Genome evolution after whole genome duplication in the baker's yeast.

Ryuichi Sugino

Doctor of Philosophy

Department of Evolutionary Studies of Biosystems

School of Advanced Sciences

The graduate university for advanced studies

2011

Genome evolution after whole genome duplication in the baker's yeast.

Ryuichi Sugino

Doctor of Philosophy

Department of Evolutionary Studies of Biosystems

School of Advanced Sciences

The graduate university for advanced studies

2011

Acknowledgement

First, I would like to express my appreciation to my superviser Dr. Hideki Innan. He invited me to this interesting scientist field, and gave me lectures about a broad range of topics. He established an active and competitive lab in Sokendai. Discussion in the lab seminar broaden my knowledge and stimulate my insight. I have been supported by the Kyoto colleagues in their friendship; Showhey, Jeff and Yasuo.

Next, I would like to thanks to Department of evolutionary studies of biosystems and Sokendai. This thesis would not have been possible unless those people. Faculty members gave many advices in progress reports. Especially, Dr. Akira Sasaki and Dr. Tastuya Ohta are my committee and Dr. Koji Hirata is superviser for my subthesis. The PhD students made my Hayama life very fun. The administrators always helped my defective paperworks.

My work was sophisticated by specialists; Ken Wolfe and his lab members, Fyodor Kondrashov, Jake Byrnes and Hiroshi Akashi and anonymous reviewers of the published papers. It was an exciting experience for me to visit Ken Wolfe Lab in Trinity college Dublin. Hiroshi gave me important comments in my difference.

Finally I have to say thanks to my family for their support and assistance. My work is supported by Japan Society for the moprmoting science (DC1). My visiting to Dublin was supported by the short-stay study abroad program in FY 2009, Sokendai.

Summary

My research interest is to understand the mechanisms of evolution on a genomic scale. Recent advances of genome sequencing technology and genome-wide experimental technology provide an excellent opportunity of studies of genome evolution. In my PhD work, using the bakers' yeast *Saccharomyces cerevisiae* as a model, I studied genome evolution after a whole-genome duplication (WGD) event. All genes were doubled at the WGD event, but only $\sim 10\ \%$ of them remain as duplicates (called ohnologs) at present and other genes have lost one of the duplicated pairs. In addition, massive genome rearrangement have occurred in this gene deletion process.

This thesis was constructed by three parts. In the first part, I studied the evolution of ohnologs. Interlocus gene conversion is a unique recombinational mechanism to duplicated genes. Because it retards the nucleotide divergence of duplicates, the standard molecular clock model can not be directly applied to infer the history of duplicates. In this chapter, a maximum likelihood method to estimate the time of the WGD was developed incorporating the effect of gene conversion. It was estimated that the WGD is almost as old as the speciation event with pre-WGD species. It is suggested that the WGD might have caused the speciation.

In the next part, I examined the role of natural selection to the duration of concerted evolution. It was found that duplicated with higher expression (especially ribosome and histone genes) prefers long-term concerted evolution, indicating gene conversion may be favored for such high-demand genes. By genome-wide data analysis with various kinds of experimental data, I found this hypothesis is a likely explanation of the observation.

In the third part, I studied the evolution of gene order in the genome rearrangement process after the WGD. In the analysis I focused on adjacent gene pairs. Comparative genome analysis indicated that newly generated adjacent gene pairs in divergent orientation are relatively rare and they have on average long intergenic distances and low coexpression. I considered that the locations of nucleosome free regions (NFRs) would explain this. It is known that transcription

start in both directions when Pol II binds to a NFR. It is predicted that such co-expression would be deleterious for a random pair of genes that happened to be adjacent to each other. If so, selection should have worked against deletion between newly created divergent gene pairs, thereby keeping them physically away so that their coexpression might be avoided. I verified this hypothesis by comparative genomic analysis of the locations of NFRs and evolutionary simulations.

Through these works, I conclude that the genome of *S. cerevisiae* undergo various types of genome-wide natural selection through the process after the WGD. This study also shows that the post-genomic biological data are useful to determine the target of natural selection.

Contents

List of Publication

- **Sugino, R.P.** and Innan, H. (2005)
  Estimating the time to the whole-genome duplication and the duration of concerted evolution via gene conversion in yeast.
  Genetics, **171**: 63-69.

- **Sugino, R.P.** and Innan, H. (2006)
  Selection for more of the same product as a force to enhance concerted evolution of duplicated genes.
  Trends in Genetics, **22**: 642-644.

- Than, C., **Sugino, R.**, Innan, H. and Nakhleh, L. (2008)
  Efficient inference of bacterial strain trees from genome-scale multilocus data.
  Bioinformatics, **24**: i123-131.

- Kitao, H., Nanda, I., **Sugino, R.P.**, Kinomura, A., Yamazoe, M., Arakawa, H., Schmid, M., Innan, H., Hiom, K. and Takata, M. (2011)
  FancJ/Brip1 helicase protects against genomic losses and gains in vertebrate cells.
  Genes to Cells, **16**: 714-727.

- **Sugino, R.P.** and Innan, H. (2011)
  Natural selection on gene order in the genome re-organization process after whole genome duplication of yeast
  Molecular biology and evolution, **22**: *in press*.

1

Introduction

1.1 Genome sequencing and post-genome studies

Evolution is the heritable change of organisms' trait. The studies of genome evolution aims to figure out the system of evolution from the change of individuals' or species' genome. An example of genome evolution on a gene scale is shown in figure 1.1. There are a variety of changes which occur in a genome. Point mutation potentially changes the character of protein or dosage of its product. Gene deletion remove a gene from the genome. Gene duplication produces a copy of itself. Gene conversion, rewrites the gene like copy-and-paste, often occurred between homologous genes (*e.g.* duplicated genes). Whole genome duplication is gene duplication on the whole genome scale.

The word, "genome", means the total genetic information of individuals. This word is produced by synthesizing "gene" and "-ome" (meaning total). A genome consists of four type of DNA, which are often symbolized by "A", "T", "C" and "G". It is considered that, if all genetic information are available, we can know everything about individuals. This is the motivation of the Human genome project (HGP). HGP is the largest paradigm shift in all fields of biology. Through the process of HGP, the genome of many different species were revealed, including *Escherichia coli*, the baker's yeast *Saccharomyces cerevisiae*, nematode, fruit fly and *Arabidopsis thaliana*. These species are often considered as "model species", and most post-genome analysis have been focused on these species.

1.2 Yeast is the best species for studying molecular evolution

The baker's yeast, *S. cerevisiae*, is a model species for eukaryotes. Its genome size is $\sim$ 12 Mb, containing 5,600 well annotated genes. The wild type of this species is prevalent in all over the world. This species is often found in oak trees. For example, the lab strain S288c was sampled from an oak tree in California (Mortimer and Johnston 1986). The most important trait is its fermentation ability. The strains of wine and sake were independently discovered and utilized in Europe

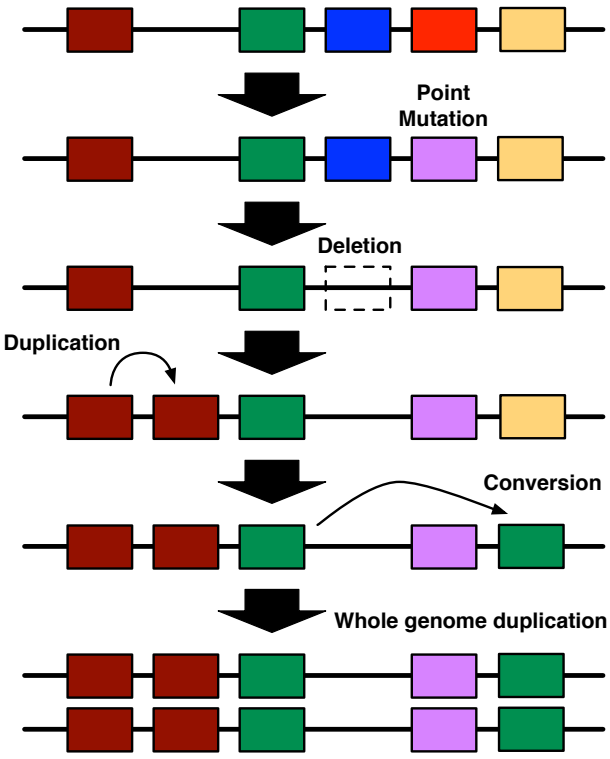


Figure 1.1: Schematic of genome evolution on a gene scale.

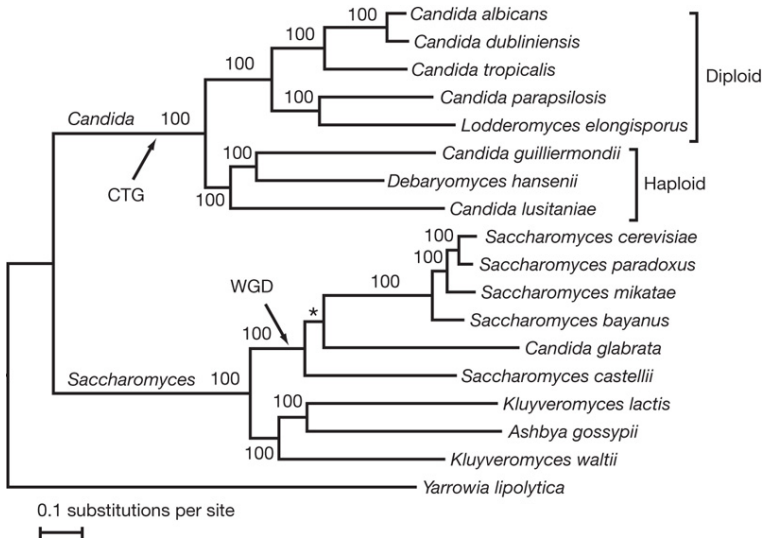


Figure 1.2: Phylogeny of the budding yeasts. Reprint from Butler et al. (2009).

and Japan. Furthermore, the ale beer is produced by this yeast. On the other hand, lager beer is fermented by an other yeast, *S. pastorianus*, which was generated by the hybridization of *S. cerevisiae* and *S. bayanus* (Dunn and Sherlock 2008). Contrary to these benefits, some strains cause infectious diseases, like *Candida* (Wei et al. 2007).

S. cerevisiae is one of the most suitable species for molecular evolution analysis for two reasons. The first is the relatively large number of genome sequences available for related species. *S. cerevisiae* was the first eukaryote to have its genome sequence determined (Goffeau et al. 1996). Because of their compactness, the genome of many related species was determined in earlier stages of the post-genome era (Scannell, Butler, and Wolfe 2007, reviewed). Figure 1.2 shows the phylogeny of the budding yeasts. Some species are deeply involved in human's life. *Candida albicans* is famous for causing infectious diseases. *Debaryomyces hansenii* is needed for cheese fermentation.

Second, there are a variety of experimental data, which have been gathered on the whole genome scale. In this post-genome era, such experiments are the major

trend of molecular biology. They reveal regulators of gene expression, chromosomal structure in nucleus, complex network of gene products, and so on. These data also allow us to survey the target of natural selection.

Furthermore, *S. cerevisiae* has experienced a unique evolutionary event, whole genome duplication (WGD). WGD is gene duplication on the whole genome scale. The WGD of yeast was identified (Kellis, Birren, and Lander 2004, Wolfe and Shields 1997). Due to the WGD, all genes became two copies. Subsequently genome reorganization with massive gene deletion occurred. In the case of *S. cerevisiae*, the 10,000 genes that were present immediately after the WGD was reduced to 5,500. Only 450 genes have remained as WGD derived duplicates (Ohnolog). Furthermore, due to the process of genome rearrangement most relative gene relationship was changed. The current 16 chromosomes of *S. cerevisiae* were corresponded to the 55 blocks from 8 chromosomes of *K. waltii*. It indicates that massive genome rearrangement have occurred. The WGD is considered a good material to study genome evolution because it generated a large number of duplicated genes simultaneously and massive genome rearrangement shuffled gene order.

Here, I studied three topics to answer above question. In Chapter 2, I estimated the duration of concerted evolution via gene conversion. Concerted evolution is a homogenizing process, which was frequently found in duplicated genes (Ohta 1980). Gene conversion is a mechanism of concerted evolution. Under concerted evolution the divergence of duplicated genes are suppressed until its termination. The large problem in the study of the evolution of duplicated genes is the disruption of molecular clock by concerted evolution. However, ohnologs outcome this problem because they were generated simultaneously. Using ohnologs, I estimated the degree of concerted evolution.

In chapter 3, I studied the effect of natural selection for ohnologs. In the previous study, I found that yeast' ohnologs didn't follow the neutral evolution model. What have caused it? Using genome-wide experimental data, I concluded that natural selection for increasing the dosage of gene product (or "more of the same products" by Ohno (1970)) works for maintaining the homology between

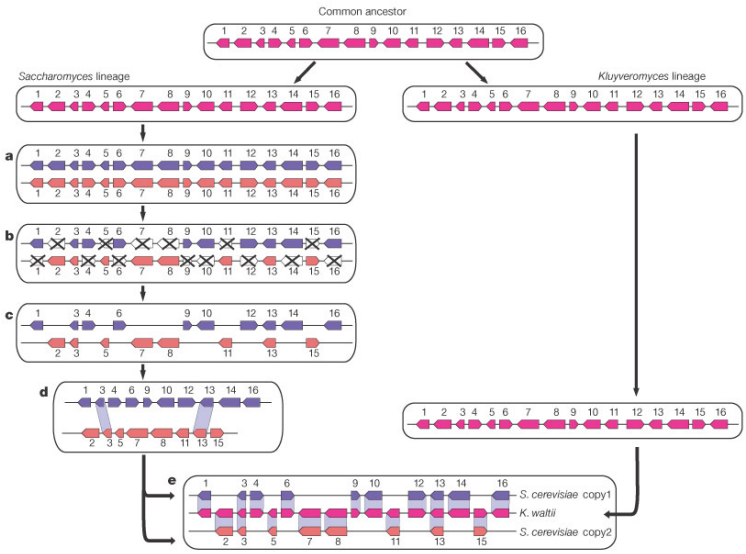


Figure 1.3: Model of whole genome duplication. Reprint from Kellis, Birren, and Lander (2004).

ohnologs.

In chapter 4, I focused on the process of genome rearrangement following WGD. Such research which considers the effect of gene location on the chromosome, is often called the study of "gene order" (Hurst, Pál, and Lercher 2004). Gene order is considered as the target of natural selection. The genome rearrangement following WGD would be a good chance to implement an optimal gene order. By comparative genomics approach, I found that natural selection also works for gene order to keep away the adjacent (or neighbor) gene pair when their expression interferes with each other.

2

Estimating the Time to the Whole-Genome Duplication and the Duration of Concerted Evolution via Gene Conversion in Yeast

2.1 Abstract

A maximum-likelihood (ML) method is developed to estimate the duration of concerted evolution and the time to the whole-genome duplication (WGD) event in bakers yeast (*Saccharomyces cerevisiae*). The models with concerted evolution fit the data significantly better than the standard molecular clock model, indicating a crucial role of concerted evolution via gene conversion after gene duplication in yeast. Our ML estimate of the time to the WGD is nearly identical to the time to the speciation event between *S. cerevisiae* and *Kluyveromyces waltii*, suggesting that the WGD occurred in very early stages after speciation or the WGD might have been involved in the speciation event.

