoter regions. A number of investigations have focused on how coding genes are organized in eukaryote genomes and found relatively weak and indirect evidence for nonrandom order. An example is a general tendency that closely located genes have similar expression patterns. There are a number of observations to support this in a wide range of species: yeasts (Cho et al. 1998), fruit flies (Boutanaev et al. 2002; Spellman and Rubin 2002), plants (Williams and Bowles 2004), nematodes (Lercher et al. 2003), and mammals (Fukuoka et al. 2004; Singer et al. 2005; Li et al. 2006; Sémon and Duret 2006).

In addition, direct evidence for the adaptive formation of new gene orders is available for a few cases, including the translocation of SSU1 genes in the winery yeast (Pérez-Ortín et al. 2002), the formation of the DAL gene cluster in the lineage of Saccharomyces sensu stricto species (Wong and Wolfe 2005), and the independent clustering of the GAL genes of many fungi species (Slot and Rokas 2010). An experimental study (Dunham et al. 2002) demonstrated that a certain gene order arose and fixed in multiple independently evolving strains, exhibiting strong evidence for positive selection on gene order. Thus, there are many lines of evidence that adaptive selection has played a role, but our knowledge on the relative contribution of natural selection in determining the order of coding genes in eukaryote genomes is still very limited and under debate.

To address this, we focused on how natural selection has worked through the evolutionary changes of gene order in yeasts including the baker’s yeast Saccharomyces cerevisiae, with special attention to gene regulation. Thus far, S. cerevisiae has been the main model species in the studies of gene order because of the availability of tremendous amounts of molecular knowledge and data. One of the key empirical findings to resolve the mystery of gene order is that two adjacent genes can be coexpressed when the promoter region between them has a single nucleosome free region (NFR), where RNA polymerase (Pol) II binds and initiates transcription (Xu et al. 2009). This fact directly indicates that coregulation of multiple genes (especially adjacent genes in divergent orientation) is a key factor in the evolution of gene order. This is also consistent with earlier finding that divergent pairs have more similar patterns of gene expression (Cohen et al. 2000; Kruglyak and Tang 2000; Herr and Harris 2004; Trinklein et al. 2004; Kenschke et al. 2008). In this study, we explore how selection works on the physical locations of cis-regulatory elements represented by NFRs. When a new adjacent gene pair is formed, the locations of NFRs in the shared promoter regions should determine the degree of coexpression, thereby affecting the fitness of the new gene order. Based on this idea, we investigate how selection is involved in the evolutionary changes of gene order in yeasts.
Another advantage of using yeasts as a model of the evolutionary study of gene order would be their unique evolutionary history; a whole-genome duplication (WGD) occurred ~100–200 Ma (Wolfe and Shields 1997; Dietrich et al. 2004; Kellis et al. 2004). Genomic sequences are now available for a number of yeast species including those whose lineages diverged prior to and after the WGD. Comparative genomic analyses revealed rearrangement occurred after the WGD (Byrne and Wolfe 2005; Scannell et al. 2007). In this post-WGD genome reorganization process, a number of coding genes and intergenic regions have been lost, resulting in a number of new gene orders. This situation together with an excellent database (yeast gene order browser (YGOB) (Byrne and Wolfe 2005) provides us exciting opportunities to explore how gene order has changed through the post-WGD process at a fine scale. Previous research on the evolution of gene order provided several new insights. For example, Hurst et al. (2002) found that highly "coexpressed" adjacent gene pairs tend to keep their adjacent relationship through the post-WGD genome reorganization process, and this could be particularly applied to adjacent gene pairs in divergent orientation (Kensche et al. 2008). However, the situation may be different for genes with high expression. Byrnes et al. (2006) found that young adjacent pairs have relatively high expression and are located apart from each other, perhaps because their transcription may interfere by their adjacent genes if they were too close to each other. Hermsen et al. (2008) found a strange bimodal distribution of the intergenic length of adjacent genes in divergent orientations and suggested that selection might have worked on the cis-regulatory elements in the intergenic regions. Thus, although recent works have improved our understanding of the evolution of gene order, its evolutionary mechanism is still poorly understood, and the direct target of natural selection on gene order is an open question. Here, we used a comparative genomic approach, which revealed that selection on the physical locations of cis-regulatory elements plays a crucial role in the post-WGD genome reorganization process in yeast. Then, we developed a simulation model of genome evolution after a WGD, from which the intensity of selection was estimated.

