SAGE-Hindawi Access to Research
International Journal of Evolutionary Biology
Volume 2011, Article ID 970768, 11 pages
doi:10.4061/2011/970768

Research Article

Ectopic Gene Conversions in the Genome of Ten Hemiascomycete Yeast Species

Robert T. Morris and Guy Drouin

Département de Biologie et Centre de Recherche Avancée en Génomique Environnementale, Université d'Ottawa, 30 Marie Curie, Ottawa, ON, Canada K1N 6N5

Correspondence should be addressed to Guy Drouin, gdrouin@science.uottawa.ca

Received 21 July 2010; Revised 18 September 2010; Accepted 15 October 2010

Academic Editor: Hiromi Nishida

We characterized ectopic gene conversions in the genome of ten hemiascomycete yeast species. Of the ten species, three diverged prior to the whole genome duplication (WGD) event present in the yeast lineage and seven diverged after it. We analyzed gene conversions from three separate datasets: paralogs from the three pre-WGD species, paralogs from the seven post-WGD species, and common ohnologs from the seven post-WGD species. Gene conversions have similar lengths and frequency and occur between sequences having similar degrees of divergence, in paralogs from pre- and post-WGD species. However, the sequences of ohnologs are both more divergent and less frequently converted than those of paralogs. This likely reflects the fact that ohnologs are more often found on different chromosomes and are evolving under stronger selective pressures than paralogs. Our results also show that ectopic gene conversions tend to occur more frequently between closely linked genes. They also suggest that the mechanisms responsible for the loss of introns in *S. cerevisiae* are probably also involved in the gene 3′-end gene conversion bias observed between the paralogs of this species.

1. Introduction

The repair of double strand DNA breaks is a critical biological process which maintains genome stability. The primary process whereby double-strand DNA breaks are repaired is via homologous recombination; this process requires the use of a repair template gene which provides a copy of the missing information caused by the double-strand DNA breaks. The repair template can either be an allele (allelic recombination) or a paralog (ectopic recombination). An end product of the homologous recombination pathway is the replacement of the broken part of the damaged gene by a homologous portion of the repair template gene. The damaged gene is therefore converted by the template gene (reviewed in [1]).

The factors affecting, and the characteristics of, ectopic and allelic gene conversions have been the focus of many studies, and sequence similarity has been shown to have a profound effect on gene conversion propensity between paralogs. In *Escherichia coli,* a 2%–4% decrease in sequence similarity between a damaged gene and its repair template can cause a 10- to 40-fold decrease in recombination frequency [2, 3]. Similarly, in *Saccharomyces cerevisiae*, larger gene conversions are limited to more similar sequences [4]. Chromosomally linked genes are converted more frequently than dispersed genes in Drosophila and humans [5, 6]. In *S. cerevisiae*, increasing distance between paralogs located on the same chromosome tends to decrease their conversion frequency [4, 7, 8]. In some genomes, different regions of genes are converted at different rates. For example, in *S. cerevisiae*, genes conversions between dispersed paralogs are more frequent at their 3′-ends [4]. This 3′-bias is likely the result of gene conversion with incomplete cDNA molecules [9].

The availability of ten hemiascomycete genomes provides the opportunity to study ectopic gene conversions within a clade with as much sequence divergence as the entire Chordate phylum [10]. The evolution of several hemiascomycetes species was affected by a whole genome duplication event (WGD) which occurred some 150 millions

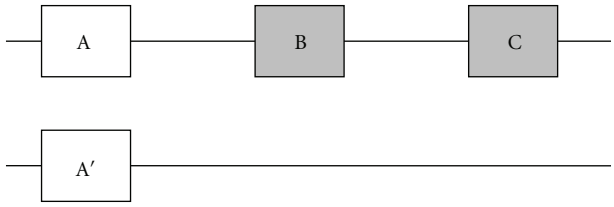

Figure 1: Schematic representation of ohnologs and paralogs. Genes A and A′ represent ohnologs created by a genome duplication. These genes are therefore located on different chromosomes. Genes B and C represent paralogs created by tandem duplications of gene A. These genes are therefore on the same chromosome as gene A.

years ago (MYA; [11–14]). The genomes of *Kluyveromyces lactis*, *Debaryomyces hansenii,* and *Yarrowia lipolytica* all diverged before the whole genome duplication event that occurred in the yeast lineage (pre-WGD species; [10]). The *S. cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces bayanus*, *Saccharomyces kudriavzevii*, *Saccharomyces castellii,* and *Candida glabrata* genomes all diverged after this whole genome duplication event (post-WGD species; [15–17]).