2.2 Introduction

Nonindependeny evolution of a multigene family is called concerted evolution (Arnheim 1983, Ohta 1980, Zimmer et al. 1980). The nucleotide divergences among copy members are likely very low during concerted evolution. Interlocus gene conversion has been thought to be the most important mechanism for the homogenization of genetic variation between duplicated genes (or small multigene families), although unequal crossing over should play a significant role in middle-size to large multigene families (reviewed in Li (1997), Ohta (1980)). Many duplicated genes in various species exhibit clear evidence for gene conversion (see Innan (2003b) and references therein), but a number of unresolved questions remain. For example, How long does concerted evolution last? How often does it occur? What is the evolutionary significance? Very little information is available to answer these questions (Gao and Innan 2004, Teshima and Innan 2004).

With concerted evolution, the behavior of the level of divergence between duplicated genes (d) does not follow the standard molecular clock model (Zuckerkandl and Pauling 1965). Teshima and Innan (2004) demonstrated that the process has three phases [see Teshima and Innan (2004)'s Figure 4]. Phase I is the time until d reaches its equilibrium value, d_0. In phase II d fluctuates around d_0,

and d increases again in phase III. Phase I and II represents the time of concerted evolution. The termination of concerted evolution occurs by either mutation or selection. Since interlocus gene conversion results from a nonreciprocal recombination between paralogous regions, the rate of gene conversion may have a positive correlation with the possibility of the pairing of the paralogous regions during meiosis. Large-size insertions or deletions may terminate concerted evolution because they might work as a barrier against the pairing of paralogs. The accumulation of point mutations could also have a similar effect (Teshima and Innan 2004, Walsh 1987) if the divergence between the paralogous regions suppresses gene conversion. Thus, the duration of concerted evolution depends primarily on the mutation and gene conversion rates, although other factors including the tract length of gene conversion also play important roles (Teshima and Innan 2004).

Additionally, selection could also work as a mechanism to terminate concerted evolution. Suppose that a new mutation with a novel function is fixed in one of the duplicated genes while the other keeps the original function (*i.e.*, neofunctionalization). If the state where the two copies have different functions is favored, this state can be maintained by strong selection even under the pressure of homogenization by gene conversion (Innan 2003a). An interesting example is seen in the RHD and RHCE loci in humans. Clear evidence for frequent gene conversion is observed in most of the coding regions of this pair of genes, and the divergence between them is low. On the other hand, $\sim$10 non-synonymous nucleotide differences (and a few synonymous ones) are fixed in exon 7 of the two genes, thereby creating a high peak of divergence. It is hypothesized that strong positive selection is operating to keep the amino acid differences in exon 7, and the termination of the concerted evolution might be about to occur in this region (Innan 2003a). The time of concerted evolution can be considered as the waiting time for a termination event by either selection or neutral mutations; therefore the time length could be approximated by an exponential distribution,

$$f(t) = \frac{1}{\tau} \exp(-t/\tau) \tag{2.1}$$

(Teshima and Innan 2004), where τ is the expected length of concerted evolution.

This article utilizes this equation to estimate the duration of concerted evolution on a genomic scale. Bakers yeast, *S. cerevisiae*, is used as a model species to take advantage of the fact that the yeast genome has experienced a whole-genome duplication (WGD) (Dietrich et al. 2004, Kellis, Birren, and Lander 2004, Wolfe and Shields 1997). Recently, Kellis, Birren, and Lander (2004) reported the genome sequence of *Kluyveromyces waltii*, which has diverged from the ancestral lineage of *S. cerevisiae* before the WGD event. They mapped two regions of *S. cerevisiae* to *K. waltii* genome. Although one copy of most duplicated gene pairs is lost after the WGD, the present *S. cerevisiae* genome has at least $\sim$ 450 pairs of genes originating from the WGD (Kellis, Birren, and Lander 2004). The DNA sequence data of these pairs from the WGD are used to estimate t together with the time to the WGD event.

2.3 Model and theory

Consider two species, I and II. Suppose that species II has experienced a gene duplication event after the speciation with species I. The three genes, one in species I and two in species II, are denoted by X, Y, and Z, respectively, as illustrated in figure 2.1A. Let T be the time to the speciation event (represented by S in figure 2.1), and R be the time to the duplication event in units of $2T$. Without concerted evolution, the divergence between the two paralogs of species II reflects the time to the duplication and the gene tree should be similar to figure 2.1A. In other words, the time to the most recent common ancestor (MRCA) of the paralogs is R. However, if the duplicated pair have undergone concerted evolution, their divergence is expected to be smaller than the prediction under the molecular clock model as illustrated in figure 2.1B and C. M represents the MRCA of the duplicates, and t is the time of concerted evolution (in units of $2T$), which is between the duplication event and M. The time length between M and present, represented by r (in units of $2T$), contributes to the nucleotide divergence between Y and Z. In figure 2.1B, concerted evolution is terminated some time ago, so that Y and Z have a relatively long divergence time. Figure 2.1C illustrates a case where concerted evolution is

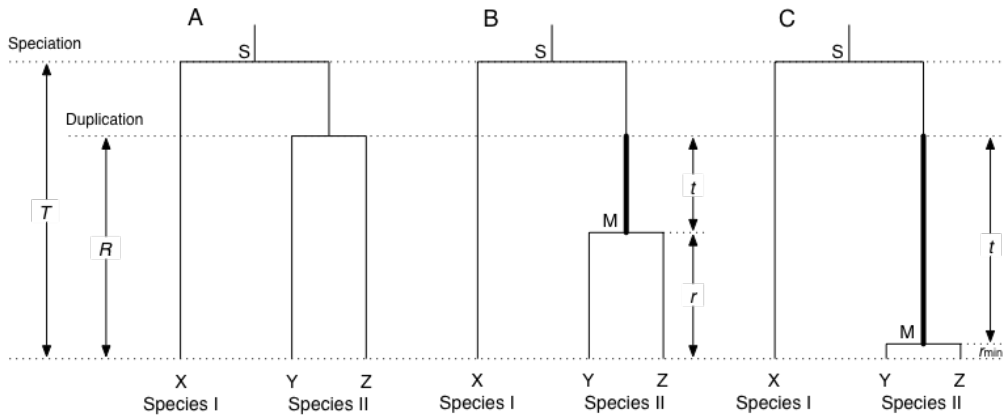


Figure 2.1: Illustration of gene trees after gene duplication. Thick lines represent the time of concerted evolution. See text for details.

ongoing. Note that, in this case, r may not be zero because the sequences of Y and Z are not always identical under concerted evolution. r_{min} represents the time to MRCA when Y and Z are under concerted evolution, which is mainly determined by the gene conversion rate (Innan 2002, 2003b, Ohta 1982).

The evolutionary history of the three genes, X, Y, and Z, is summarized by a simple relationship as shown in Figure 2.2, regardless of how long concerted evolution continues. Focus on a particular nucleotide site, at which x, y, and z represent the nucleotides at the site on X, Y, and Z, respectively. Mutations occur at a constant rate m per site. A simple two-allele model is considered first. Let 0 be the nucleotide at M, say "G," and 1 be the other three nucleotides ("A," "T," and "C"). Under the Jukes-Cantor model (Jukes and Cantor 1969), the probability that $x = 0$ is

$$p_1 = \frac{1 + 3\exp[-8\mu T(1-r)/3]}{4}. \tag{2.2}$$

Likewise, the probability that $y = 0$ is given by

$$p_2 = \frac{1 + 3\exp[-8\mu Tr/3]}{4}, \tag{2.3}$$

which is identical to the probability that $z = 0$. Then, it is straightforward to

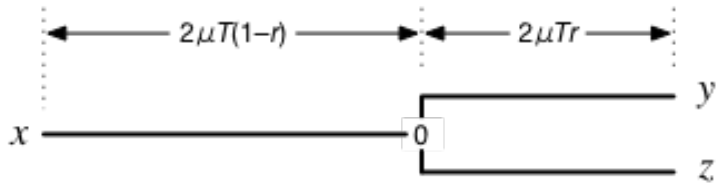


Figure 2.2: The evolutionary relationship among three homologous sites, x, y and z.

obtain the joint probability for x, y, and z as summarized in table 1. There are eight possible allelic states, (000), (001), (010), (100), (011), (101), (110), and (111), where the three numbers represent x, y, and z. For example, the probability that $x = y = z = 0$ is $P_{000} = p_1 p_2^2$, where the subscript of P represents the allelic state.

The model is extended to a four-allele model, in which x, y, and z could be one of the four alleles, 'A', 'T', 'G', and 'C'. Let P_{AAA} be the probability that $x = y = z$. P_{AAA} is given by $P_{000} + \frac{1}{9}P_{111}$ because the three nucleotides can be the same with probability $1/9$ when $x = y = z = 1$. In a similar way, we have the probabilities for the other four states, P_{BAA}, P_{ABA}, P_{AAB}, and P_{ABC} as shown in table 2.1.

Suppose that there are L nucleotides in a focal gene, and let l_{AAA}, l_{BAA}, l_{ABA}, l_{AAB}, and l_{ABC} be the number of nucleotides of the five allelic states. When P_{BAA}, P_{ABA}, P_{AAB}, and $P_{\text{ABC}} \ll 1$, the joint probability of l_{AAA}, l_{BAA}, l_{ABA}, l_{AAB}, and l_{ABC} is given by a function of r and $m = 2\mu T$:

$$Prob(\delta|r,m) = Q(l_{\text{BAA}}, P_{\text{BAA}}L)Q(l_{\text{ABA}}, P_{\text{ABA}}L) \\ Q(l_{\text{AAB}}, P_{\text{AAB}}L)Q(l_{\text{ABC}}, P_{\text{ABC}}L), \quad (2.4)$$

where $\delta = (l_{\text{AAA}}, l_{\text{BAA}}, l_{\text{ABA}}, l_{\text{AAB}}, l_{\text{ABC}})$ and $Q(l,s)$ is the Poisson probability to observe l when its expectation is s:

$$Q(l,s) = \frac{s^l}{e^s l!}. \quad (2.5)$$

Table 2.1: Probabilities of Allelic States

Allelic state	Probability
Two-allele model	
000	$p_1 p_2^2$
001 and 010	$p_1 p_2 (1-p_2)$
100	$(1-p_1) p_2^2$
101 and 110	$(1-p_1) p_2 (1-p_2)$
011	$p_1 (1-p_2)^2$
111	$(1-p_1)(1-p_2)^2$
Four-allele model	
AAA ($x = y = z$)	$P_{000} + \frac{1}{9} P_{111}$
BAA ($x \neq y = z$)	$P_{100} + \frac{1}{3} P_{011} + \frac{2}{9} P_{111}$
ABA ($x = z \neq y$)	$P_{010} + \frac{1}{3} P_{101} + \frac{2}{9} P_{111}$
AAB ($x = y \neq z$)	$P_{001} + \frac{1}{3} P_{110} + \frac{2}{9} P_{111}$
ABC ($x \neq y \neq z$)	$\frac{2}{3}(P_{011} + P_{101} + P_{110}) + \frac{2}{9} P_{111}$

This approximation works well because we use conserved regions such that the proportion of variable sites is about 10% (see below).

Although (2.4) involves the mutation rate (m) that is unknown, it is possible to estimate m from the divergence between (X and Y) or (X and Z). Let d_y and d_z be the numbers of nucleotide differences between (X and Y) and (X and Z), respectively. A point estimate of m is easily obtained by the Jukes-Cantor equation:

$$\hat{m} = -\frac{3}{4} \ln \left(1 - \frac{4}{3} \frac{d_y + d_z}{2L} \right). \tag{2.6}$$

It is also possible to obtain the mutation rate as a probability density distribution, which is given by

$$G(m) = \frac{Prob(d_y, d_z | m)}{\int_0^\infty Prob(d_y, d_z | m)\, dm}, \tag{2.7}$$

where

$$Prob(d_y, d_z|m) \approx Q(d_y|\bar{d}L)Q(d_z|\bar{d}L), \tag{2.8}$$

and

$$\bar{d} = \frac{1 + 3\exp(-4m/3)}{4}. \tag{2.9}$$

Then, the unconditional probability of δ given r can be obtained from (2.4) by replacing m with a point estimate given by (2.6), or by averaging $Prob(\delta|r, m)$ weighted by $G(m)$:

$$Prob(\delta|r) = \int_0^\infty G(m)Prob(\delta|r, m)dm. \tag{2.10}$$

Eq. 2.10 is used in the following analysis although almost identical results are obtained by (2.4) with a point estimate of m from (2.6).

2.4 Maximum likelihood

Data: Using Equation 2.10, we develop a maximum-likelihood (ML) method to estimate the time to the WGD and the duration of concerted evolution in yeast. We use the DNA sequence data for the ∼450 pairs of genes from the WGD in *S. cerevisiae* plus their orthologs in *K. waltii* (Kellis, Birren, and Lander 2004). The aligned sequences of the 450 trios were downloaded from http:// www. nature. com/ nature/ journal/ v428/ n6983/ extref/ nature02424-s1.htm, and well-aligned regions were extracted (*i.e.*, .90% identity at the first and second positions of the codon). Third positions are not used because the speciation event is so old that nucleotide substitutions at the third positions are almost saturated. The advantage of using the first and second positions is that the effect of multiple mutations at a single site is small, because the first and second positions are more conserved . At the first position, ∼95% of nucleotide changes result in amino acid changes and 100% for the second position. For each of the trios, we count the numbers of the five types of sites, $\delta = (l_{\text{AAA}}, l_{\text{BAA}}, l_{\text{ABA}}, l_{\text{AAB}}, l_{\text{ABC}})$, at the first and second

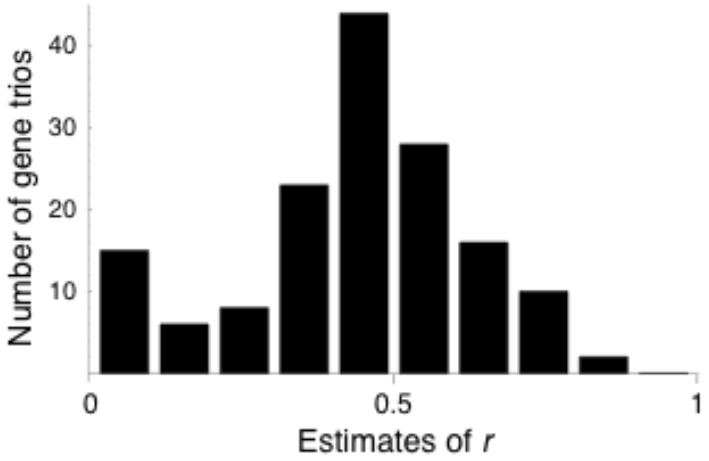


Figure 2.3: The distribution of estimates of r from 152 gene trios for which $l_{\mathrm{BAA}} + (l_{\mathrm{ABA}} + l_{\mathrm{AAB}})/2 \geq 20$.

positions of the codon in the well-aligned regions. In the following analysis, we use the data of $n = 329$ trios, for which 50 bp of the well-aligned regions (*i.e.*, 25 codons) are available. For each trio, r is roughly estimated as $\frac{(l_{\mathrm{ABA}}+l_{\mathrm{AAB}})/2}{l_{\mathrm{BAA}}+(l_{\mathrm{ABA}}+l_{\mathrm{AAB}})/2}$, and the distribution is shown in figure 2.3. Although the major peak is $\sim$0.5, there is another peak for very low r, which might reflect genes that have experienced extensive concerted evolution.