Materials and Methods

Genome Sequence Data

Our analysis of genomic sequences of multiple yeast species is based on the data in the YGOB version 3.0 (Byrne and Wolfe 2005; Gordon et al. 2009), which includes ~5600 coding genes of S. cerevisiae. With this database, it is straightforward to trace the evolutionary changes of gene order along the genome evolution of yeast species. Our following analysis is based on the ancestral genome at the WGD event inferred by Gordon et al. (2009), which is also included in the YGOB. This ancestral genome is referred to as the pre-WGD genome (fig. 1).

The number of the coding genes in the YGOB is slightly smaller than that in the Saccharomyces Genome Database (SGD) (Cherry et al. 1997) because the YGOB has eliminated dubious annotations in the SGD. The YGOB database contains synteny (gene order) with the transcription orientation of coding genes (i.e., divergent (head–head), tandem (head–tail), or convergent (tail–tail)). We excluded tandem duplicated genes from the analysis because of their potential problems as repeatedly pointed out (Lercher et al. 2002; Williams and Hurst 2002; Batada et al. 2007). Tandem duplicated genes were detected using the BlastP algorithm with a cut-off value of E value <10^{-5}.

Locations of NFRs

We used the data of nucleosome positions in the S. cerevisiae genome, estimated by using H2AZ and H3/H4 (Albert et al. 2007; Mavrich et al. 2008). Given these data of nucleosome positions, we identified NFRs where the interspaces between nucleosomes are over 80 bp, according to the definition of Xu et al. (2009).

Gene Expression Data

The similarity score index of gene expression pattern for all adjacent gene pairs were computed using the data in ExpressDB (Aach et al. 2000), where timescale expression data of various (~40) conditions are available (Spellman et al. 1998; Gasch et al. 2000; Hughes et al. 2000; Roberts et al. 2000; Gasch et al. 2001; Iyer et al. 2001; Lieb et al. 2001; Natarajan et al. 2001; Olesen et al. 2002; Williams et al. 2002; Bulik et al. 2003; Fry et al. 2003). We downloaded data from http://longitude.weizmann.ac.il/BackUpCircuits/, which are normalized data of ExpressDB by Kafri et al. (2005). The similarity score was measured by Pearson’s correlation coefficient (r).

Results

Comparing the Genomes of Pre- and Post-WGD Species

We here demonstrate that the action of selection on gene order has dramatically changed after the WGD event in yeast, that occurred roughly 100–200 Ma (Wolfe and Shields 1997; Dietrich et al. 2004; Kellis et al. 2004; Sugino and Innan 2005). In relation to this event, different yeast species were classified into two categories, pre- and post-WGD species (fig. 1A). We first compared several properties of the genomes between the two categories. The current genomes of pre-WGD species have on average ~5,000 genes and the genome size are roughly 10 Mb (fig. 1A see also Clifton et al. 2003; Kellis et al. 2003; Byrnes et al. 2006; Génolevures Consortium et al. 2009). Because these numbers are quite constant in all pre-WGD species, it is straightforward to predict that the ancestral genome before the WGD event also had a similar gene number and genome size. This is indeed supported by the pre-WGD genome inferred by Gordon et al. (2009) (see also fig. 1). After the WGD, the ancestral genome was doubled, but the current post-WGD species exhibit only a 10% increase both in the genome size and in the number of genes. The length of intergenic regions is only slightly longer in the post-WGD species, indicating that the current genomes of post-WGD species are almost as compact as the pre-WGD genome,
Gene Order Evolution in Yeast

FIG. 1. Summary of the evolutionary analysis of adjacent genes. (A) Phylogenetic relationship of yeast species. The star represents the WGD event, which occurred ∼ 100–200 Ma. The table summarizes the genome sizes and the number of genes. Data of genome sizes are according to Liti et al. (2009), Cliften et al. (2003), Scannell et al. (2007), and Génoleves Consortium et al. (2009). (B) Illustration of a typical pattern of the post-WGD genome reorganization process in a hypothetical region, where the ancestral pre-WGD chromosome has nine genes, labeled A–I (open arrows). There are two descendant chromosomes in the post-WGD species. Genes lost in the post-WGD process are shown by open arrows with broken lines. From this pattern, the relationships of adjacent gene pairs in the current post-WGD species are inferred by a simple parsimonious algorithm (see text for details). (C) Distributions of $l$ for the three categories of adjacent gene pairs.

most likely because massive reduction in the genome size and gene number occurred in the early stages of the post-WGD process (Kellis et al. 2004; Scannell et al. 2006). Thus, it can be suggested that drastic genome reorganization occurred since the WGD event.