The advantage of separating these genomes into two groups is that we are able to perform two comparisons. The first compares the characteristics of ectopically converted ohnologs and paralogs between the post-WGD species. The post-WGD ohnologs are composed of the duplicated gene pairs that resulted from the whole genome duplication [11, 18]. The post-WGD paralogs data set is composed of the genes from multigene families containing at least three members in the genome of the seven post-WGD species but excluding all ohnologs (Figure 1). The second comparison involves the contrast of the characteristics of ectopically converted paralogs between pre- and post-WGD species. The pre-WGD paralogs data set is composed of the genes from multigene families containing at least three members in the genome of the three pre-WGD species.

The previous studies have shown that the reason why many ohnologs are still found in yeast genomes is because they provide a selective advantage [19, 20]. Ohnologs are maintained by selection either because they carry out a subset of the functions that were previously assumed by their preduplication ancestor (subfunctionalization), assume new functions (neofunctionalization), or provide increased gene product dosage. We therefore expect that most ectopic gene conversions between ohnolog genes will be deleterious and removed by selection. If so, ectopic gene conversions between ohnologs should be less frequent than those between paralogs. In addition, based on the previous studies, we expect that gene conversion frequency should decrease as the distance between related genes increases (and be least frequent for genes situated on different chromosomes), that the length of gene conversion tracts should be positively correlated with sequence similarity and that converted regions should be more frequent at the 3′-end of genes [4].

2. Materials and Methods

2.1. Genome Sequences. The *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. bayanus*, *S. kudriavzevii*, and *S. castellii* genome sequences were retrieved from the Saccharomyces Genome Database (SGD; ftp://genome-ftp.stanford.edu/pub/yeast/sequence/). The *C. glabrata*, *K. lactis*, *D. hansenii*, and *Y. lipolytica* genome sequences and distance files (*.ptt files) were retrieved from the NCBI ftp website (ftp://ftp.ncbi.nih.gov/).

2.2. Gene Family Data Sets. We used three different data sets of protein coding genes. To retrieve the post-WGD ohnologs from the seven post-WGD species, we used the 551 *S. cerevisiae* duplicated gene pairs (1102 ohnologs) identified by Byrne and Wolfe [21] as queries. Sequences from *C. glabrata* and *S. castellii* were retrieved using the Yeast Gene Order Browser (http://wolfe.gen.tcd.ie/ygob/), and those from the other 4 species were retrieved from the Saccharomyces Genome Database (ftp://genome-ftp.stanford.edu/pub/yeast/sequence/fungal_genomes/Multiple_species_align/other/fungalAlignCorrespondance.txt). Our data set of ohnologs in post-WGD species is therefore only composed of the ohnologs pairs also found in *S. cerevisiae*. We used this subset of ohnologs because the efficient detection of gene conversion events using the GENECONV method requires that at least three sequences be available [4]. To detect gene conversions in ohnologs, we therefore needed ohnologs from at least two species and we used the ohnologs of *S. cerevisiae* to retrieve ohnologs pairs from the other 6 post-WGD species. Retrieving common ohnologs also allowed us to study gene conversions between similar genes in seven different genomes.

The post-WGD paralog data set was constructed using the BLASTCLUST program available at the NCBI FTP site. Gene families were defined as being composed of sequences having at least 60% amino acid identity over at least 50% of their length. If genes previously identified as ohnologs were grouped into paralog multigene families, then these genes were removed from the family to ensure that there was no redundancy between the ohnolog and paralog data sets (see Figure 1). The pre-WGD paralog data set was also constructed using the BLASTCLUST program, and gene families were also defined as being composed of sequences having at least 60% amino acid identity over at least 50% of their length.

2.3. Sequence Alignments and Gene Conversion Detection. ClustalW was used to align the protein sequences of multigene families' members [22]. DNA sequences were then fitted to the protein alignments using a PERL script.

Gene conversions were detected using the GENECONV method [23]. Redundant gene conversions within a multigene family were detected by examining the phylogenetic tree of each family and removed from the analysis [4]. If the same gene conversion was detected at the same location in the multigene family alignment in closely related descendents of a common ancestor then the most parsimonious

Table 1: Number of ohnologs and paralogs in the pre- and post-WGD genomes.

Genome	Number of ohnolog families	Number of paralog families
Post-WGD		
S. cerevisiae	551 (2)	30 (3–40)
S. paradoxus	436 (2)	80 (3–68)
S. mikatae	412 (2)	86 (3–37)
S. kudriavzevii	226 (2)	13 (3–20)
S. bayanus	462 (2)	75 (3–23)
C. glabrata	300 (2)	16 (3–7)
S. castellii	398 (2)	17 (3–10)
Pre-WGD		
K. lactis	n.a.	15 (3–9)
D. hansenii	n.a.	43 (3–9)
Y. lipolytica	n.a.	60 (3–26)

Notes. The range of multigene family sizes is provided in brackets. n.a.: not applicable.