The drawback in using the first and second positions of the codon is that they are sensitive to selective pressure, which varies across genes. However, this variation may not cause a serious bias in the theory described above because R is estimated on the basis of the ratio of the divergence from M to X to that from M to Y and Z. In other words, the variation in the substitution rates among genes is allowed (see Eq. 2.7).

If we assume a constant rate of substitution over time, R can be between r_{min} and 0.5. However, if the selective pressure is relaxed after gene duplication (Lynch and Conery 2000, Ohno 1970), the substitution rate may be higher on the lineage leading to species II than that to species I. If so, R could exceed 0.5. We examine

this possibility using the *Debaryomyces hansenii* genome (Lépingle et al. 2000) as an outgroup of *S. cerevisiae* and *K. waltii*. For each of the analyzed trios, their orthologous gene in *D. hansenii* is identified by BLAST (Altschul et al. 1997). The four amino acid sequences are aligned by CLUSTALW (Thompson, Higgins, and Gibson 1994), and reverse transcribed into nucleotide sequences. Then, the substitution rates from S to X and from S to Y and Z are estimated from well aligned regions. Because the two estimates are roughly the same, we find no evidence for such acceleration of the substitution rate on the lineage leading to Y and Z (see DISCUSSION). Therefore, in the following maximum likelihood analysis, we investigated R up to 0.5, unless otherwise noted. It is also possible that the acceleration of substitution rate occurs on one of the duplicated copy, for example, under the scenario of neofunctionalization (Ohno 1970). This problem will also be discussed in DISCUSSION.

Model I: First, we consider a model with no concerted evolution as a null model. The evolutionary relationship for all trios follows figure 2.1A. Under this model, it is straightforward to obtain an ML estimate of the time to the WGD, R. The log likelihood of the data given R is given by

$$LL_1(R) = \sum_{i=1}^{n} \ln Prob(\delta_i|R), \tag{2.11}$$

where $Prob(\delta|R)$ is from (2.10).

Model II: Model II allows concerted evolution. The duration of concerted evolution is approximated by an exponential distribution with mean τ (see Eq. 2.1). τ is assumed to be constant for all duplicated genes. Under this model, the probability density distribution of r is given by

$$F(r) = \begin{cases} f(R-r) & \text{when } r_{\text{min}} < r \leq R \\ \int_{R-r_{\text{min}}}^{\infty} f(t)dt & \text{when } r = r_{\text{min}} \end{cases}. \tag{2.12}$$

Then, the probability to observe δ is given by a function of R and τ:

$$Prob(\delta|R,\tau) = \int_{r_{\text{min}}}^{R} F(r)Prob(\delta|r)dr, \tag{2.13}$$

and the log likelihood of the data is given by

$$LL_2(R,\tau) = \sum_{i=1}^{n} \ln Prob(\delta_i|R,\tau). \tag{2.14}$$

Model III: This model relaxes the assumption of a constant τ for all genes. It is assumed that τ follows a Gamma function with mean = τ_{ave} and $SD = k\tau_{\text{ave}}$, which is denoted by $\Gamma(\tau|\tau_{\text{ave}}, k)$. Then, the probability to observe δ is given by a function of R, τ_{ave} and k:

$$Prob(\delta|R,\tau_{\text{ave}},k) = \int_0^\infty \Gamma(\tau|\tau_{\text{ave}},k) Prob(\delta|R,\tau) d\tau, \tag{2.15}$$

and the log likelihood of the data given R, τ_{ave} and k is

$$LL_3(R,\tau_{\text{ave}},k) = \sum_{i=1}^{n} \ln Prob(\delta_i|R,\tau_{\text{ave}},k). \tag{2.16}$$

2.5 Results

Using the data from 329 trios, the maximum likelihood analysis is performed. We assume r_{min} is known. r_{min} represents the time to the most recent common ancestor of the duplicated genes when they are under concerted evolution, therefore r_{min} is very small. We assume $r_{min} = 0.01$ in the following analysis, but the effect of this assumption is negligible. Almost identical results are obtained for $r_{min} = 0.002$ (results not shown). We calculate the likelihood numerically under the three models, I, II, and III. Numerical calculation of likelihood is carried out for R and τ (τ_{ave}) with intervals 0.002 and 0.01, respectively. For k, the likelihood is calculated with an interval of 0.01 when $k < 0.1$ and with an interval of 0.1 when $k \geq 0.1$.

Model I

The time to the WGD (R) is estimated without concerted evolution. Figure 2.4 shows the log likelihood curve as a function of R. We obtain the maximum likelihood estimate of $R = 0.428$ (95% C.I. = $0.420 - 0.436$) with the maximum log likelihood $MLL_1 = -4641.01$.

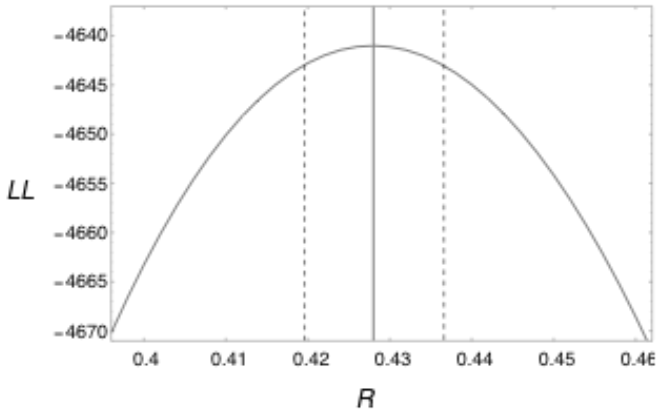


Figure 2.4: The log likelihood curve as a function of R under model I. The maximum likelihood estimate of R is represented by the vertical slid line. The 95% C.I. is represented by the two broken lines.

Model II

The time to the WGD (R) and the duration of concerted evolution (τ) are simultaneously estimated under the model with concerted evolution. When the rate of substitution is constant over time, R should be a variable between r_{min} and 0.5. Under this assumption, we have the maximum likelihood estimate of $R = 0.5$ (95% C.I. = $0.498 - 0.5$) with $\hat{\tau} = 0.12$ (95% C.I. = $0.10 - 0.13$). The maximum log likelihood is $MLL_2 = -3934.82$, which is significantly larger than MLL_1 (likelihood ratio test: $P \approx 0$), indicating that model II with concerted evolution provides a much better explanation of the observation than model I. It is suggested that concerted evolution via gene conversion plays a crucial role after genome duplication in yeast.

The assumption of a constant rate of nucleotide substitution rate over time may not hold if the selective pressure is relaxed shortly after gene duplication (Lynch and Conery 2000, Ohno 1970). Although this may not be the case for our data, the assumption can be easily relaxed by investigating the likelihood up to R_{max}. For example, if $R_{max} = 0.6$ is set, we find that maximum log likelihood

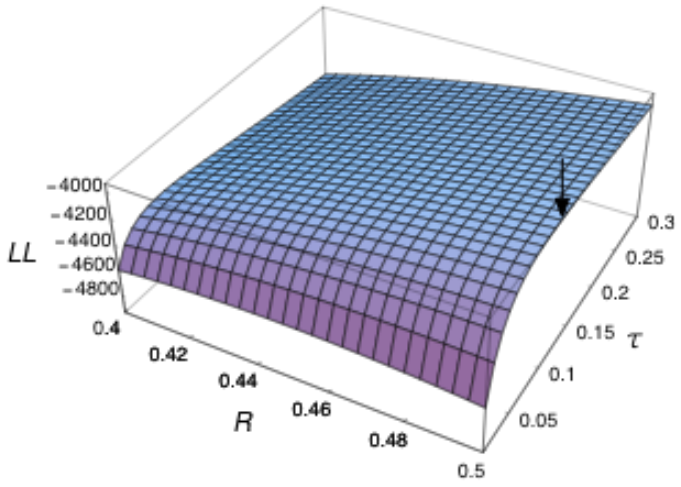


Figure 2.5: The log likelihood surface as a function of R and τ under model II. The maximum likelihood estimate is shown by the arrow.

$MLL_2 = -3786.62$ is obtained at $R = 0.6$ and $\tau = 0.18$ (Table 2). With a more unrealistic setting ($R_{max} = 1$), we find the best fit to the data when $R = 0.696$ and $\hat{\tau} = 0.25$ with $MLL_2 = -3753.41$. It is indicated that for any value of R_{max} the data fit model II significantly better than model I.

Model III

Model III incorporates the variation in τ assuming τ follows a gamma distribution. For $R_{max} = 0.5$ the maximum likelihood estimates are $R = 0.5$ (95% C.I. = $0.498 - 0.5$), $\tau = 0.18$ (95% C.I. = $0.11 - 0.27$) and $k = 2.4$ (95% C.I. = $2.0 - 3.0$) with the maximum log likelihood $MLL_3 = -3859.44$. MLL_3 is significantly larger than MLL_2 (likelihood ratio test: $P \approx 0$), indicating that the data fit model III significantly better than model II. Similar results are obtained for $R_{max} = 0.6$, but maximum likelihood for the models II and III are nearly identical when $R_{max} = 1$ (Table 2.2).

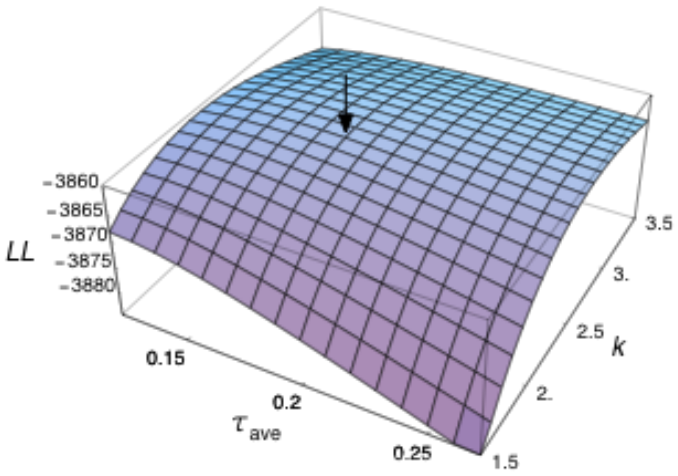


Figure 2.6: The log likelihood surface as a function of τ_{ave} and k under model III. R is fixed to be 0.5. The maximum likelihood estimate is shown by the arrow.

Table 2.2: Summary of ML analysis

Model	No. of parameters	MLL
I	1	-4641.01
$R_{max} = 0.5$		
II	2	-3934.82
III	3	-3859.44
$R_{max} = 0.6$		
II	2	-3786.62
III	3	-3777.23
$R_{max} = 1$		
II	2	-3753.41
III	3	-3753.41

2.6 Discussion

A maximum-likelihood method is developed to estimate the duration of concerted evolution and the time to the WGD event of yeast. The method utilizes the theoretical results by Teshima and Innan (2004), who demonstrated that the time of concerted evolution approximately follows an exponential distribution. Estimation of the duration of concerted evolution is extremely difficult when we do not know the date of the duplication event. To overcome this problem, we use many duplicated genes which appeared at the same time (*i.e.,* whole genome duplication). Yeast is one of the ideal species to apply this method to because of the availability of the genome sequences of *S. cerevisiae* (Goffeau et al. 1996) and its relatives (Kellis, Birren, and Lander 2004).

The application of our ML method demonstrates a crucial role of concerted evolution via gene conversion after gene duplication in yeast because the models with concerted evolution (models II and III) fit the data significantly better than the null molecular clock model (model I). It is also shown that the time to the WGD is underestimated under the molecular clock model. In models II and III, the ML estimate of R is 0.5, suggesting that the WGD occurred in very early stages after speciation with *K. waltii* or the WGD might have been involved in the speciation event.

When the expected duration of concerted evolution (τ) is assumed to be constant (model II), the ML estimate of τ is 0.12. If we assume that the WGD event occurred about $100-150$ million years ago, τ is $24-36$ million years. Gao and Innan (2004) have estimated τ to be about $25-86$ million years from different methods, in which the time of concerted evolution in *S. cerevisiae* is considered directly on the species tree of *S. cerevisiae* and its six relatives. Our estimate is roughly in agreement with that of Gao and Innan (2004).

Model III incorporates the variation in τ in model II. Model III is more realistic because τ depends on many parameters (Teshima and Innan 2004), which may not be constant over the genome. Selection is one of the most important factors to cause variation in τ among genes. For example, selection could work such that a larger amount of a gene product is favored (Kondrashov and Koonin 2004), which

is likely for ribosomal and histon genes. For such genes, τ might be larger than other genes. In fact, the $\sim$450 yeast genes pairs identified by Kellis, Birren, and Lander (2004) include many ribosomal and histon genes. We find that model III explains the data significantly better than model II. The ML estimate of SD of τ is $2.4 \times \tau_{ave}$, indicating τ is very variable.

There are several limitations in our model. First, we assume a constant evolutionary rate over time, but it could fluctuate by the changes of selective pressure. For example, Lynch and Conery (2000) suggested that selective pressure might be relaxed shortly after gene duplication. This possibility was somehow incorporated by investigating the likelihood up to $R_{max}(> 0.5)$. However, we found that R_{max} may not be much larger than 0.5. We modified the ML equation to estimate R_{max} using the *D. hansenii* sequence as an outgroup, and it turned out that the ML estimate of R_{max} is 0.49. Another possible scenario is that selective pressure could be relaxed on only one of the duplicated genes, for example, under a neofunctionalization model. Ohno (1970) describes this process such that a redundant copy created by duplication could be "freed" from selective pressure. Since it is very difficult to incorporate this effect into our system, as a proxy, we repeated the same analysis after excluding 63 trios, for which the evolutionary rates on the lineages leading to the two yeast duplicates are significantly different (Tajima 1992). Note that this treatment may not be very fair because the trios excluded are biased toward those with higher r because of the statistical power. Nevertheless, very similar results are obtained.

Second, we assume a Gamma distribution to take into account the variation in the expected duration of concerted evolution, τ. Unfortunatley, almost no prior information on this distribution is available. There are many factors to determine τ, including mutation, gene conversion, recombination rate and selection. Therefore, our Gamma approximation might oversimplify the situation.

This study demonstrates a significance role of concerted evolution after gene duplication on a genomic scale in yeast. We have successfully estimated the duration of concerted evolution via gene conversion in yeast duplicated genes, indicating that gene conversion is a very important mechanism in the evolution of

duplicated genes. The results suggest the importance of the analysis of duplicated genes incorporating the effect of gene conversion rather than simple analysis based on the molecular clock model. As discussed in Teshima and Innan (2004) and Gao and Innan (2004), molecular clock-based analysis causes a bias when the effect of gene conversion is not negligible. Examples of genome-wide analysis of duplicated genes with the molecular clock model include estimation of the age distribution of duplicated genes (Gu, Wang, and Gu 2002, McLysaght, Hokamp, and Wolfe 2002) and estimation of the rates of gene duplication and loss (Lynch and Conery 2000). Together with recent evidence for frequent gene conversion in various species (see Innan (2003b) and references therein), such analysis should be understood carefully, especially when applied to gene conversion-rich species such as yeast. The extent of interlocus gene conversion on a genomic scale in other organisms is an open question. The development of theories that incorporates gene conversion is also needed to better understand the evolution of duplicated genes.