Evolution of Adjacent Gene Pairs

To explore the action of selection on gene order in the post-WGD genome reorganization process, we focused on the orientations of physically adjacent gene pairs in the genome. All adjacent gene pairs in the genome were classified into three categories in terms of orientation: divergent, tandem, and convergent pairs (see fig. 1B). We found that in post-WGD species, roughly half (47–48%) of the adjacent gene pairs are in tandem orientation and the others are in divergent and convergent orientations (∼26% for each) (fig. 1). In the pre-WGD genome, these proportions are also similar, although the proportions of divergent and convergent gene pairs (∼28% for each) are slightly larger than those of post-WGD species (fig. 1). Thus, the genome context of the post-WGD species is quite similar to that of the pre-WGD genome.

However, a closer look at the changes of gene order exhibits several lines of evidence that extensive selection operated after the WGD. To investigate the evolutionary changes of gene order, the adjacent gene pairs in the current genome of S. cerevisiae were further classified according to when they were formed, that is, those that were newly created after the WGD (referred to as “new” gene pairs) and those that already existed at the WGD event (referred to as “conserved” gene pairs). As illustrated in figure 1B, a number of new adjacent gene combinations arose after the WGD, providing an excellent opportunity for studying the evolution of gene order.

We used data from the YGOB (Byrne and Wolfe 2005), which clearly visualizes the post-WGD process through the comparison of multiple post- and pre-WGD genomes. By applying a simple parsimony algorithm to the YGOB data, we inferred the evolutionary histories of the current adjacent gene pairs in the S. cerevisiae genome. In practice, for each adjacent gene pair in the current S. cerevisiae genome, we estimated $l$, the number of genes lost in the lineage of S. cerevisiae since WGD. The basic idea of our method is described in figure 1. For each adjacent gene pair in S. cerevisiae, we identified the locations and orientations of their orthologous genes in the pre-WGD genome. We inferred $l$ for adjacent gene pairs whose orthologous genes in the pre-WGD genome are on the same chromosome with conserved relative orientations. For the example of the A$_2$-C$_2$ pair in figure 1B there has been a single gene loss in the lineage to S. cerevisiae after WGD, so we estimate $l = 1$ (the situation is identical for the E$_1$-G$_1$ and D$_2$-F$_2$ pair). No gene
Table 1. Coexpression and Intergenic Distance for Adjacent Gene Pairs in S. cerevisiae.

<table>
<thead>
<tr>
<th></th>
<th>Number of adjacent genes</th>
<th>Intergenic distance (in bp)</th>
<th>Coexpression (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Conserved</td>
<td>New</td>
</tr>
<tr>
<td><strong>CDS data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Divergent</td>
<td>751 (28.3%)</td>
<td>351 (22.0%)</td>
<td>581.85</td>
</tr>
<tr>
<td>Tandem</td>
<td>1179 (44.4%)</td>
<td>809 (50.7%)</td>
<td>487.20</td>
</tr>
<tr>
<td>Convergent</td>
<td>727 (27.4%)</td>
<td>435 (27.3%)</td>
<td>249.42</td>
</tr>
<tr>
<td><strong>UTR data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Divergent</td>
<td>555 (27.4%)</td>
<td>264 (21.5%)</td>
<td>414.99</td>
</tr>
<tr>
<td>Tandem</td>
<td>859 (42.3%)</td>
<td>589 (48.0%)</td>
<td>305.00</td>
</tr>
<tr>
<td>Convergent</td>
<td>614 (30.3%)</td>
<td>374 (30.5%)</td>
<td>−27.38</td>
</tr>
</tbody>
</table>

Note.—Data for i = 1, 2, and 3 are pooled. Very similar results were obtained when we restricted the analysis to i = 1.

*p < 0.0001

loss is needed to explain the four pairs (A₁-B₁, G₁-H₁, C₁-D₂, H₂-I₂); therefore l is estimated to be 0, indicating the adjacency of these pairs has been conserved since WGD.