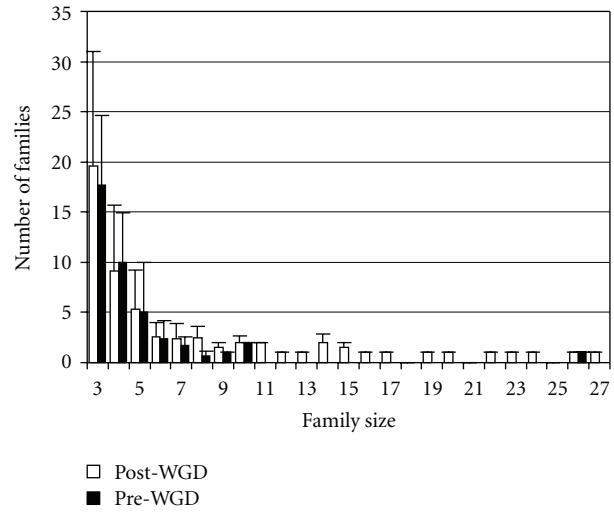

Figure 2: The distribution of the average number of paralog gene families (mean ± S.D.) within the seven postduplication genomes and three preduplication genomes is shown. Five outlier families including two families of size 63 and 68 from *S. paradoxus*, two families of size 32 and 40 from *S. cerevisiae*, and a single family of 38 genes from *S. mikatae* are not shown in the figure to improve the visual clarity of the data.

explanation is that the conversion event occurred within the common ancestor, therefore only one of the conversions detected in the set of descendents was retained for further analysis. To control for false positives, gene conversions between sequences having less than 80% maximum flanking similarity were removed from the analysis [24].

2.4. Gene Conversion Characteristics. The gene conversion frequency for each species was calculated using two different methods. The first method calculates the conversion frequency as the ratio of the number of conversions divided by the total number of gene comparisons between multigene family members. The second method calculates the frequency as the ratio of the number of gene conversions divided by the total number of multigene family members. Intra- and interchromosomal gene conversion frequencies were calculated for the *S. cerevisiae*, *C. glabrata* ohnolog and paralog multigene families. In addition intra- and interchromosomal conversion frequencies were calculated for the paralog multigene families of *K. lactis*, *D. hansenii,* and *Y. lipolytica* genomes. These frequencies are calculated as the ratio of intra- (or inter-) chromosomal conversions divided by the total number of intra- (or inter-) chromosomal gene comparisons. The gene conversion length was obtained from the GENECONV output. The maximum similarity for the flanking 100 nucleotides was calculated for each converted gene pair using an in-house PERL script. The locations of the converted regions were calculated as the correlation between the positions of each conversion with respect to the length of the converted genes. A positive correlation indicates a bias towards the $3'$-end of genes, and a negative correlation indicates a bias towards the $5'$-end of genes. The distance between converted genes was calculated only for conversions detected within *S. cerevisiae*, *C. glabrata*, *K. lactis, D. hansenii*, and *Y. lipolytica* because position data for the other five species was not available. Data tabulation and analysis was done using Microsoft Excel (Microsoft, Redmond, WA, USA) and S-plus v 7.0 (Insightful, Seattle, WA, USA). The G-Power program was used to calculate the power of the ANOVA tests [25]. Power calculations for correlation tests were done using an online application (http://calculators.stat.ucla.edu/powercalc/correlation/) and SAS 9.1.3 (SAS Institute Inc., Cary, NC, USA).

2.5. Numbers of Substitutions per Site and Gene Ontology. The number of nonsynonymous substitutions per nonsynonymous site (Ka) and synonymous substitutions per synonymous site (Ks) and their ratio (Ka/Ks) were calculated for the protein coding regions (excluding the converted regions) of each pair of converted genes using the YN00 program from the PAML software [26, 27].

The processes in which the *S. cerevisiae* ohnologs and paralogs are involved were analyzed using the gene ontology annotations of the Saccharomyces Genome Database at http://db.yeastgenome.org/cgi-bin/GO/goTermFinder.pl [28].

3. Results

3.1. Ohnolog and Paralog Multigene Families. Ohnolog and paralog multigene families were analyzed to determine whether the number and size of these two types of families were different in different yeast genomes (Table 1, Figure 2). The genomes of the six post-WGD species from which we retrieved ohnolog pairs using the *S. cerevisiae* ohnologs contain an average of 372 ± 90 ohnolog pairs. The number of ohnolog pairs found in each of these six different genomes is not significantly different from this average when using a

Table 2: Percentage of gene comparisons between multigene family members located on the same chromosome.