3

Selection for more of the same product as a force to enhance concerted evolution of duplicated genes

3.1 Abstract

The duration of concerted evolution after gene duplication is highly variable across genes. To identify the cause of the variation, we analyzed of duplicated genes in yeast that originate from a whole genome duplication event. There appears to be a strong positive correlation between the duration of concerted evolution and the gene expression level. This observation can be explained by selection favoring more of the same product, which could enhance concerted evolution in dosage-sensitive genes.

3.2 Introduction

After a gene duplication event, the duplicated genes might be subject to concerted evolution, which is the phenomenon whereby duplicated copies coevolve by homogenizing DNA sequences between each other. Gene conversion is considered to be the major mechanism of concerted evolution of duplicated genes (*i.e.* a small multigene family) (Li 1997, Ohta 1980). Recently, it was demonstrated that concerted evolution by gene conversion might be quite common on a genomic scale in yeast (Gao and Innan 2004). However, there is little knowledge about the role of concerted evolution in the evolution of duplicated genes. Here, we considered how natural selection works for or against concerted evolution after gene duplication by using the bakers yeast *S. cerevisiae* as a model. The current *S. cerevisiae* genome has $\sim$450 pairs of duplicated genes, which originated from a whole-genome duplication (WGD) event (figure 3.1), and many of these genes underwent concerted evolution by gene conversion (Gao and Innan 2004, Sugino and Innan 2005). The fact that these duplicated genes appeared at the same time makes it possible to estimate the duration of concerted evolution, $\hat{c}$, for each gene pair (see below). We previously estimated that the time of the WGD event is very close to the speciation event with *K. waltii*, and that concerted evolution lasted roughly $24-36$ million years, on average (Sugino and Innan 2005). Furthermore, it was found that the duration of concerted evolution is extremely variable among

gene pairs (Sugino and Innan 2005). Here, we sought to discover the mechanism explaining the large variance of $\hat{c}$.

The *S. serevisiae* genome experienced a whole genome duplication (WGD) event roughly 100-150 MYA after the speciation with *Kluyveromyces waltii* (Dietrich et al. 2004, Kellis, Birren, and Lander 2004, Wolfe and Shields 1997). Kellis *et. al.* (Kellis, Birren, and Lander 2004) found that after substantial amounts of gene loss following the WGD, the current *S. serevisiae* genome have $\sim$450 pairs of duplicated genes originated from the WGD. Using these gene pairs, we have recently estimated that the WGD occurred very shortly after the speciation with *K. waltii* or the WGD was involved in the speciation event (Sugino and Innan 2005). Here, we estimate the duration of concerted evolution using *Debaryomyces hansenii* as an outgroup (Figure 3.1). Concerted evolution retards the divergence between two duplicates as illustrated in figure 3.1, in which the thick line represents the time under concerted evolution. We consider $c = T_a/(T_a + T_b)$ as a measure of the duration of concerted evolution. Note that c does not exactly represent the duration of concerted evolution if the WGD occurred after speciation. Nevertheless, c should be a good summary statistic to measure the relative effect of concerted evolution because the gap between the WGD and speciation is constant for all gene pairs. c was estimated from the alignment of the four sequences (S1, S2, K and D in figure 3.1). From the alignment, we extracted well-aligned regions and identified the numbers of mutations unambiguously mapped in T_a and T_b in a parsimonious way. Only the first and second positions of the codon were considered (Sugino and Innan 2005). Let a be the observed number of mutations in time T_a, and b_1 and b_2 represent those on the external branches leading to S_1 and S_2, respectively. Then, c could be approximately estimated by $\hat{c} = a/d$, where $d = a + (b_1 + b_2)/2$. In the analysis, we used the 132 gene pairs for which $\hat{c}$ is estimated from $d \geq 5$.

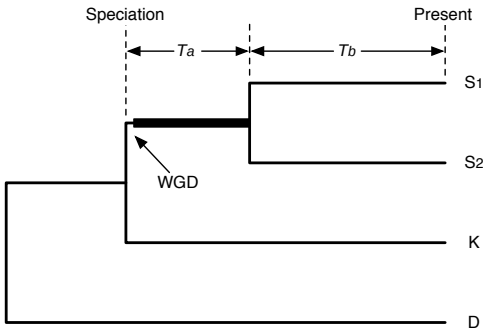


Figure 3.1: The phylogenetic relationship of four genes. S1 and S2: two copies of duplicates from the whole genome duplication (WGD) in *S. serevisiae*. K: the ortholog of S1 and S2 in *K. waltii*. D: the ortholog in *D. hansenii*.

3.3 Result and Discussion

3.3.1 The effects of gene conversion rate and sequence conservation on $\hat{c}$

We first considered two factors that could directly affect the duration of concerted evolution according to recent theoretical studies (Innan 2002, Teshima and Innan 2004). One is the gene conversion rate, and it is easy to imagine that $\hat{c}$ could be large for gene pairs with a high gene conversion rate. The other is sequence conservation due to purifying selection. If a gene pair is under the pressure of strong purifying selection, the sequence divergence between them is restricted, providing a situation in which gene conversion occurs efficiently. Therefore, it was predicted that highly conserved genes might have a large $\hat{c}$. However, we found that these two factors do not have strong correlations with $\hat{c}$, indicating that there might be other factors to explain the observed large variation of $\hat{c}$.

As potential causes to explain the observed large variation in $\hat{c}$, we consider the gene conversion rate and sequence conservation due to purifying selection (Sugino and Innan 2005, Teshima and Innan 2004). Figure 3.2a shows the relationship between $\hat{c}$ and estimates of local recombination rate, which is considered

to have a positive correlation with the gene conversion rate because gene conversion is a non-reciprocal recombination process. It should be noted that almost all analyzed gene pairs from the WGD are located on different chromosomes. Although the rate of interlocus gene conversion may be high between tandemly duplicated genes, we assume that there may not be such location effect in our data set without tandem duplications. In figure 3.2a, there is no significant correlation between $\hat{c}$ and estimates of recombination rate from the data of Gerton et al. (2000) ($r = 0.01$, $p = 0.91$, permutation test). It is suggested that the contribution of the variation in the gene conversion rate to the observed variation in $\hat{c}$ may not be large, although the correlation between the interlocus gene conversion and recombination rates is not very clear.

3.3.2 Highly expressed genes favor the long duration of concerted evolution

Next, we investigate the effect of sequence conservation due to purifying selection. First, to measure the intensity of selection at the amino acid level, we use the level of nucleotide identity at the first and second positions of the codon between *S. serevisiae* and *K. waltii*. Figure 3.2b shows that there is no significant correlation between the identity and $\hat{c}$ ($r = 0.05$, $p = 0.60$, permutation test). It is suggested that the variation in the level of sequence conservation alone cannot account for the observed variation in $\hat{c}$. However, a slightly different result is obtained if we use the identity at all three positions of the codon (Figure 3.2c), that is, there is a significant positive correlation between them ($r = 0.61$, $p < 0.0001$). This discrepancy may be explained by strong codon bias at the third position of the codon especially in gene pairs with large $\hat{c}$ (see below).

We then considered the effect of selection on the duration of concerted evolution. Selection could favor or disfavor concerted evolution, and we first considered the effect of the former $-$ that is, selection favoring gene conversion, so that concerted evolution can be enhanced. Dosage-sensitive genes are probably subject to such selection because producing more of the same product is advantageous

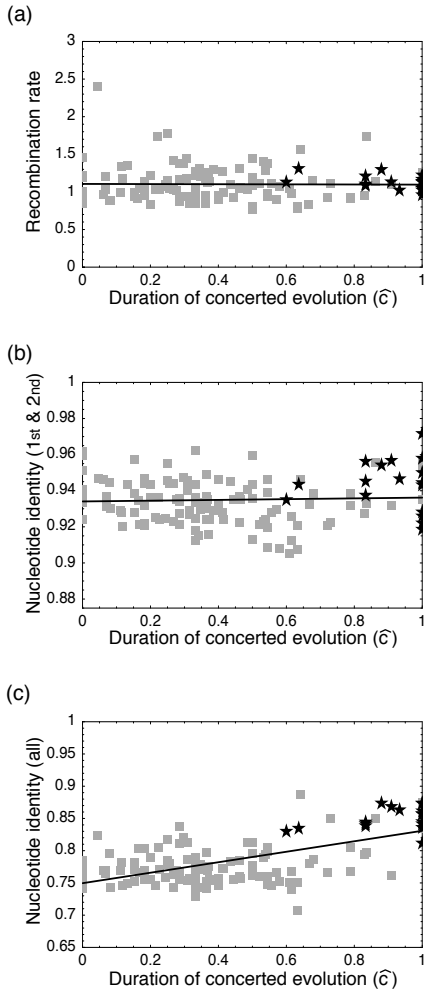


Figure 3.2: (a) The relationship between the duration of concerted evolution ($\hat{c}$) and recombination rate. The stars represent ribosomal genes and gray squares are for others. The regression line is also shown. (b) The relationship between $\hat{c}$ and sequence conservation between ortholog of *K. waltii* and *S. cerevisiae* (nucleotide identity). The first and second positions of the codon are used for the computation of the nucleotide identity. (c) The relationship between and sequence conservation (nucleotide identity). All three positions of the codon are used.

(Ohno 1970). Concerted evolution by gene conversion should be potentially beneficial because it helps to keep the sequence identity (Ohno 1970, Ohta 1989). A typical example is ribosomal genes, which have been known frequently to be under concerted evolution in various species since the first demonstration in the African toad *Xenopus* (Brown, Wensink, and Jordan 1972). Therefore, we hypothesized that selection for gene dosage might cause long-term concerted evolution. To test whether this hypothesis is generally true, we investigated the relationship between $\hat{c}$ and the level of gene expression. The hypothesis predicts that highly expressed genes should have a larger $\hat{c}$. The gene expression level was measured at the protein level using the data of (Ghaemmaghami et al. 2003). Because the mRNA hybridization technique was not involved, this dataset should be robust to the problem of cross-hybridization between duplicated genes in microarray data. As shown in figure 3.3a, there was a significant positive correlation between $\hat{c}$ and the protein expression level ($r = 0.23$; $P = 0.0004$, permutation test), supporting our hypothesis. Excluding ribosomal genes did not change the general trend ($r = 0.08$; $P = 0.0751$). A similar result was also obtained when the codon adaptation index (CAI) (Sharp and Li 1987) was used as a measure of gene expression (Figure 3.3b; $r = 0.67$, $P < 0.0001$, permutation test).

As demonstrated in figure 3.3b (see also figure 3.2a), there is a strong positive correlation between $\hat{c}$ and CAI ($r = 0.67$, $p < 0.0001$, permutation test), where CAI is a measure of codon bias (Sharp and Li 1987). Because it is known that codon bias is positively correlated with gene expression level (Ikemura 1981), this correlation could be considered to support our hypothesis: highly expressed genes have larger $\hat{c}$. However, this result should be interpreted carefully because CAI is directly related to GC-content, which could be increased by gene conversion if gene conversion is GC-biased (*i.e.*, biased gene conversion (Galtier 2003, Marais, Charlesworth, and Wright 2004)). Recently, it was reported that GC3 (GC-content at the third position of the codon) is elevated in duplicated genes that were subject to concerted evolution for a long time, concluding that GC-biased gene conversion has increased GC3 in those genes (Benovoy et al. 2005). Indeed, we also observe a positive correlation between $\hat{c}$ and GC3 (Figure 3.4b). There-

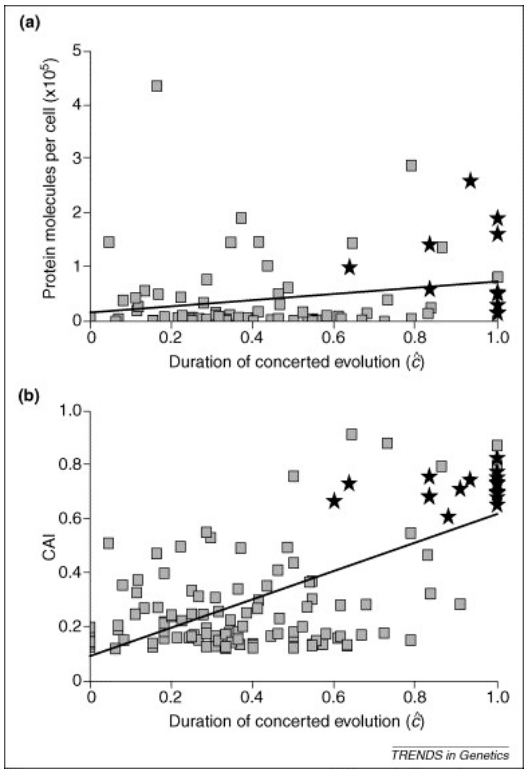


Figure 3.3: (a) The relationship between $\hat{c}$ and protein expression level. The stars represent ribosomal genes and the gray squares are for others. The regression line is also shown. (b) The relationship between $\hat{c}$ and CAI.

fore, if GC-biased gene conversion is the major cause to create high CAI in genes that experienced long-term concerted evolution, CAI may not be a good measure of the level of gene expression. To examine this possibility, we focus on the 44 genes (22 pairs) that underwent concerted evolution for a long time ($\hat{c} > 0.8$). The average CAI and GC3 for these genes are 0.68 and 0.41, respectively. It is expected that this observed GC3 $= 0.41$ may be higher than the average GC3 in singleton genes with similar levels of CAI, if gene conversion is highly GC-biased. We define singletons as genes with no BLASTP hit in the *S. cerevisiae* genome when the e-value cutoff is 0.1. It is important to note that our interest is in interlocus gene conversion that occurs between duplicated genes, while gene

conversion also occurs between homologous regions. Our hypothesis that GC3 in duplicated genes is higher than that in singletons is based on the prediction that duplicated genes are subject to interlocus gene conversion in addition to homologous gene conversion. We find that the average GC3 for singleton genes whose average CAI $= 0.68$ is 0.44. This observation is in the opposite direction expected under the hypothesis of GC-biased gene conversion, suggesting the effect of GC-biased gene conversion on the observed positive correlation between $\hat{c}$ and CAI may be small. One of the explanations for the observed correlation between $\hat{c}$ and GC3 may be an artifact due to the strong correlation between GC3 and CAI. This highly positive correlation between $\hat{c}$ and CAI can also explain the positive correlation in figure 3.4c, because preferred codons of *S. cerevisiae* is nearly identical to those of *K. waltii*, creating high sequence identity at the third position of the codon. Thus, it could be concluded that the high CAI in gene pairs with large $\hat{c}$ should be due to high gene expression rather than GC-biased gene conversion.

To verify the above conclusion, we also analyze CAI and GC3 in *K. waltii*s orthologous genes of the duplicates in *S. cerevisiae* Considering the importance of selection on dosage, it is hypothesized that the level of gene expression (*i.e.*, CAI) would also be high in *K. waltii*s orthologous genes of those underwent long-term concerted evolution in the *S. cerevisiae* lineage. As expected, we observe a significant positive correlation between $\hat{c}$ and CAI in *K. waltii* (Figure 3.4c). We also observe a positive correlation between $\hat{c}$ and GC3 in *K. waltii* (Figure 3.4d), but this correlation cannot be explained by the GC-biased interlocus gene conversion hypothesis because the orthologous genes in *K. waltii* are not duplicated. With these results, it is concluded that the relative contribution of the GC-biased gene conversion to the observed positive correlation between $\hat{c}$ and CAI may be small.