The YGOB database consists of ~5,600 coding genes (i.e., ~5,600 adjacent pairs) in the S. cerevisiae genome and their orthologs in other yeast species and the inferred pre-WGD genome (Gordon et al. 2009). We successfully identified the orthologous gene pairs in the pre-WGD genome for ~80% of the adjacent genes in S. cerevisiae (n = 4, 617). We found that ~90% of them (n = 4, 440) have their orthologous genes on the same chromosomes in the same genomes in the pre-WGD genome, for which we estimated l. Figure 1C shows the distribution of l, indicating that ~60% (n = 2, 657) have been conserved as adjacent pairs since WGD (i.e., l = 0). For the remaining new pairs, the distribution of l is L-shaped and over 95% are explained by losing up to three genes between them. In the following analysis, to make the situation simple, we only focus on these new pairs with l ≤ 3, although we obtained almost identical results when those with l < 3 were included.

We first found that the proportion of divergent gene pairs in the conserved category (28.3%) is almost identical to that of the pre-WGD genome (~28.0%), but it is significantly reduced in the new adjacent gene pairs (20.7%, P < 0.0001, exact test, table 1). Given that roughly a quarter of newly arisen gene pairs would be divergent if random, it can be suggested that new divergent pairs might have been more likely selected against through the post-WGD deletion process. Because this analysis is based on the comparison between S. cerevisiae and the pre-WGD genome, we repeated the same analysis using other genomes. We first performed comparison between S. cerevisiae and six pre-WGD species (Zygosaccharomyces rouxii, K. lactis, Ashbya gossypii, K. waltii, K. thermotolerans, and S. kluiveri). We next compared the pre-WGD genome and four post-WGD species (S. bayanus, Candida glabrata, S. castellii, and Vanderwaltozyma polysporus). In all comparisons, we obtained very similar results (not shown). We also repeated the same analysis excluding genes that still remain as duplicates. Most of these genes are ribosomal genes, which generally make random clusters and might cause a bias in our analysis. However, our result hardly changed, indicating that the result is robust to this factor.

We confirmed selection against new divergent gene pairs by a simple simulation. To model the pattern of gene loss after a WGD, we assumed that a WGD event doubles the ancestral genome with 5,000 coding genes each, and that random gene loss occurs after WGD so that the number of coding genes in the duplicated genome decreases from 10,000 to eventually 5,500. This is because the model follows the assumption that one of the duplicated copy can be pseudogenized as long as the other is functional. It was found that the behavior of the proportions of the three orientations of adjacent genes (i.e., divergent, tandem, and convergent) depended on their initial proportions (i.e., at the event of WGD) and selection.

We started a simulation with simple neutral assumptions; gene order is completely random at the initial state (such that the proportion of the divergent, tandem, and convergent orientations are 25%, 50%, and 25%, respectively). A neutral gene loss process is assumed. That is, one of the two duplicated copies is randomly removed at a constant rate until the total number of genes became 5,500, which represents the current S. cerevisiae genome. The rate of gene loss is adjusted such that the number of genes decreases to 5,500 in 10,000 generations (fig. 2A). In figure 2B, the change of the proportion of tandem gene pairs is shown by the gray dashed line and that for divergent gene pairs is shown by the dashed black line (the result of convergent gene pairs is identical to that of divergent gene pairs). The averages over 100 replications of the simulations are plotted. Under neutrality, the proportions of tandem and divergent (convergent) gene pairs stay at 50% and 25% over generations, respectively (broken lines in fig. 2B).

We next employed the proportions of the three orientations in the pre-WGD genome, which should provide a more realistic initial condition of the genome at the WGD event. It is assumed that the proportions of divergent, tandem, and convergent are 28%, 44%, and 28% (see fig. 1A), respectively. We found that the proportion of tandem orientation approaches 50%, whereas that of divergent (convergent) orientation approaches to 25% through this random gene loss process (solid line in fig. 2B). The proportions of new divergent and convergent pairs stay at 25% through the simulation. Thus, we conclude that the two neutral simulations...
and 25%, and solid lines when 44% and 28%.

when the initial proportions of tandem and divergent pairs are 50% (gray and black lines, respectively). The result is shown by broken lines of the proportions of tandem and divergent (conserved) gene pairs for divergent vs. convergent, permutation test), indicating of the total number of genes through the simulation. (B)

(A) Decrease of the total number of genes through the simulation. (B) The changes of the proportions of tandem and divergent (conserved) gene pairs (gray and black lines, respectively). The result is shown by broken lines when the initial proportions of tandem and divergent pairs are 50% and 25%, and solid lines when 44% and 28%.

cannot explain the observed reduction in the proportion of new divergent gene pairs (20.7%) without considering selection against new divergent pairs.