Genome	Ohnologs	Paralogs
Post-WGD		
S. cerevisiae	4.0% (22/551)	8.4% (163/1930)
C. glabrata	5.0% (15/300)	38.6% (29/75)
Pre-WGD		
K. lactis	n.a.	21% (26/124)
D. hansenii	n.a.	31% (86/270)
Y. lipolytica	n.a.	18% (158/884)

Notes. The ratios in brackets are the number of gene comparisons between genes found on the same chromosome divided by the total number of gene comparisons. n.a.: not applicable.

Bonferroni-corrected α-value of 0.0083 (Wilcoxon rank sum test; [29]).

For post-WGD paralogs, only the *S. mikatae* genome has significantly more paralog families than average (45.28 ± 33.36; Wilcoxon rank sum test, $P = 0.009$) and only the *S. kudriavzevii* genome has significantly fewer paralog families than average ($P = 0.009$). The mean size of the paralog families (5.7 ± 5.2 genes/family) is similar in all post-WGD genomes except that of *C. glabrata* which has significantly smaller paralog families than average (3.3 ± 0.99 genes/family, Wilcoxon rank sum test, $P = 0.003$).

For the pre-WGD paralogs, the numbers of paralog families in the three pre-WGD genomes are not significantly different from the population mean (39.33 ± 22.72; Wilcoxon rank sum test, $P \geq 0.27$). The mean size of all paralog families in these three genomes (4.41 ± 2.59 paralogs per family) is similar to the mean family size of each pre-WGD genome (Wilcoxon rank sum test, $P \geq 0.09$). Finally, there is no statistical difference between the number (Wilcoxon rank sum test, $P = 0.83$) and the mean size of paralog families (Wilcoxon rank sum test, $P = 0.17$) or between pre- and post-WGD species.

3.2. Organization of Gene Families. The organization of the multigene families can be measured as the proportion of multigene family members located on the same chromosome. Since most paralogs originate from unequal crossover events, they are expected to be most often found on the same chromosome. In contrast, since ohnologs are remnants of ancient genome duplication events, they are expected to be most often found on different chromosomes. The higher percentage of paralogs found on the same chromosome is therefore consistent with the likely mode of origin of these two types of duplicated genes (Table 2). The percentage of paralogs found on the same chromosome is also similar between pre- and post-WGD genomes (Table 2).

3.3. Gene Conversion Frequency and Distance between Converted Genes. In post-WGD genomes, intrachromosomal gene conversions tend to occur more frequently than interchromosomal conversions. In the paralog families of *S. cerevisiae* and *C. glabrata*, genes located on the same chromosome are converted 2 to 10 times more frequently than genes found on different chromosomes (Table 3). Similarly, in the ohnolog families of *S. cerevisiae*, genes located on the same chromosome are converted 4 times more frequently than genes found on different chromosomes (Table 3). In contrast, there is an almost complete absence of gene conversions between the ohnologs found within the *C. glabrata* genome (Table 3).

In pre-WGD genomes, the paralogs found on the same chromosomes of *K. lactis* and *Y. lipolytica* are not converted more frequently than paralogs found on different chromosomes but the *D. hansenii* paralogs found on the same chromosomes are converted roughly 3 times more frequently than those found on different chromosomes (Table 3).

The mean number ($\pm$S.D.) of conversions detected within the paralog gene families of the pre- (38 ± 33) and post-WGD (30 ± 16) genomes is not statistically different (t-test, $P = 0.67$; Table 4). Although the ohnolog families of post-WGD genomes contain only an average of 7 ± 5 conversions, this number is also not significantly different from the average number of conversions found in post-WGD paralog families (t-test, $P = 0.06$).

When considering gene conversion frequencies with respect to the total number of comparisons, gene conversions of post-WGD species are either equally frequent in paralog and ohnolog families (in the *S. paradoxus*, *S. mikatae,* and *S. bayanus* genomes) or significantly more frequent in paralog than in ohnolog families in the four other post-WGD families (t-test, $P = 0.046$; Table 4).

When considering gene conversion frequencies with respect to the total number of multigene family members, the mean conversion frequency for paralogs ($19.03 \pm 16.29\%$) is significantly larger than the frequency for ohnologs (0.74 ± 0.46; Wilcoxon two sample test, $P = 0.0006$).