Thus, it was demonstrated that selection for higher gene dosage probably prefers concerted evolution. It is known that selection for dosage works not only for higher dosage, but also for dosage balance. Papp, Pál, and Hurst (2003) argued that selection could work on genes producing subunits of the same protein to maintain their dosage balance. In such genes, it might be possible to consider that concerted evolution is preferred, although it might be difficult to demonstrate

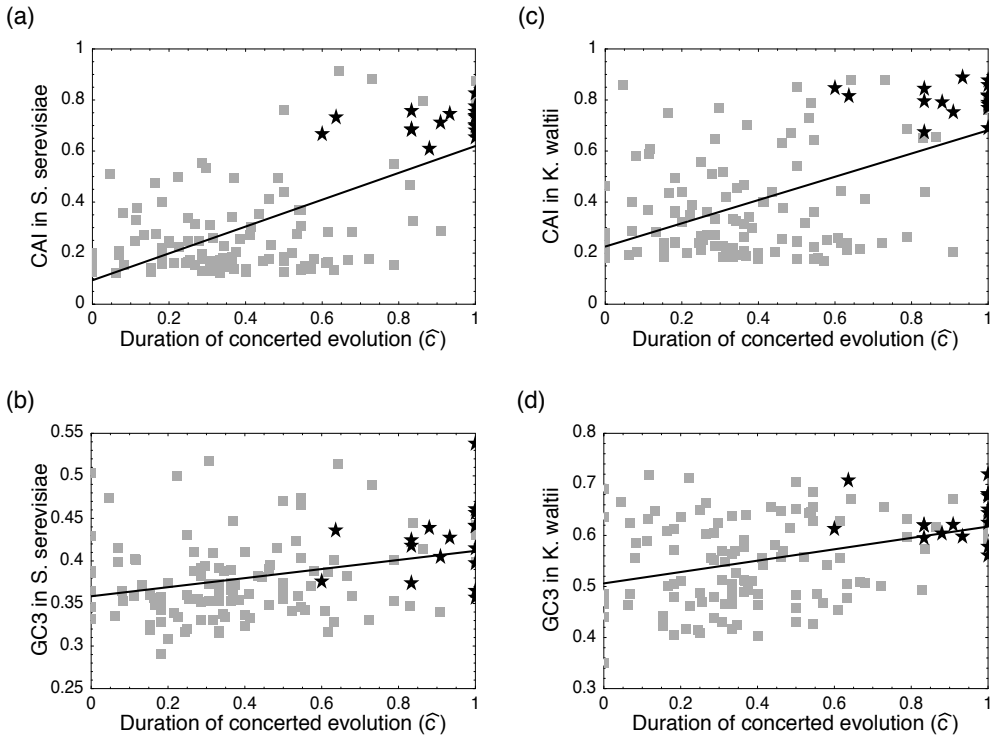


Figure 3.4: (a) The relationship between the duration of concerted evolution ($\hat{c}$) and CAI in *S. cerevisiae*. The stars represent ribosomal genes and the gray squares are for others. The regression line is also shown. This figure is identical to figure 3.3b. (b) The relationship between $\hat{c}$ and GC3 in *S. cerevisiae*. (c) The relationship between $\hat{c}$ and CAI in *K. waltii*. (d) The relationship between $\hat{c}$ and GC3 in *K. waltii*.

the effect of this type of selection with our data.

It was pointed out by (Lin et al. 2006) that concerted evolution via gene conversion is effective only in the presence of strong codon-bias and protein sequence conservation. They draw phylogenic tree of ohnologs and found that irregular tree, which is the evidence of concerted evolution, was observed in strong codon biased genes and concluded codon bias is the constant force of slow down of sequence evolution. Our result is not inconsistent with their result because long-term concerted evolution have occurred in highly expressed genes. In addition, it is theoretically proved that selection works more efficiently in duplicated genes with gene conversion between them (Innan and Kondrashov 2010, Mano and Innan 2008). We conclude that concerted evolution via gene conversion play a significant roles in the evolution of ohnologs.

3.3.3 Dosage sensitive genes also favor the long duration of concerted evolution

Selection for higher gene dosage could also have an important role in gene duplication. It is thought that genes for which selection works for higher dosage would have high duplicability that is, the probability that a duplication fixes in the population is high (*i.e.* 'duplication for the sake of producing more of the same', as stated by Ohno (Ohno 1970)). Under the framework of gene duplication proposed by Kondrashov and Koonin (2004), such genes are likely to be haploinsufficient genes rather than haplosufficient genes. Haploinsufficient genes are those for which the fitness of the heterozygote of the wild-type allele and mutant allele exhibit a sufficient reduction in fitness in comparison with the homozygote of the wild-type allele, whereas haplosufficient genes are those for which there is no significant reduction in the fitness of the heterozygote. We proposed that genes with higher duplicability are more likely to undergo long-term concerted evolution if selection for more of the same product has a crucial role before and after duplication. For duplicated genes that originated from the WGD in yeast, it might be considered that duplicated copies of haploinsufficient genes would be more

preferentially retained than haplosufficient genes (Davis and Petrov 2004). To test this hypothesis, the data from a systematic experimental gene knockout study by Steinmetz et al. (2002) were used. In this study, the authors developed gene deletion strains, and their growth rates were measured under five different medium conditions (two fermentable and three non fermentable substrates). In our analysis, we focused on genes with a significant reduction (5%) in the fitness of the deletion homozygote under at least one medium, and these genes were classified into haplosufficient and haploinsufficient genes. We defined the former as those with the minimum fitness of the five media $> 99\%$ and the latter as the minimum fitness $< 97.5\%$. We first found that the ratio of haploinsufficient to haplosufficient genes was significantly higher for the duplicated genes from the WGD than that for singleton genes (84:45 for WGD versus 138:159 for singletons; $P = 0.005$, exact test), confirming the results of Kondrashov and Koonin (2004). It was also found that the ratio in genes with a large $\hat{c}$ (>0.8) was 16:4, which was higher than that for the others (18:11), as expected, although not significant. A significant excess of haploinsufficient genes for $\hat{c} > 0.8$ was observed when the data for the two fermentable substrates were considered (15:4 versus 7:10; $P = 0.039$), whereas no significant difference was obtained for the three non fermentable substrates. Thus, selection for dosage might be an important evolutionary force both before and after gene duplication. Selection might be more important in regular conditions rather than in nutrient-limited conditions.

3.3.4 The possibility of disfavoring concerted evolution

Finally, we considered selection that works in the other direction. Concerted evolution could be disfavored when selection works to maintain the genetic divergence between duplicated genes. Evolution of genetic novelty by gene duplication occurs such that one copy keeps the original function, whereas the other acquires a new function created by beneficial mutations (Ohno 1970). In such a case, gene conversion could be deleterious because it could erase the beneficial mutation (Innan 2003a). The population can maintain the beneficial mutation for a reasonable length of time only when selection is so strong that deleterious gene conversions

are quickly eliminated, and consequently concerted evolution can be terminated (see Innan (2003a) for an interesting example in the human RH genes). To investigate how often such events occurred in a relatively short time after the WGD, we investigated whether the proportion of haplosufficient genes was elevated in genes of low $\hat{c}$ (<0.2) in comparison with the others ($\hat{c} > 0.2$). It was expected that if the duplicated copies had diverged functionally, and both have crucial roles at present, these genes could be haplosufficient rather than haploinsufficient. However, we did not find strong evidence for this mode of selection, although the proportion of haplosufficient genes was slightly higher for genes of $\hat{c} < 0.2$ (the numbers of haplosufficient and haploinsufficient genes were three and five, respectively, for $\hat{c} < 0.2$, and 12 and 29 for $\hat{c} > 0.2$; $P = 0.69$, exact test). Although this non significant result might have been partly due to a lack of statistical power because of the small amounts of data, neofunctionalization might not be a likely fate of duplicated genes under the pressure of homogenization by gene conversion (Innan 2003a).

3.4 Conclusion

Here, we have considered the relationship between natural selection and the duration of concerted evolution of duplicated genes. Selection could work for or against concerted evolution. Gene conversion can be preferred by selection that works for higher gene dosage, whereas it can be disfavored when divergence between duplicates is advantageous. Our analysis demonstrates that selection is likely to be one of the main factors determining the duration of concerted evolution. Concerted evolution could be enhanced in genes in which a higher gene dosage is required.

4

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

4.1 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.

4.2 Introduction

It is widely accepted that the order of coding genes is not random (Hurst, Pál, and Lercher 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 gene 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,

Blumenthal, and Hurst 2003) and mammals (Fukuoka, Inaoka, and Kohane 2004, Li et al. 2006, Sémon and Duret 2006, Singer et al. 2005). 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 *S. 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 co-expressed 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, Herr and Harris 2004, Kensche et al. 2008, Kruglyak and Tang 2000, Trinklein et al. 2004). 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 $\sim 100 - 200$ million years ago (mya) (Dietrich et al. 2004, Kellis, Birren, and Lander 2004, Wolfe and Shields 1997). 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, Williams, and Pál (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, Morris, and Li (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, ten Wolde, and Teichmann (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.

4.3 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, Byrne, and Wolfe 2009), which includes $\sim 5,600$ 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, Byrne, and Wolfe (2009), which is also included in the YGOB. This ancestral genome is referred to as the pre-WGD genome (figure 4.1).

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 (Batada, Urrutia, and Hurst 2007, Lercher, Urrutia, and Hurst 2002, Williams and Hurst 2002). 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 H2A.Z 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).

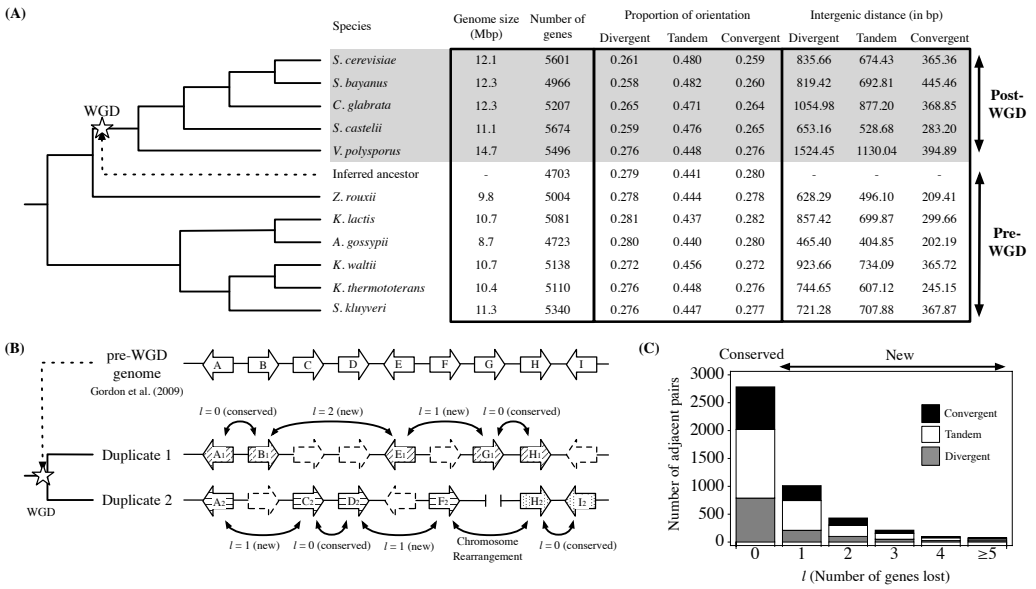
Gene expression data

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

4.4 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 $100-200$ mya (Dietrich et al. 2004, Kellis, Birren, and Lander 2004, Sugino and Innan 2005, Wolfe and Shields 1997). In relation to this event, different yeast species were classified into two categories, pre- and post-WGD species (figure 4.1A). We first compared several properties of the genomes between the two categories. The current genomes of pre-WGD species have on average $\sim$5,000 genes and the genome size are roughly 10 Mb (figure 4.1A, see also Cliften et al. (2003), Kellis et al. (2003), Byrnes, Morris, and Li (2006) and 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, Byrne, and Wolfe (2009) (see also figure 4.1). 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.



Species	Genome size (Mbp)	Number of genes	Proportion of orientation			Intergenic distance (in bp)			
			Divergent	Tandem	Convergent	Divergent	Tandem	Convergent	
S. cerevisiae	12.1	5601	0.261	0.480	0.259	835.66	674.43	365.36	Post-WGD
S. bayanus	12.3	4966	0.258	0.482	0.260	819.42	692.81	445.46	
C. glabrata	12.3	5207	0.265	0.471	0.264	1054.98	877.20	368.85	
S. castelii	11.1	5674	0.259	0.476	0.265	653.16	528.68	283.20	
V. polysporus	14.7	5496	0.276	0.448	0.276	1524.45	1130.04	394.89	
Inferred ancestor	-	4703	0.279	0.441	0.280	-	-	-	Pre-WGD
Z. rouxii	9.8	5004	0.278	0.444	0.278	628.29	496.10	209.41	
K. lactis	10.7	5081	0.281	0.437	0.282	857.42	699.87	299.66	
A. gossypii	8.7	4723	0.280	0.440	0.280	465.40	404.85	202.19	
K. waltii	10.7	5138	0.272	0.456	0.272	923.66	734.09	365.72	
K. thermototerans	10.4	5110	0.276	0.448	0.276	744.65	607.12	245.15	
S. kluyveri	11.3	5340	0.276	0.447	0.277	721.28	707.88	367.87	

Figure 4.1: Summary of the evolutionary analysis of adjacent genes. (A) Phylogenic relationship of yeast species. The star represents the WGD event, which occurred $\sim 100-200$ mya. 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 Génolevures Consortium et al. (2009). (B) Illustration of a typical pattern of the post-WGD genome re-organization 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.

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

4.4.1 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 figure 4.1B). We found that in post-WGD species, roughly half (47$\sim$ 48%) of the adjacent gene pairs are in tandem orientation and the others are in divergent and convergent orientations ($\sim$ 26% for each) (figure 4.1). In the pre-WGD genome, these proportions are also similar, although the proportions of divergent and convergent gene pairs ($\sim$ 28% for each) are slightly larger than those of post-WGD species (figure 4.1). 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 4.1B, 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 4.1. For each adjacent gene pair in *S. cerevisiae*, we identified the locations and orientations of their orthlogous 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 4.1B 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 loss is needed to explain the four pairs (A_1-B_1, G_1-H_1, C_2-D_2, H_2-I_2); therefore l is estimated to be 0, indicating the adjacency of these pairs has been conserved since WGD.