Target of Selection

Our comparative genomics analysis thus far demonstrated that selection against new divergent gene pairs should have worked in the post-WGD genome reorganization process. To address the question of what would be the actual target of selection, we focused on the intergenic regions, which should play a crucial role to regulate the expression of the genes nearby. We found that the average length of intergenic regions of new divergent gene pairs is generally longer than those of new tandem and convergent gene pairs. As illustrated in figure 1C, new adjacent gene pairs arose by losing genes between them. Therefore, it is predicted that the intergenic region is generally long in the initial state because of pseudogenic sequence in the new intergenic region. Then it is subject to strong pressure of deletion to keep the genome compact, and it will shrink over time. If this process works equally for the new gene pairs in three orientations, we expect that the speed of shrinkage would be similar for the three orientations. However, it seems that this does not hold in the S. cerevisiae genome as shown in table 1. New divergent pairs have on average ~300 bp longer intergenic sequences than conserved ones, whereas new tandem and convergent gene pairs have only 100-bp longer intergenic sequences. This difference is statistically significant ($P < 0.0001$ for divergent vs. tandem, $P < 0.0001$ for divergent vs. convergent, permutation test), indicating that there could be a reason to keep new divergent pairs physically apart. In this analysis, a coding gene is defined as the region between the translation initiation and termination positions, and an intergenic region is defined as the region between two adjacent coding genes: this is a commonly used definition in yeast because of a lack of transcriptome data. However, transcriptome data are increasing recently although the amount is still limited (Miura et al. 2006; Nagalakshmi et al. 2008). Therefore, we repeated the same analysis by redefining an intergenic region as between untranslated regions and confirmed that the same trend holds (table 1).

Here, we hypothesize that natural selection works to keep newly divergent gene pairs physically away because their coregulation may be deleterious and/or because it takes a long evolutionary time to reduce the intergenic region length between a new divergent pair in a short region. In either case, selection should work against deletion, so that the shrinkage process is slowed down. Then what makes deletion deleterious? It is quite straightforward to imagine that the chromatin state of intergenic region should be a key factor (Batada et al. 2007). We focused on the locations of NFRs in intergenic regions, where RNA Pol II binds and initiates transcription (Neil et al. 2009; Xu et al. 2009). It is known that at least in yeast, two adjacent genes in divergent orientation can be coexpressed when the promoter region between them has a single NFR (Xu et al. 2009). If such coexpression of a newly created divergent gene pair is disfavored, selection would work against deletions that made the intergenic region so short that it could accommodate only one NFR.

This scenario is further explained by using a very simplified model illustrated in figure 3. It is assumed that the ancestral genome (state 1) is nearly as compact as possible, so that each gene has one NFR. It is also assumed that a single NFR is shared if an adjacent gene pair are in divergent orientation. Then, there are only four patterns for the formation of a new adjacent gene pair by a single gene loss. The first and second patterns create new divergent and convergent pairs (fig. 3A and B), and the other two create new tandem pairs (fig. 3C and D). In all cases, the middle gene is lost (state 2) and DNA deletions occur to shrink the intergenic region of the new adjacent gene pair (state 3). Eventually, the intergenic region becomes as short as possible (state 4). This process should be different between (A) and the other three because deletions in case (A) can potentially force the new divergent pair to share one of the NFRs while this should not happen to the other three. As a significant amount of time has passed since the WGD, we suppose that the current genome of the post-WGD species is very close to state 4. However, our hypothesis is that case (A) is an exception because sharing one NFR by a new divergent gene pair would often be deleterious. If so, it is possible that only in case (A), the situation may be stuck or delayed in state 3, where the two genes have their own NFRs.