We believe that using gene conversion frequencies with respect to the total number of multigene family members is more appropriate to compare gene frequencies between ohnologs and paralogs because it better reflects the much larger number of conversions found in paralogs when compared to ohnologs. For example, in the case of *S. cerevisiae* with 13 conversions between ohnologs and 110 conversion between paralogs (Table 4), the conversion frequency for ohnologs is 2.35% (13/551) and 5.71% for paralogs (110/1930) when frequencies are calculated with respect to the total number of comparisons. However, these frequencies do not take into account the fact that 1102 ohnolog sequences were compared (551 pairs) whereas only 212 paralog sequences (i.e., less than the fifth of the number of ohnolog sequences) were compared (for a total of 1930 pairwise comparisons) to obtain the 5.71% frequency of paralogs. In contrast, if one compares the frequencies calculated with respect with the number of genes, the frequency of conversions is 1.17% (13/1102) for ohnologs and 51.40% for paralogs (110/212). The large difference between the two ways of calculating frequencies is due to the fact that frequencies calculated with respect to the total number of comparisons have a much larger denominator which biases the comparisons between ohnologs and paralogs. For example, for a family with 10 paralogous sequences, the number of pairwise comparisons

Table 3: Intra- and interchromosomal gene conversion frequencies for pre- and post-WGD genomes.

Genome	Ohnologs		Paralogs	
Post-WGD	Intrachromosomal frequency	Interchromosomal frequency	Intrachromosomal frequency	Interchromosomal frequency
S. cerevisiae	9.1% (2/22)	2.1% (11/529)	9.2% (15/163)	5.4% (95/1767)
C. glabrata	0% (0/15)	0.007% (2/285)	24.1% (7/29)	2.2% (1/46)
Pre-WGD				
K. lactis	n.a.	n.a.	11.5% (3/26)	12.2% (12/98)
D. hansenii	n.a.	n.a.	36% (31/86)	11.4% (21/184)
Y. lipolytica	n.a.	n.a.	1.9% (3/158)	2.9% (21/726)

Notes. Values in brackets indicate the ratio of the number of gene conversions divided by the number of gene comparisons. Data for *S. paradoxus, S. mikatae, S. kudriavzevii, S. bayanus,* and *S. castellii* are not provided because position data was not available for the genes of these genomes. n.a.: not applicable.

Table 4: The number and frequency of gene conversions in ohnologs and paralogs.

Genomes	Ohnologs			Paralogs		
	Number	Frequency (%) with respect to total number of comparisons	Frequency (%) with respect to total number of multigene family members	Number	Frequency (%) with respect to total number of comparisons	Frequency (%) with respect to total number of multigene family members
Post-WGD						
S. cerevisiae	13	2.35	1.17	110	5.71	51.40
S. paradoxus	7	1.60	0.80	44	1.54	9.20
S. mikatae	6	1.45	0.73	26	1.50	4.80
S. kudriavzevii	2	0.88	0.44	20	7.96	29.80
S. bayanus	14	3.03	1.51	50	3.60	12.40
C. glabrata	2	0.67	0.33	8	10.67	14.80
S. castellii	2	0.50	0.25	8	5.06	10.80
Pre-WGD						
K. lactis	n.a.	n.a.	n.a.	15	12.09	23.80
D. hansenii	n.a.	n.a.	n.a.	52	19.25	31.70
Y. lipolytica	n.a.	n.a.	n.a.	24	2.71	8.20

Notes. n.a.: not applicable.

will be 45 ([10(10−1)]/2) whereas it will only be 5 for 10 ohnologs.

Ectopic gene conversions between paralogs are equally frequent in both pre- and post-WGD genomes. Median gene conversion frequencies relative to both total number of comparisons and number of multigene family members are not statistically different between pre-WGD (12.09%, 23.8%) and post-WGD (5.06%, 12.4%) paralogs (Table 4; Wilcoxon two sample test, $P = 0.26$ with respect to the number of gene comparisons and $P = 0.82$ with respect to the number of multigene family members).

There is a significant negative correlation (Spearman rank correlation test) between gene conversion frequency and distance between paralogs located on the same chromosomes in the genomes of S. *cerevisiae* ($r = -0.54$; $P = 0.008$), *C. glabrata* ($r = -0.74$; $P = 0.048$), and *D. hansenii* ($r = -0.45$; $P = 0.008$). Correlations could not be calculated for the other paralog and/or ohnolog data sets either because gene distance information was not available for some species (see above) or because less than four genes were found on the same chromosomes (statistical power analyses require at least 4 data points).