The YGOB database consists of $\sim 5,600$ coding genes (*i.e.,* $\sim 5,600$ adjacent pairs) in the *S. cerevisiae* genome and their orthologs in other yeast species and the inferred pre-WGD genome (Gordon, Byrne, and Wolfe 2009). We successfully identified the orthologous gene pairs in the pre-WGD genome for $\sim$ 80% of the adjacent genes in *S. cerevisiae* ($n = 4,617$). We found that $\sim 90\%$ of them ($n = 4,440$) has their orthologous genes on the same chromosomes in the pre-WGD genome, for which we estimated l. Figure 4.1C shows the distribution of l, indicating that $\sim 60\%$ ($n = 2,657$) 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 $l \leq 3$, 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 ($\sim$28.0%), but it is significantly reduced in the new adjacent gene pairs (22.0%, $P < 0.0001$, exact test, table 4.1). 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 (*Z. rouxii, K. lactis, A. gossypii, K. waltii, K. thermototerans and S. kluyveri*). We next compared the pre-WGD genome and four post-WGD species (*S. bayanus, C. glabrata, S. castelii and V. 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 tandem 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 are 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 (figure 4.2A). In figure 4.2B, 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

Table 4.1: Coexpression and Intergenic Distance for Adjacent Gene Pairs in *S. cerevisiae*.

	Number of adjacent genes			Intergenic distance (in bp)			Coexpression (r)		
CDS data	Conserved	New	Total	Conserved	New	difference	Conserved	New	difference
Divergent	751 (28.3 %)	351 (22.0 %)	25.9 %	581.85	967.37	385.52	0.235	0.187	−0.047**
Tandem	1179 (44.4 %)	809 (50.7 %)	46.8 %	487.20	613.01	125.81	0.162	0.158	−0.003
Convergent	727 (27.4 %)	435 (27.3 %)	27.3 %	249.42	333.52	84.10	0.206	0.203	−0.002
UTR data	Number of adjacent genes			Intergenic distance (in bp)					
Divergent	555 (27.4 %)	264 (21.5 %)	25.2 %	414.99	873.69	458.70			
Tandem	859 (42.3 %)	589 (48.0 %)	44.5 %	305.00	397.15	92.15			
Convergent	614 (30.3 %)	374 (30.5 %)	30.3 %	−27.38	25.79	53.17			

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

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 figure 4.2B).

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 figure 4.1A), 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 figure 4.2B). The proportions of new divergent and convergent pairs stay at 25% through the simulation. Thus, we conclude that the two neutral simulations cannot explain the observed reduction in the proportion of new divergent gene pairs (20.7%) without considering selection against new divergent pairs.

4.4.2 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 reorga-

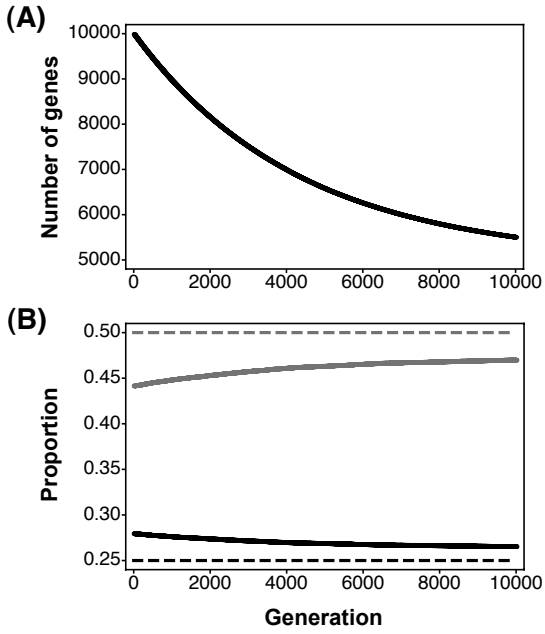


Figure 4.2: The behavior of the proportions of the three orientations after WGD through the decrease of the total number of genes. (A) Decrease of the total number of genes through the simulation. (B) The changes of the proportions of tandem and divergent (conserved) gene pairs (gray and black lines, 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%.

nization 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 4.1C, 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 4.1. New diver-

gent pairs have on average ∼400 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 untranlated regions and confirmed that the same trend holds (table 4.1).

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, Urrutia, and Hurst 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 4.3. 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 (figure 4.3A and B), and the other two create new tandem pairs (figure 4.3C 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 4.1) (this is also pointed out by a recent empirical study 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 4.2). 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 4.1. As shown in table 4.2, 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

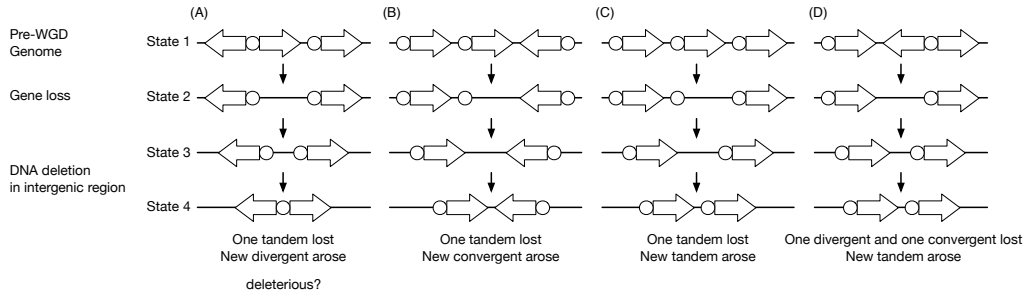


Figure 4.3: Simple illustration of gene loss and shrinkage of intergenic region by DNA deletion. Under our simple assumptions (see text), there are only four possible cases from (A) to (D). In all cases, the loss of the middle gene is considered. Coding genes and NFRs are presented by open arrows and 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).

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 vursus 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 4.2).

4.4.3 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 indels causing a frameshift

Table 4.2: Relationship Between the number of NFRs on the Coexpression and Intergenic Distance

	One NFR			Two NFRs			$\geq$ Two NFRs		
	No. of gene pairs	Coexpression (r)	Intergenic distance	Number of gene pairs	Coexpression (r)	Intergenic distance	Number of gene pairs	Coexpression (r)	Intergenic distance
Divergent									
Conserved	286	0.280	339.46	69	0.225	637.52	132	0.199	850.74
New	69	0.295	345.04	42	0.186	603.02	118	0.159	1197.85
Tandem									
Conserved	411	0.167	363.42	65	0.125	589.89	114	0.144	863.09
New	279	0.180	346.33	40	0.120	609.43	100	0.102	1327.06
Covergent									
Conserved	446	0.211	216.08	3	-0.126	570.33	9	0.101	654.67
New	232	0.216	262.47	6	0.160	506.50	19	0.132	991.74

(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. Selection should work against DNA deletion when it deletes an important functional part of the intergenic region. To incorporate this effect in the simulation, we suppose that there is a minimum length of intergenic region to accommodate all necessary regulatory elements, $L_{\min}$. Then, it is assumed that selection works against deletion especially when the intergenic region becomes short and close to the minimum length. This selection pressure is relaxed when the intergenic region is very long. Therefore, f is designed such that the intensity of selection increases (hence, the fitness increases) as the length of the intergenic region decreases. f involves a parameter g, which determines the shape of the function. When g is large, the fitness dramatically decreases as the intergenic length approaches $L_{\min}$. For a small g, the fitness increases nearly

linearly when the inetergenic length is similar to L_{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.

The simulation starts at the WGD event, where all chromosomes are doubled in a single genome, and the subsequent evolutionary process through pseudogenization and DNA deletion is simulated. We designed the simulation by taking advantage of the fact that the genome organization of pre-WGD species has been quite conserved for a very long time, and that extensive rearrangements occurred only in post-WGD species (Byrne and Wolfe 2005). Therefore, we can predict that in the ancestral genome at the WGD event, the proportion of the three orientations (divergent, tandem, convergent) should be roughly 28%, 44% and 28% (see figure 4.1A), respectively, which is assumed in our simulation. In practice, we randomly created a genome with ∼5000 genes, which represents the genome before the WGD event. Those genes are randomly arranged such that the proportions of the three orientations are consistent with the observation (*i.e.,* 28%, 44% and 28%). Then, the entire genome is doubled (*i.e.,* whole genome duplication), and the subsequent reorganization process is simulated with random pseudogenization and DNA deletion. To estimate the intensity of selection against deletion, multiple steps of simulations are involved as outlined below. The symbols used in this simulation are listed in table 4.3.

1. Simulate the pseudogenization process alone to estimate m, which is defined as the rate at which a pseudogenizing mutation occurs per gene per time unit. In this simulation, DNA deletion is ignored because it is not relevant. We assume pseudogenization is a neutral process except for one condition: for each duplicated gene pair, one gene can be pseudogenized if the other is functional. With this condition, the genome will eventually have only one copy for all gene pairs. This process can be described in a population genetic framework as follows. If we suppose a random-mating diploid population with size N, then a pseudogenizing mutation fixes in the population with probability $1/(2N)$ if the other copy is functional, and with probability 0 if the other copy is already pseudogenized. This simulation is

Table 4.3: Symbols used in the gene loss-deletion simulations

Symbol	Description
N	Diploid population size
m	Gene loss rate per gene per time unit
k	Number of the WGD-derived duplicated genes
u	Rate of DNA deletion per 1 kb time unit
λ	Mean size of DNA deletion
L	Length of intergenic region. Subscript (D, T or C) specifies the orientation of the gene pair; D: divergent, T: tandem, and C: convergent.
f	Fitness of an individual determined by the length of intergenic region.
g	Parameter to determine the shape of fitness function. Subscript follows those for L.

almost identical to what is described earlier, and similar to that of Byrnes, Morris, and Li (2006). A time unit is defined such that the time interval from the WGD event to present is divided into 10,000 time units; therefore, one time unit corresponds to 10,000 years if the WGD event occurred 100 mya.

2. Simulate the pseudogenization and DNA deletion processes simultaneously, in which the estimated m in the previous step 1 is used. The purpose of this simulation is to estimate the DNA deletion rate and the selective pressure against deletion. It is assumed that the selection intensity against deletion is very strong when the intergenic region is short and close to the minimum, while almost no selection works when the intergenic region is very long. Although this logic should apply to all three orientations, we hypothesized that this selective pressure is particularly strong in the intergenic region between a divergent gene pair. This step of simulation is to estimate the selection intensity for the tandem and convergent gene pairs, which will be a control to compare with that of the divergent gene pairs. We assume that the deletion rate is constant for all three orientations.

3. Simulate the pseudogenization and DNA deletion processes simultaneously, in which the estimated m in step 1 and the estimated DNA deletion rate and selection parameters in step 2 are used. In this last step, we aim to estimate the intensity of selection against DNA deletion for divergent gene pairs, which will be compared to those for tandem and convergent pairs.

In step 1 we ask what rate of pseudogenization explains the gene content in the current *S. cerevisiae* genome. In the current *S. cerevisiae* genome, only 10% of the genes from the WGD event still remain as two copies (Dietrich et al. 2004, Kellis, Birren, and Lander 2004). The rate, m, is estimated by using a simulation-based approximate likelihood method (Marjoram et al. 2003). In our simulation, for simplicity, we assume that the initial state of the genome has 5120 pairs of duplicated genes on the same lengths of eight chromosomes (each has 640 genes). In practice, for a single m value, 10,000 replications of simulations of 10,000 time units are performed, in which we focus on k, the number of two-copy genes. Obviously, the initial value of k is 5,120, and k decreases over time. A simulation run is accepted if the simulated genome has a similar value of k to the observation (*i.e.,* $k_{obs} = 5120 \times 0.1 = 512$), and we consider that the proportion of accepted replications should approximately represent the likelihood of m. To evaluate the similarity, we use the approach of weighted acceptance following Beaumont, Zhang, and Balding (2002), in which a replication is accepted with probability:

$$\begin{cases} c\,\delta^{-1}(1 - t^2\,\delta^{-2}), & t \leq \delta \\ 0, & t > \delta \end{cases} \tag{4.1}$$

where $t = |k - k_{obs}|$. c is a normarilizing constant, which is assumed to be $c = 3/4$ following Beaumont, Zhang, and Balding (2002). δ is tolerance of acceptance and assumed to be 10 (our additional simulations confirmed that this choice of $\delta = 10$ provides quite a good estimate, and that $\delta < 10$ does not necessarily improve our estimate). We evaluated the approximate likelihood for a wide range of m, and obtained a maximum likelihood estimate of $m = 1.15 \times 10^{-4}$ per gene per time unit (the 95% confidence interval (C. I.): $1.00 - 1.33 \times 10^{-4}$). When the time

of the WGD event is assumed to be 100 mya, the psuedegenization rate can be translated to $m = 1.15 \times 10^{-8}$ per gene per year.

Next (step 2), we estimate the rate of DNA deletion and selection intensity for tandem and convergent gene pairs. In this simulation, the initial state of the simulated genome is again assumed to have 5,120 genes, following the simulation in step 1. The orientations of those genes are randomly determined such that the proportions of the three orientations (divergent, tandem and convergent) are consistent with that in the pre-WGD species. It is assumed that the lengths of all coding genes are 1,450 bp, which is the average over those of the pre-WGD species. The length of each intergenic region is determined by drawing a random number from the empirical distribution from the *S. cerevisiae* genome, because they should reflect the stable ancestral genomic state. We created the empirical distributions for the divergent, tandem, and convergent orientations, whose averages are 581.85, 478.20 and 249.42, respectively.

In this step 2, the gene loss and deletion processes are simulated simultaneously. The gene loss process is identical to that in step 1, except that the rate is fixed to our estimate, $m = 1.15 \times 10^{-4}$. It is assumed that DNA deletion occurs at any location in the intergenic regions. The rate of deletion is assumed to be u per 1 kb per time unit, and the deletion length follows a geometric distribution with mean length λ. We use three different lengths, λ = 10, 100, and 1000 bp, but because we obtained essentially identical results for the three values, our simulation procedure is here explained by using only λ = 100 bp. We assume that the genome is always always under selective pressure to make it compact, which is obvious from the observed extensive genome shrinkage in many post-WGD species (Byrne and Wolfe 2005, Dietrich et al. 2004, Kellis, Birren, and Lander 2004). Selection should work against DNA deletion when it deletes an important functional part of the intergenic region. To incorporate this effect in the simulation, we suppose that there is a minimum length of intergenic region to accommodate all necessary regulatory elements. Then, it is assumed that selection works against deletion especially when the intergenic region becomes short and close to the minimum length. This selection pressure is relaxed when the intergenic region is very long.

To incorporate this effect, we assume that the intensity of selection increases as the length of the intergenic region decreases. We set the minimum length of intergenic regions for the divergent, tandem and convergent orientations, denoted by $L_{D,min}$, $L_{T,min}$ and $L_{C,min}$. For each intergenic region, the minimum length was randomly determined from the empirical distribution in the *S. cerevisiae* genome.

Suppose that a deletion event occurred to change the intergenic length L to L'. Let f and f' be the fitness before and after this deletion, respectively. We assume that the fitness is given by a function of the selection parameter g, L, L', and L_{min}:

$$f = \begin{cases} 1 + \left\{ \left(1 - e^{-(L' - L_{mim})\, g}\right) - \left(1 - e^{-(L - L_{mim})\, g}\right) \right\} \\ \qquad\qquad L' - L_{mim} > 0 \\ 0, \qquad\qquad L' - L_{mim} \leq 0 \end{cases} \tag{4.2}$$

where L_{mim} can be either $L_{D,min}$, $L_{T,min}$ or $L_{C,min}$ depending on the orientation of the two adjacent genes. g can also be either g_D, g_T or g_C, representing the selection intensities of the three orientations. The relationship between f and $L - L_{min}$ (or f' and $L' - L_{min}$) is shown in figure 4.4, indicating that f is a monotonically increasing function of $L' - L_{min}$, and that g determines the slope shape. When g is large, the fitness dramatically decreases as $L - L_{min}$ gets close to zero. For a small g, the fitness increases nearly linearly when $L - L_{min}$ is small, and saturates at 1.