Our hypothesis was supported by expression data. Using microarray data, we measured the similarity in the expression pattern using the correlation coefficient, $r$. We found that the mean $r$ for conserved divergent gene pairs is much higher than those of tandem and convergent gene pairs (table 1) (this is also pointed out by a recent empirical study.
Fig. 3. Simple illustration of gene loss and shrinkage of intergenic region by DNA deletion. Under our simple assumptions (see text), there are only four possible cases from (A) to (D). In all cases, the loss of the middle gene is considered. Coding genes and NFRs are presented by open arrows and circles, respectively. When a circle is attached on an allow, it is meant that the NFR works as a promoter of the attached. Once the middle gene is lost (pseudogenized), it immediately becomes a part of the intergenic region of the new gene pair (state 2). DNA deletions make the intergenic region shorter (state 3), and eventually the intergenic region will be composed of the minimum elements including a single NFR in our simplified setting (state 4).

by Xu et al. [2009]). In addition, we found that new divergent gene pairs have on average significantly lower $r$ than conserved ones, while there is no such difference for tandem and convergent categories.

To further verify our hypothesis, we compared the number of NFRs between the new and the conserved divergent gene pairs (table 2). We first considered the cases with one and two NFRs. As expected, we found that about 80% (286/355) of conserved divergent gene pairs share a single NFR while this proportion is significantly reduced to 62% (69/111) for new divergent gene pairs ($P = 0.0001$, exact test). It is important to notice that this difference accounts for the difference in the intergenic distance and the correlation ($r$) in expression pattern between the new and conserved divergent gene pairs demonstrated in table 1. As shown in table 2, whether it is new or conserved, gene pairs with one NFR have on average higher $r$ (roughly 0.29) and shorter intergenic distances (roughly 340 bp), whereas gene pairs with two NFRs have lower $r$ (roughly 0.20) with longer intergenic distances (roughly 600 bp). Thus, it can be concluded that the observed new versus conserved differences in the intergenic distance and in $r$ are very well explained by a reduced number of new divergent gene pairs with one NFR. Such differences were not observed for tandem or convergent gene pairs. We also included the cases with more than two NFRs and obtained a very similar result (table 2).

Estimating the Intensity of Selection

Based on these observations, we developed a simulation model of the reorganization process of the yeast genome after WGD and estimated the intensity of selection against deletion in intergenic regions. Our prediction was that negative selection is stronger for new divergent gene pairs than for new tandem and convergent ones. The process involves at least two major mutational processes: pseudogenization of one of the two duplicated copies and genome-size shrinkage by deletion of DNA fragments. To simplify the model, we assume that pseudogenization occurs by a point mutation or very small insertions and deletions causing a frameshift (this event itself has little effect on the genome size). A pseudogenized gene and its regulatory regions then become less important, which will be a target of DNA deletion. By using this model, we inferred the intensity of selection against DNA deletion.

The selection intensity is parameterized by introducing a function $f$, which describes the fitness effect of a deletion.

### Table 2. Relationship Between the Number of NFRs on the Coexpression and Intergenic Distance.

<table>
<thead>
<tr>
<th></th>
<th>One NFR</th>
<th>Two NFRs</th>
<th>≥ Two NFRs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of gene pairs</td>
<td>Coexpression ($r$)</td>
<td>Intergenic distance</td>
</tr>
<tr>
<td>Divergent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conserved</td>
<td>286</td>
<td>0.280</td>
<td>339.46</td>
</tr>
<tr>
<td>New</td>
<td>69</td>
<td>0.295</td>
<td>345.04</td>
</tr>
<tr>
<td>Tandem</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conserved</td>
<td>411</td>
<td>0.167</td>
<td>363.42</td>
</tr>
<tr>
<td>New</td>
<td>279</td>
<td>0.180</td>
<td>346.33</td>
</tr>
<tr>
<td>Convergent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conserved</td>
<td>446</td>
<td>0.211</td>
<td>216.08</td>
</tr>
<tr>
<td>New</td>
<td>232</td>
<td>0.216</td>
<td>262.47</td>
</tr>
</tbody>
</table>
proximate likelihood (acceptance rate) for \( g_D \) is similar to \( L \). Fitness increases nearly linearly when the intergenic length increases as the length of the intergenic region decreases. That the intensity of selection increases (hence, the fitness is large, the fitness dramatically decreases. When \( g \) is large, the fitness dramatically decreases as the intergenic length approaches \( L_{\text{min}} \). For a small \( g \), the fitness increases nearly linearly when the intergenic length is similar to \( L_{\text{min}} \) and saturates at 1. Under this model, we developed a simulation-based approximate likelihood algorithm to estimate the selection parameter \( g \) for the three orientations, \( g_D \), \( g_T \), and \( g_C \).