3.4. Gene Conversion Length and Flanking Similarity. The median lengths of the gene conversions between ohnologs are identical in all seven post-WGD genomes (Table 5; multiple comparison ANOVA test, $P = 0.86$, $\alpha = 0.05$). The median lengths of gene conversions between the paralogs of pre-WGD genomes are also equal ($P = 0.34$). However, the median length of the gene conversions between paralogs are significantly longer in *S. cerevisiae* than in *S. paradoxus*, *S. mikatae*, *S. kudriavzevii,* and *S. bayanus* (multiple comparison ANOVA, $P < 0.0001$). In post-WGD genomes, the median length of gene conversion in paralogs and ohnolog (182 and 186.5 bp, resp.) are not significantly different (pairwise Wilcoxon rank tests, Table 5). Finally, the median lengths of gene conversions are significantly different from each other between pre-WGD (150 bp) and post-WGD (182 bp) paralogs (Wilcoxon two sample test, $P = 0.02$,

Table 5: Gene conversion lengths of pre- and post-WGD species.

Genome	Ohnologs (bp)					Paralogs (bp)					Wilcoxon test
	Median	1st quartile	3rd quartile	Min	Max	Median	1st quartile	3rd quartile	Min	Max	*P*-value
Post-WGD											
S. cerevisiae	272	107	465	60	773	382.5	141	869	8	2642	0.22
S. paradoxus	235	98	354	50	531	106	51.5	232	14	1060	0.17
S. mikatae	165.5	95	431	68	568	167	83	366	14	535	0.64
S. kudriavzevii	270.5	146	395	146	395	136	85	172	25	391	0.19
S. bayanus	149.5	71	315	45	905	126	76	203	21	724	0.50
C. glabrata	83.5	27	140	27	140	130	83.5	386	59	668	0.36
S. castellii	144	118	170	118	170	226	73.5	581.5	44	862	0.69
Pre-WGD											
K. lactis	n.a.	n.a.	n.a.	n.a.	n.a.	99	40	236	32	1127	n.a.
D. hansenii	n.a.	n.a.	n.a.	n.a.	n.a.	183	104.5	310.5	18	1309	n.a.
Y. lipolytica	n.a.	n.a.	n.a.	n.a.	n.a.	83	27.5	196	16	1770	n.a.

Note. Wilcoxon two-sample tests were used to detect differences between the median gene conversion lengths of ohnologs and paralogs. n.a.: not applicable.

Table 6: Maximum flanking similarity of gene conversions in pre and post-WGD species.

Genome	Ohnolog maximum flanking similarity (%)					Paralog maximum flanking similarity (%)					Wilcoxon test
	Median	1st quartile	3rd quartile	Min	Max	Median	1st quartile	3rd quartile	Min	Max	*P*-value
Post-WGD											
S. cerevisiae	88	84	94	80	97	95.6	91	99	80	100	0.001
S. paradoxus	89	83	94	82	97	90.3	87	97.5	80	100	0.24
S. mikatae	87.5	82	92	81	96	91.7	86.6	95.6	81	100	0.15
S. kudriavzevii	86.8	85.7	88	85.7	88	94	93	97	85	99	0.07
S. bayanus	87.6	85	92	80	98	92.9	86	99	81	100	0.04
C. glabrata	84.5	83	86	83	86	92.6	90	99.5	86	100	0.08
S. castellii	87	86	88	86	88	93	87	97	85	100	0.35
Pre-WGD											
K. lactis	n.a.	n.a.	n.a.	n.a.	n.a.	90	86	97	81	98	n.a.
D. hansenii	n.a.	n.a.	n.a.	n.a.	n.a.	93	86.3	97	80	100	n.a.
Y. lipolytica	n.a.	n.a.	n.a.	n.a.	n.a.	86.5	83.3	93.5	80	100	n.a.

Note. Wilcoxon two sample tests were used to detected differences between the median flanking similarities of ohnologs and paralogs. n.a.: not applicable.

$\alpha = 0.05$). These median lengths are similar to the average length of the *S. cerevisiae* conversions observed in a previous study (173 bp, [4]).

The median sequence similarities of regions flanking gene conversions between ohnologs are equal in all seven post-WGD genomes (Table 6; multiple ANOVA tests, $P = 0.97$, $\alpha = 0.05$). Furthermore, the median sequence similarities of regions flanking gene conversions between paralogs are equal in all seven genomes (multiple comparison ANOVA test, $P = 0.18$, $\alpha = 0.05$).

Although the median flanking similarity of the converted paralogs of post-WGD species is always higher than that of their ohnologs, this difference is only significant in the genome of *S. cerevisiae* and *S. bayanus* (Table 6). However, this lack of statistical significance is likely the result of the relatively low power of these statistical tests because the power of each test was ≤61% (results not shown).

The median sequence similarities of regions flanking gene conversions between the paralogs of pre-WGD genomes are equal (Table 6; multiple ANOVA tests, $P = 0.21$, $\alpha = 0.05$). However, converted genes within pre-WGD paralogs have significantly less flanking similarity (pooled median of 90%) than converted paralogs in post-WGD genomes (pooled median of 94%; Wilcoxon two sample test, $P = 0.0004$, $\alpha = 0.05$, Table 6). We do not know whether this difference has any biological significance.