This selection can be incorporated in the simulation by translating f into q, the fixation probability. According to the standard theory of population genetics (Kimura 1983), q is given by

$$q = \frac{1 - e^{-2(f-1)}}{1 - e^{-4N(f-1)}}. \tag{4.3}$$

Because equation 4.3 includes the population size, q may be sensitive to the difference of population size. We here use two different values, $N = 10^6$ and 10^7, according to the following observations: from single-nucleotide polymorphism (SNP) data (Liti et al. 2009, Schacherer et al. 2009), the population mutation rate $\theta = 4N\mu$ (μ: mutation rate per site per generation) in *S. cerevisiae* had been estimated in different samples; 2.08×10^{-3} (essential genes) and 2.69×10^{-3}

(non-coding regions) by Schacherer et al. (2009), and 1.11×10^{-3} (Wine/European yeast) and 5.93×10^{-3} (global sample) by Liti et al. (2009). The mutation rate per site per generation (or cell division) was estimated to be 3.3×10^{-10} (Lynch et al. 2008). Assuming this, moment estimates of the population size ranges from 10^6 to 10^7. Here, we show the results with $N = 10^6$ because we obtained almost identical results for $N = 10^7$.

Thus, in our model the DNA deletion process is primarily characterized by two parameters, u and g. From our analysis above, we presume that deletion is more disadvantageous for divergent pairs than the others, that is, $g_D < g_T$ and g_C. To evaluate this effect, we first estimate g_T, g_C, and u simultaneously (step 2). Then, in step 3 we test whether g_D is smaller than g_T and g_C.

For estimating g_T, g_C, and u, a simulation-based approximate likelihood algorithm was designed. Two summary statistics are used to evaluate the likelihood in the algorithm, the average lengths of intergenic regions for new tandem and convergent pairs, denoted by L_T and L_C. For a single parameter set (g_T, g_C and u), the likelihood is approximately computed as the proportion of simulation runs with $L_{T,sim}$ and $L_{C,sim}$ similar to the corresponding averages in the *S. cerevisiae* genome ($L_{T,obs} = 613.01$ and $L_{C,obs} = 333.52$). We again use equation (4.1) with δ = 20 bp to determine the probability to accept a simulation run. For each simulation replication, we compute d for new tandem and convergent pairs ($d = L_{T,sim} - L_{T,obs}$ for tandem and $d = L_{C,sim} - L_{C,obs}$ for convergent). The acceptance probability is given by the product of two acceptance probabilities from equation (4.1); one is that for tandem and other is that for convergent. Approximate likelihoods are obtained for a wide range of the three parameters. By using pre-simulation runs with small numbers of replications (1,000 for each parameter set), we found that the parameter set that produces the maximum likelihood should be covered if we consider $g_T = [0.01, 0.30]$, $g_C = [0.01, 0.30]$, $u = [1.0 \times 10^6, 5.0 \times 10^7]$. In these intervals, we obtained approximate likelihood by 10,000 replications of simulations, changing g_T and g_C with increment 0.01 and u with increment 10^6.

It should be noted that this simulation requires an unknown value of g_D.

Therefore, we performed preliminary simulations with several different conditions, including $g_D = \{0.001, 0.01, 0.1, 1, 10, 100\}$. It was found that those simulations resulted in almost identical likelihood surfaces of g_T, g_C and u, indicating the effect of g_D on L_T and L_C may be small. This is not surprising because L_T and L_C are used as summary statistics, not L_D. g_D plays a significant role to determine L_D as will be shown below. In the following, we show the results of step 2 assuming $g_D = 0.1$.

From these simulations, we obtained the maximum likelihood estimates of $\{u, g_T, g_C\} = \{3.5, 0.11, 0.16\}$. The profiled likelihood for $\{g_T, g_C\}$ is shown in figure 4.5A in the main text, indicating that the selective pressure against deletion is similar for tandem and convergent gene pairs. We used these values for estimating g_D in the next step (see below).

Finally in step 3, we estimate the selection intensity against DNA deletion for new divergent gene pairs, g_D. Our prediction is that g_D should be significantly smaller than g_T and g_C, that is, the effect of DNA deletion on fitness is stronger for the divergent pair when the intergenic length is small (figure 4.4). Now, we have an estimate of g_T and g_C together with m and u from steps 1 and 2. By using these estimates, it is possible to test if our prediction holds by estimating g_D.

Here we use the identical simulation method used in the previous step, except that u, g_T and g_C are fixed to the estimated values, $\{u, g_T, g_C\} = \{3.5, 0.11, 0.16\}$. Then, g_D can be estimated by the same approximate likelihood method. We 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 figure 4.5B in the main text). This suggests that selection against DNA deletion is particularly strong for divergent gene pairs.

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 4.5A shows the profiled likelihood for $\{g_T, g_C\}$ (see Supplementary materials online for details), 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

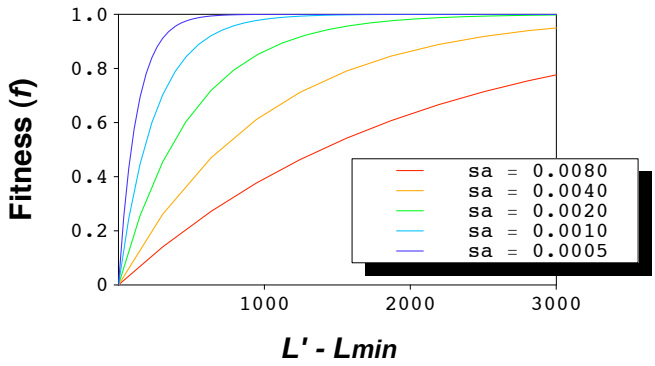


Figure 4.4: Illustrating the fitness effect of DNA deletion defined by equation 4.2.

$\sim$ 0.03 to 0.06, which is significantly smaller than the estimates of g_D and g_C (see also figure 4.5B). This suggests that selection against DNA deletion is particularly strong for divergent gene pairs, as we suspected.

4.5 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 4.1). In addition, we found that new divergent gene pairs had significantly longer intergenic regions than the

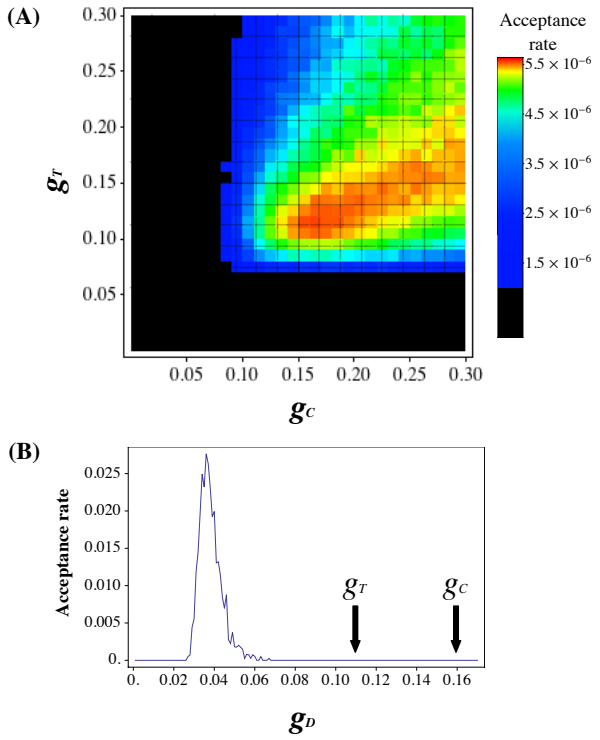


Figure 4.5: (A) Profiled likelihood surface for g_T and g_C. (B) Approximate likelihood (acceptance rate) for g_D conditional on $\{g_T, g_C\} = \{0.11, 0.16\}$.

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 (figure 4.5 and table 4.1). 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 (figure 4.5). Disadvantage of closely located genes have been suggested by Byrnes, Morris, and Li (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, Williams, and Pál 2002, Kensche et al. 2008). Hurst, Williams, and Pál (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, Williams, and Pál (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, Hermsen, ten Wolde, and Teichmann (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.

5

Conclusion and perspectives

5.1 Conclusion

Following the seminal work of Martin Kreitman (1983), many geneticists have analyzed single nucleotide polymorphisms in many individual genes to find evidence of natural selection. In the post-genome era, we can use the whole-genome sequence (WGS), which enable us to do population genetics on a genomic scale. Amino acid changes, expression, gene order, gene number, gene repertoires and so on, have been considered as the target of natural selection (Hurst 2009, Koonin and Wolf 2010, for review).

In this PhD work, in order to look for evidence of natural selection in genome evolution, I focused on whole genome duplication (WGD). WGD of the budding yeasts were first documented by Wolfe and Shields (1997) with a followup of genome sequences of other related species (Kellis, Birren, and Lander 2004). There are two major evolutionary process associated with WGD: gene duplication on whole genome scale (often called Ohnologs, named by Ken Wolfe) and subsequent genome rearrangement with massive gene deletion. WGD has been one of the major focuses in molecular evolution in the post-genome studies (Davis and Petrov 2004, Gao and Innan 2004, Wong and Wolfe 2005, for example).

In chapter 2 and 3, I focused on gene duplication on whole genome scale. In chapter 2, I estimated the duration of concerted evolution via gene conversion of the ohnologs. Concerted evolution is the non-independent evolution of copy members in a multigene family, and interlocus gene conversion is one of the major mechanisms of concerted evolution (Li 1997). The extent of concerted evolution in genome evolution was unclear, meaning that the standard molecular clock theory doesn't work under concerted evolution. However, we found that ohnologs overcome this problem, because they were generated simultaneously (see figure 2.1). Using the evolutionary model of duplicated genes (Teshima and Innan 2004) and maximum likelihood methods, I estimated the duration of concerted evolution via gene conversion in the *S. cerevisiae* ohnologs. I also compared neutral and selection models and examined if they fit the observed data. The neutral model assumed that the expected duration of concerted evolution is same between ohnologs, while the selection model allowed some variation. I found that

the observed distribution of the duration cannot be explained by the neutral model. This raises the possibility that natural selection has influenced on the duration of concerted evolution in ohnologs.

In the next work (chapter 3), I tested some hypotheses for explaining the observed distribution of the duration of concerted evolution. The previous work (chapter 2) suggested that the expectation of the duration of concerted evolution is variable between the ohnologs. In a neutralist's view, the variance is caused by the variation in local gene conversion and mutation rates. In a selectionist's view, natural selection works to favor ohnologs to be homogenized. I first found that local gene conversion and mutation rates cannot explain the data. Then, I examined the possibility of natural selection for "more of the same products". This mode of selection was pointed out by Ohno's seminal book, "Evolution by Gene Duplication"(Ohno 1970). The logic is here. Gene duplication is beneficial for the genes, which high gene expression level is required. Then, sequence divergence would diminish the advantage, because it often causes the change of the gene's function. However, when gene conversion occurred, the genes are homogenized and recover the advantage by gene duplication. See also Innan and Kondrashov (2010). Two genome-wide experimental data supported this hypothesis. The gene expression levels, measured by protein dosage (Nagalakshmi et al. 2008) and codon bias (Sharp and Li 1987), correlated to the duration of concerted evolution (figure 3.3). It is suggested that this mode of selection is a likely explanation for the variation in the duration of concerted evolution between ohnologs.

While I focused on duplicates in the first two chapters, in the following chapter (chapter 4), I alternatively focused on genes lost after WGD, where drastic genome rearrangement associating with gene deletion occurred. With the increasing genome sequence data, we now accept that gene order is not random even in eukaryote genomes, which do not have operon structure except for nematodes (Hurst, Pál, and Lercher 2004). Contrary to increasing evidences for non-random gene order, natural selection on gene order was demonstrated only in few cases (Slot and Rokas 2010, Wong and Wolfe 2005). In chapter 4, I looked for natural selection on gene order after WGD. The process of genome rearrangement

after WGD would be a good opportunity to obtain more advantageous gene order. This process is well summarized in Yeast Gene Order Browser (YGOB; http:// wolfe.gen.tcd.ie /ygob/) by Ken Wolfe and his coallegue (Byrne and Wolfe 2005, Gordon, Byrne, and Wolfe 2009, Scannell et al. 2006, 2007). Using this information, I traced the evolutionary history of each adjacent gene pairs of post-WGD species to elucidate the evolution caused by the WGD. Here, the adjacent pairs that are conserved through WGD is called "conserved" and the pairs of newly generated through WGD is called "new". Compared within these classes and each transcription orientation, I found that the number of new divergent pairs are lower and the distance of new divergent pairs are longer than neutral expectations (table 4.1). These observation suggests that some natural selection disfavors new divergent pairs. Why are they disfavored? I propose that transcription interference would be one of the major causes (Shearwin, Callen, and Egan 2005). If the transcripts of adjacent pairs interfere with each other and disrupt efficient transcription, it is advantageous to keep its partner away. To test my hypothesis, I used nucleosome free region (NFR) as the regulator of gene expression. Empirical data shows that coexpression likley occurs only when there is a single NFR between adjacent pair and that multiple NFRs buffer their coexpression (Xu et al. 2009). I showed that conserved pairs tend to have a single NFR and new pairs tend to have multiple NFRs. Furthermore, I also showed that the number of NFRs explained the observed negative correlations between coexpression level and intergenic distance. From these observations, it is suggested that selection against new divergent gene pairs made a great contribution to the evolution of gene order.

Through my phD work, I focused on two modes of natural selection that have worked on the yeast genome evolution after the WGD event. Both of these two modes are related to gene expression; one is selection for more dosage and the other is selection on the coexpression of adjacent genes, suggesting the importance of of the changes of gene expression in genome evolution. This idea, which was first proposed by King and Wilson (1975), is involved in one of the central controversy in recent molecular evolutionary studies (Carroll 2005). It is argued that evolution caused by the change of coding region. On the other hand, the

change of regulatory region, or change of gene expression levels, is major factor in evolution. Here, I showed that the change of gene expression is highly related to the evolution of *S. cerevisiae*. Selection on concerted evolution for maintaining the homology between ohnologs works to keep the dosage of their products (chapter 3). Selection on adjacent gene pairs to keep their neighbors away works to diminish the interference of their transcripts (chapter 4). These results supports the hypothesis that evolution of gene expression causes species evolution.

5.2 Perspectives

In the pre-genome era, most research focused on a single locus because of the lack of data. The genome data of multiple species and large genome-wide experimental data allow us to survey the locus related to phenotypes (*eg.* genome-wide association studies) and to analyze the interaction of multiple loci (*eg.* epi-genetics) on the genome-wide scale. My work is one of the pionnering works of the post-genome era. Most of new technology has been introduced in the yeast because of its simple system. I used both evolutionary theory and experimental data to estimate the natural selection in the genome evolution. In chapter 3, I used the rate of DSB as the proxy of gene conversion rate, which is hard to estimate empirically. In chapter 4, NFRs were used to represent regulators. Compared to transcription factor and its binding sites, NFRs would be the stable data. I used gene expression level in both studies. These data allow us to analyze how a change in the DNA sequence affects gene expression, gene networks and phenotypes.