We have hypothesized that deletion is more disadvantageous for divergent pairs than the others, that is, \( g_D < g_T \) and \( g_C \). To verify this hypothesis, we first estimated \( g_T \) and \( g_C \). Then, we tested whether \( g_D \) is smaller than \( g_T \) and \( g_C \). Figure 4A shows the profiled likelihood for \( \{g_T, g_C\} \), from which we obtained estimates \( \{g_T, g_C\} = \{0.11, 0.16\} \).

Next, given these estimates, we examined whether \( g_D \) is significantly larger than \( g_T \) and \( g_C \). Our rejection-sampling method found that \( g_D \) roughly distributes from \( \sim 0.03 \) to 0.06, which is significantly smaller than the estimates of \( g_D \) and \( g_C \) (see also fig. 4B). This suggests that selection against DNA deletion is particularly strong for divergent gene pairs, as we suspected.

**Discussion**

A genome locates its coding genes on the chromosomes in a nonrandom manner, but the mechanism of gene order evolution is poorly understood. How is natural selection involved in the evolution of gene order? To address this question, we focused on the evolutionary changes of gene order in yeasts with special attention to pairs of adjacent genes. There are many lines of empirical evidence that adjacent genes (especially in divergent orientation) in yeasts can be coexpressed and coregulated; therefore, one can imagine that cis-regulatory elements would be one of the major factors to determine the order of coding genes, and that selection should work particularly when a new pair of genes in divergent orientation is formed.

The first finding of our comparative genomic analysis was that the proportion of newly arisen divergent gene pairs is significantly reduced in comparison with new gene pairs in the other two orientations (table 1). In addition, we found that new divergent gene pairs had significantly longer intergenic regions than the other two. For all post-WGD species, the DNA of intergenic regions has been under strong selective pressure to be compacted by deletion. Although this applies to intergenic regions of newly formed gene pairs in all three orientations, it seems that the rate of shrinkage is particularly slow for new divergent gene pairs (fig. 4 and table 1). From these observations, we concluded that newly arisen divergent gene pairs are generally disfavored most likely because their coexpression and/or coregulation may be deleterious. Accordingly, when such new adjacent pairs arose in the population, they usually did not become fixed immediately. Once fixation occurred, the shrinkage of intergenic regions was slowed down, perhaps because selection worked against deletion to keep them physically separate, so that they would be less likely coexpressed and/or coregulated. Our further analyses of the locations of NFRs supported our conclusion. By using simulations, we demonstrated that very strong selection against deletion has worked in the intergenic regions of new divergent gene pairs (fig. 4). Disadvantage of closely located genes have been suggested by Byrnes et al. (2006) and Liao and Zhang (2008).

Once beneficial divergent pairs are formed, it is expected that selection should work to maintain them, as supported by earlier genome analyses (Hurst et al. 2002; Kensche et al. 2008). Hurst et al. (2002) found that there is a trend that adjacent gene pairs that are conserved between S. cerevisiae and C. albicans have higher correlation (\( r \) in the expression pattern and (Kensche et al. 2008) confirmed this by using additional genome sequences. Hurst et al. (2002) further found that conserved gene pairs have significantly shorter intergenic regions, and multivariate analysis using logistic regression of Poyatos and Hurst (2007) found that the
distance of intergenic region is highly correlated with gene pair conservation. Recently, Hermens et al. (2008) reported a strange bimodal distribution of the intergenic regions of adjacent genes in divergent orientations. Thus, there have been several lines of indirect evidence that cis-regulatory elements in the intergenic regions play a crucial role in the evolution of gene order. Consistent with these studies, we showed that the physical locations of NFRs could be potential targets of selection, suggesting that gene regulation would be one of the major factors to determine the order of coding genes.

Acknowledgments

The authors thank Jeff Fawcett and two anonymous reviewers for helpful comments. This work was in part supported by grants from the Japan Society for the Promotion of Science (JSPS), Japan Science and Technology Agency (JST), and the Graduate University for Advanced Studies to H.I. RPS was also supported by JSPS.

References