Analysis of the relationship between the length of gene conversions and flanking similarity indicates a significant positive correlation within the ohnologs of the seven post-WGD genomes (Spearman rank correlation test, $r = 0.44$, $P = 0.005$; Figure 3(a)), the paralogs of the seven post-WGD genomes ($r = 0.36$, $P = 0$; Figure 3(b)) and the paralogs of the three pre-WGD genomes ($r = 0.35$, $P = 0$; Figure 3(c)).

3.5. Ka, Ks, Ka/Ks Ratios and Ontology of Ohnolog and Paralog Converted Genes. In post-WGD genomes, the fact that synonymous substitutions (Ks) are lower for converted paralogs than for converted ohnologs suggests that paralogs have a more recent origin (Table 7). Therefore, the higher Ka/Ks ratio of paralogs clearly indicates that paralogs are under less selection constraints than ohnologs. Furthermore, the similar Ka/Ks ratios of pre- and post-WGD paralogs indicate that the paralogs of pre- and post-WGD evolve under similar selective constraints (Table 7).

The ohnologs and paralogs of *S. cerevisiae* are involved in different processes. Although many of the GO terms shown in Table 8 are not mutually exclusive (e.g., "transposition" and "transposition, RNA-mediated"), analyses of the processes in which these genes are involved show that ohnologs are involved in regulation, essential biosynthetic processes, and metabolic processes whereas paralogs are involved transposition, transport, and nonessential biosynthetic processes.

3.6. Location of Converted Regions. When considering pre-WGD paralogs, post-WGD paralogs and post-WGD ohnologs, only the post-WGD paralogs of *S. cerevisiae* show a significant bias of gene conversions towards the 3′-end of genes (Table 9). However, the fact that the power of all nonsignificant tests is smaller than 15% suggests that this bias might also exist in the data sets where it was not detected but that our data are not sufficient to detect it (Table 9).

4. Discussion

Using *S. cerevisiae* ohnologs as queries allowed us to retrieve an average of 372 ohnolog pairs from the other six post-WGD genomes (Table 1). Although these seven species are phylogenetically related (see [13] for a phylogenetic tree of these fungi species), and therefore did not evolve independently, it is very unlikely that species as divergent as *S. cerevisiae* and *C. glabrata* (which diverged soon after the whole genome duplication, some 150 MYA), would have kept 300 pairs of common ohnologs by chance. In fact, assuming that the ancestral pre-WGD genome had 5000 genes and that current post-WGD genomes have 5500 genes [13], one would expect them to have kept only 50 ohnologs in common ($0.1 \times 0.1 \times 5000$) by chance alone. As we discuss further below, this suggests that common ohnologs provide a selective advantage and evolve under strong selective constraints.

Since the number and the mean size of paralog multigene families are not significantly different between pre- and post-WGD species, the genome duplication event in the post-WGD genome ancestor did not significantly increase the number or mean size of paralog multigene families in post-WGD species (Table 1, Figure 2). The small number and size of gene families in *C. glabrata* have already been noticed and are likely the result of reductive evolution and gene loss through relatively high genome instability [10, 12, 30].

The chromosomal distribution of ohnologs and paralogs is very different. Whereas, on average, 23.4% of paralogs are found on the same chromosomes, only 4.5% of ohnologs are