I think genome-wide biological data will become more important in the study of genome evolution. In these days, the hottest technology is next-generation sequencing (NGS) (Shendure and Ji 2008). NGS allows us not only to intra-species' genome sequence with low cost, but also to study some important biological features on the whole-genome scale. I used NFR in chapter 4. This data were obtained using NGS by sequencing the DNA which are attached to histones. The chromosome interaction data were also obtained by NGS with the DNA-chip technology (Duan et al. 2010). The problem of the study of genome evolution has been

that it is often unknown what kind of selection is in action even if some evidence of natural selection were found. Until now a good understanding of the nature of selection has been restricted to a small number of well-studied genes. However, genome-wide experimental data would overcome this problem. One example is called integrative analysis (Zhu et al. 2008). Recently, such integrative analysis has been done in species other than yeast (Gerstein et al. 2010, modENCODE Consortium et al. 2010).

In future, a huge amount of data would be generated by NGS and other technology. How do we treat them? One of the answer is evolutionary studies. The approach of population genetics and molecular evolution allow us to identify the evidence of natural selection. The model and statistics of them are very useful to extract biological knowledge from these data. I am looking forward to analyzing the general evolutionary mechanism using population genetics, genome sequence data and experimental data.

Bibliography

Aach J, Rindone W, Church G. 2000. Systematic management and analysis of yeast gene expression data. Genome Res. 10:431–445.

Albert I, Mavrich T, Tomsho L, Qi J, Zanton S, Schuster S, Pugh B. 2007. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. Nature. 446:572–576.

Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, Lipman D. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Arnheim N. 1983. *Concerted evolution of multigene families, pp. 38–61 in* Evolution of Genes and Proteins*, edited by* M. NEI *and* R. K. KOEHN. Sinauer, Sunderland, MA.

Batada N, Urrutia A, Hurst L. 2007. Chromatin remodelling is a major source of coexpression of linked genes in yeast. Trends Genet. 23:480–484.

Beaumont M, Zhang W, Balding D. 2002. Approximate Bayesian computation in population genetics. Genetics. 162:2025–2035.

Benovoy D, Morris R, Morin A, Drouin G. 2005. Ectopic gene conversions increase the G + C content of duplicated yeast and *Arabidopsis* genes. Mol Biol Evol. 22:1865–1868.

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

Brown D, Wensink C, Jordan E. 1972. A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. J Mol Biol. 63:57–73.

Bulik D, Olczak M, Lucero H, Osmond B, Robbins P, Specht C. 2003. Chitin synthesis in *Saccharomyces cerevisiae* in response to supplementation of growth medium with glucosamine and cell wall stress. Eukaryot Cell. 2:886–900.

Butler G, Rasmussen M, Lin M, et al. (51 co-authors). 2009. Evolution of pathogenicity and sexual reproduction in eight Candida genomes. Nature. 459:657–662.

Byrne K, Wolfe K. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. 15:1456–1461.

Byrnes J, Morris G, Li W. 2006. Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. Mol Biol Evol. 23:1136–1143.

Carroll S. 2005. Evolution at two levels: on genes and form. PLoS Biol. 3:e245.

Cherry J, Ball C, Weng S, et al. (11 co-authors). 1997. Genetic and physical maps of *Saccharomyces cerevisiae*. Nature. 387:67–73.

Cho R, Campbell M, Winzeler E, et al. (11 co-authors). 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell. 2:65–73.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen B, Johnston M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science. 301:71–76.

Cohen B, Mitra R, Hughes J, Church G. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet. 26:183–186.

Davis J, Petrov D. 2004. Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol. 2:E55.

Dietrich F, Voegeli S, Brachat S, et al. (14 co-authors). 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. Science. 304:304–307.

Duan Z, Andronescu M, Schutz K, McIlwain S, Kim Y, Lee C, Shendure J, Fields S, Blau C, Noble W. 2010. A three-dimensional model of the yeast genome. Nature. 465:363–367.

Dunham M, Badrane H, Ferea T, Adams J, Brown P, Rosenzweig F, Botstein D. 2002. Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. Proc Natl Acad Sci U S A. 99:16144–16149.

Dunn B, Sherlock G. 2008. Reconstruction of the genome origins and evolution of the hybrid lager yeast *Saccharomyces pastorianus*. Genome Res. 18:1610–1623.

Fry R, Sambandan T, Rha C. 2003. DNA damage and stress transcripts in *Saccharomyces cerevisiae* mutant sgs1. Mech Ageing Dev. 124:839–846.

Fukuoka Y, Inaoka H, Kohane I. 2004. Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. BMC Genomics. 5:4.

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

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

Gasch A, Huang M, Metzner S, Botstein D, Elledge S, Brown P. 2001. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. Mol Biol Cell. 12:2987–3003.

Gasch A, Spellman P, Kao C, Carmel-Harel O, Eisen M, Storz G, Botstein D, Brown P. 2000. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell. 11:4241–4257.

Génolevures Consortium S, JL D, B G, et al. (55 co-authors). 2009. Comparative genomics of protoploid Saccharomycetaceae. Genome Res. 19:1696–1709.

Gerstein M, Lu Z, Van Nostrand E, et al. (131 co-authors). 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science. 330:1775–1787.

Gerton J, DeRisi J, Shroff R, Lichten M, Brown P, Petes T. 2000. Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. Proc Natl Acad Sci U S A. 97:11383–11390.

Ghaemmaghami S, Huh W, Bower K, Howson R, Belle A, Dephoure N, O'Shea E, Weissman J. 2003. Global analysis of protein expression in yeast. Nature. 425:737–741.

Goffeau A, Barrell B, Bussey H, et al. (16 co-authors). 1996. Life with 6000 genes. Science. 274:546, 563–547.

Gordon J, Byrne K, Wolfe K. 2009. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. PLoS Genet. 5:e1000485.

Gu X, Wang Y, Gu J. 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat Genet. 31:205–209.

Hermsen R, ten Wolde P, Teichmann S. 2008. Chance and necessity in chromosomal gene distributions. Trends Genet. 24:216–219.

Herr D, Harris G. 2004. Close head-to-head juxtaposition of genes favors their coordinate regulation in *Drosophila melanogaster*. FEBS Lett. 572:147–153.

Hughes T, Marton M, Jones A, et al. (22 co-authors). 2000. Functional discovery via a compendium of expression profiles. Cell. 102:109–126.

Hurst L. 2009. Fundamental concepts in genetics: genetics and the understanding of selection. Nat Rev Genet. 10:83–93.

Hurst L, Pál C, Lercher M. 2004. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet. 5:299–310.

Hurst L, Williams E, Pál C. 2002. Natural selection promotes the conservation of linkage of co-expressed genes. Trends Genet. 18:604–606.

Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol. 151:389–409.

Innan H. 2002. A method for estimating the mutation, gene conversion and recombination parameters in small multigene families. Genetics. 161:865–872.

Innan H. 2003a. A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. Proc Natl Acad Sci U S A. 100:8793–8798.

Innan H. 2003b. The coalescent and infinite-site model of a small multigene family. Genetics. 163:803–810.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. 11:97–108.

Iyer V, Horak C, Scafe C, Botstein D, Snyder M, Brown P. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature. 409:533–538.

Jukes TH, Cantor DR. 1969. *Evolution of protein molecules, pp. 21–132 in* Mammalian Protein Metabolism*, edited by* H. N. MUNRO. Academic Press, New York.

Kafri R, Bar-Even A, Pilpel Y. 2005. Transcription control reprogramming in genetic backup circuits. Nat Genet. 37:295–299.

Kellis M, Birren B, Lander E. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature. 428:617–624.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander E. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. 423:241–254.

Kensche P, Oti M, Dutilh B, Huynen M. 2008. Conservation of divergent transcription in fungi. Trends Genet. 24:207–211.

Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

King M, Wilson A. 1975. Evolution at two levels in humans and chimpanzees. Science. 188:107–116.

Kondrashov F, Koonin E. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. Trends Genet. 20:287–290.

Koonin E, Wolf Y. 2010. Constraints and plasticity in genome and molecular-phenome evolution. Nat Rev Genet. 11:487–498.

Kreitman M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature. 304:412–417.

Kruglyak S, Tang H. 2000. Regulation of adjacent yeast genes. Trends Genet. 16:109–111.

Lépingle A, Casaregola S, Neuvéglise C, Bon E, Nguyen H, Artiguenave F, Wincker P, Gaillardin C. 2000. Genomic exploration of the hemiascomycetous yeasts: 14. *Debaryomyces hansenii* var. *hansenii*. FEBS Lett. 487:82–86.

Lercher M, Blumenthal T, Hurst L. 2003. Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. Genome Res. 13:238–243.

Lercher M, Urrutia A, Hurst L. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet. 31:180–183.

Li WH. 1997. *Molecular Evolution*. Sinauer, Sunderland, MA.

Li Y, Yu H, Guo Z, Guo T, Tu K, Li Y. 2006. Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. PLoS Comput Biol. 2:e74.

Liao B, Zhang J. 2008. Coexpression of linked genes in Mammalian genomes is generally disadvantageous. Mol Biol Evol. 25:1555–1565.

Lieb J, Liu X, Botstein D, Brown P. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet. 28:327–334.

Lin Y, Byrnes J, Hwang J, Li W. 2006. Codon-usage bias versus gene conversion in the evolution of yeast duplicate genes. Proc Natl Acad Sci U S A. 103:14412–14416.

Liti G, Carter D, Moses A, et al. (26 co-authors). 2009. Population genomics of domestic and wild yeasts. Nature. 458:337–341.

Lynch M, Conery J. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151–1155.

Lynch M, Sung W, Morris K, et al. (11 co-authors). 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A. 105:9272–9277.

Mano S, Innan H. 2008. The evolutionary rate of duplicated genes under concerted evolution. Genetics. 180:493–505.

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

Marjoram P, Molitor J, Plagnol V, Tavare S. 2003. Markov chain Monte Carlo without likelihoods. Proc Natl Acad Sci U S A. 100:15324–15328.

Mavrich T, Ioshikhes I, Venters B, Jiang C, Tomsho L, Qi J, Schuster S, Albert I, Pugh B. 2008. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res. 18:1073–1083.

McLysaght A, Hokamp K, Wolfe K. 2002. Extensive genomic duplication during early chordate evolution. Nat Genet. 31:200–204.

Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, Ito T. 2006. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. Proc Natl Acad Sci U S A. 103:17846–17851.

modENCODE Consortium R, S E, J K, et al. (97 co-authors). 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science. 330:1787–1797.

Mortimer R, Johnston J. 1986. Genealogy of principal strains of the yeast genetic stock center. Genetics. 113:35–43.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 320:1344–1349.

Natarajan K, Meyer M, Jackson B, Slade D, Roberts C, Hinnebusch A, Marton M. 2001. Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. Mol Cell Biol. 21:4347–4368.

Neil H, Malabat C, d'Aubenton Carafa Y, Xu Z, Steinmetz L, Jacquier A. 2009. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature. 457:1038–1042.

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

Ohta T. 1980. *Evolution and Variation of Multigene Families*. Springer-Verlag, Berlin/New York., New York.

Ohta T. 1982. Allelic and nonallelic homology of a supergene family. Proc Natl Acad Sci U S A. 79:3251–3254.

Ohta T. 1989. The mutational load of a multigene family with uniform members. Genet Res. 53:141–145.

Olesen K, Felding T, Gjermansen C, Hansen J. 2002. The dynamics of the *Saccharomyces* carlsbergensis brewing yeast transcriptome during a production-scale lager beer fermentation. FEMS Yeast Res. 2:563–573.

Papp B, Pál C, Hurst L. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature. 424:194–197.

Pérez-Ortín J, Querol A, Puig S, Barrio E. 2002. Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. Genome Res. 12:1533–1539.

Poyatos J, Hurst L. 2007. The determinants of gene order conservation in yeasts. Genome Biol. 8:R233.

Roberts C, Nelson B, Marton M, et al. (13 co-authors). 2000. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science. 287:873–880.

Scannell D, Butler G, Wolfe K. 2007. Yeast genome evolution–the origin of the species. Yeast. 24:929–942.

Scannell D, Byrne K, Gordon J, Wong S, Wolfe K. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature. 440:341–345.

Scannell D, Frank A, Conant G, Byrne K, Woolfit M, Wolfe K. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci U S A. 104:8397–8402.

Schacherer J, Shapiro J, Ruderfer D, Kruglyak L. 2009. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. Nature. 458:342–345.

Sémon M, Duret L. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol. 23:1715–1723.

Sharp P, Li W. 1987. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.

Shearwin K, Callen B, Egan J. 2005. Transcriptional interference–a crash course. Trends Genet. 21:339–345.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat Biotechnol. 26:1135–1145.

Singer G, Lloyd A, Huminiecki L, Wolfe K. 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Mol Biol Evol. 22:767–775.

Slot J, Rokas A. 2010. Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. Proc Natl Acad Sci U S A. 107:10136–10141.

Spellman P, Rubin G. 2002. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. J Biol. 1:5.

Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, Eisen M, Brown P, Botstein D, Futcher B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol Biol Cell. 9:3273–3297.

Steinmetz L, Scharfe C, Deutschbauer A, et al. (11 co-authors). 2002. Systematic screen for human disease genes in yeast. Nat Genet. 31:400–404.

Sugino R, Innan H. 2005. Estimating the time to the whole-genome duplication and the duration of concerted evolution via gene conversion in yeast. Genetics. 171:63–69.

Tajima F. 1992. Statistical method for estimating the standard errors of branch lengths in a phylogenetic tree reconstructed without assuming equal rates of nucleotide substitution among different lineages. Mol Biol Evol. 9:168–181.

Teshima K, Innan H. 2004. The effect of gene conversion on the divergence between duplicated genes. Genetics. 166:1553–1560.

Thompson J, Higgins D, Gibson T. 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.

Trinklein N, Aldred S, Hartman S, Schroeder D, Otillar R, Myers R. 2004. An abundance of bidirectional promoters in the human genome. Genome Res. 14:62–66.

Walsh J. 1987. Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? Genetics. 117:543–557.

Wei W, McCusker J, Hyman R, et al. (22 co-authors). 2007. Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. Proc Natl Acad Sci U S A. 104:12825–12830.

Williams E, Bowles D. 2004. Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Res. 14:1060–1067.

Williams E, Hurst L. 2002. Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes. J Mol Evol. 54:511–518.

Williams R, Primig M, Washburn B, Winzeler E, Bellis M, Sarrauste de Menthiere C, Davis R, Esposito R. 2002. The Ume6 regulon coordinates metabolic and meiotic gene expression in yeast. Proc Natl Acad Sci U S A. 99:13431–13436.

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

Wong S, Wolfe K. 2005. Birth of a metabolic gene cluster in yeast by adaptive gene relocation. Nat Genet. 37:777–782.

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz L. 2009. Bidirectional promoters generate pervasive transcription in yeast. Nature. 457:1033–1037.

Zhu J, Zhang B, Smith E, Drees B, Brem R, Kruglyak L, Bumgarner R, Schadt E. 2008. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet. 40:854–861.

Zimmer E, Martin S, Beverley S, Kan Y, Wilson A. 1980. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. Proc Natl Acad Sci U S A. 77:2158–2162.

Zuckerkandl E, Pauling L. 1965. *Evolutionary divergence and convergence in proteins, pp. 97–166 in* Evolving Genes and Proteins*, edited by* V. BRYSON *and* H. J. VOGEL. Academic Press, New York.