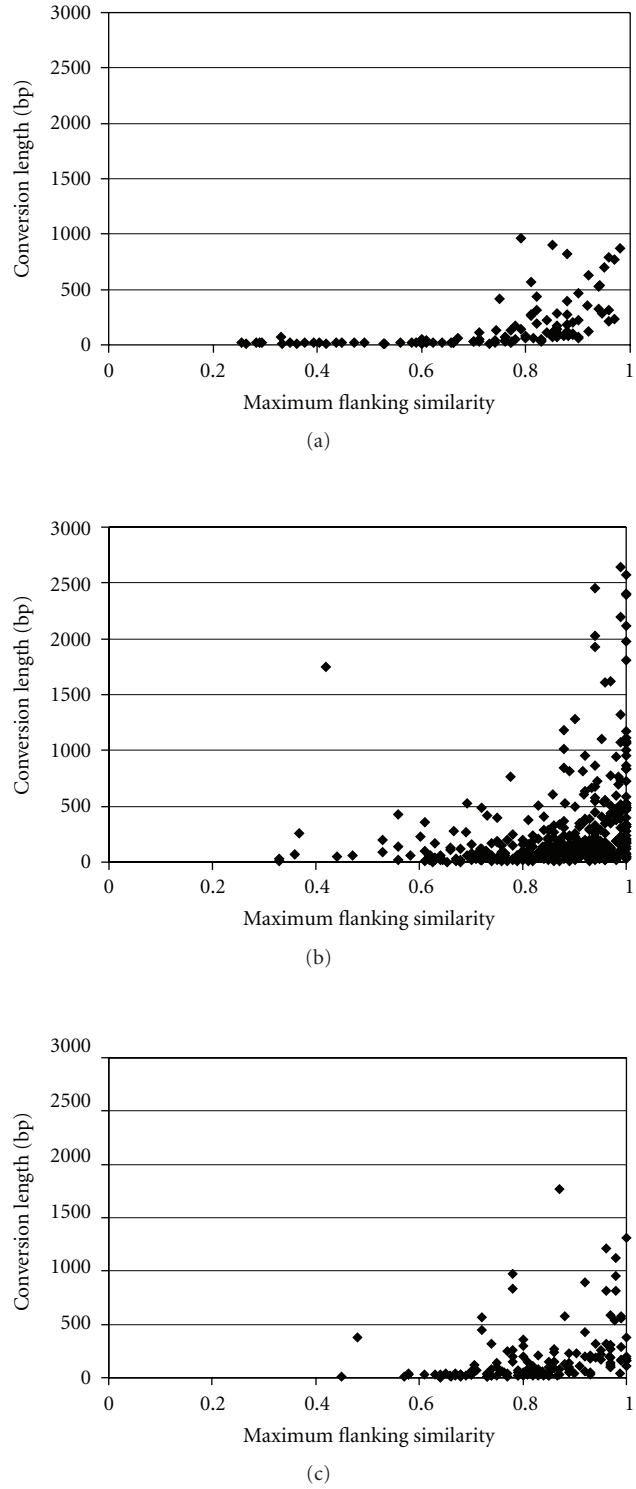

Figure 3: Correlation between gene conversion length and maximum flanking sequence similarity. (a) Conversions detected between the ohnologs of the six Saccharomyces species and *C. glabrata*. There are 107 conversions, 46 of which have ≥80% flanking similarity. (b) Conversions detected between the paralogs of the six Saccharomyces species and *C. glabrata*. There are 401 conversions, 311 of which have ≥80% flanking similarity. (c) Conversions detected the paralogs of the three pre-WGD genomes. There are 147 conversions, 91 of which have ≥80% flanking similarity.

Table 7: Nonsynonymous substitutions per nonsynonymous site (Ka), synonymous substitutions per synonymous site (Ks), and Ka/Ks ratios (± standard deviations) for pairs of converted genes in pre- and post-WGD species.

Genome	Ka		Ks		Ka/Ks	
	Ohnologs	Paralogs	Ohnologs	Paralogs	Ohnologs	Paralogs
Post-WGD						
S. cerevisiae	0.04 ± 0.03	0.09 ± 0.08	0.96 ± 0.49	0.37 ± 0.44	0.04 ± 0.02	0.38 ± 0.27
S. paradoxus	0.09 ± 0.11	0.18 ± 0.20	0.91 ± 0.76	0.56 ± 0.40	0.10 ± 0.05	0.46 ± 0.57
S. mikatae	0.09 ± 0.11	0.17 ± 0.19	1.87 ± 1.06	0.56 ± 0.31	0.04 ± 0.04	0.34 ± 0.45
S. kudriavzevii	0.06 ± 0.04	0.08 ± 0.04	0.95 ± 0.46	0.47 ± 0.59	0.06 ± 0.01	0.38 ± 0.34
S. bayanus	0.11 ± 0.09	0.13 ± 0.12	1.91 ± 1.68	0.40 ± 0.46	0.07 ± 0.05	0.40 ± 0.28
C. glabrata	0.25 ± 0.17	0.04 ± 0.04	1.32 ± 0.08	0.36 ± 0.55	0.18 ± 0.12	0.37 ± 0.45
S. castellii	0.18 ± 0.09	0.13 ± 0.07	2.80 ± 1.18	0.29 ± 0.12	0.06 ± 0.01	0.61 ± 0.46
Pre WGD						
K. lactis	n.a.	0.20 ± 0.26	n.a.	0.61 ± 0.40	n.a.	0.49 ± 0.58
D. hansenii	n.a.	0.10 ± 0.07	n.a.	0.50 ± 0.40	n.a.	0.31 ± 0.17
Y. lipolytica	n.a.	0.25 ± 0.19	n.a.	1.12 ± 0.38	n.a.	0.46 ± 0.38

n.a.: not applicable.

Table 8: GO terms associated with biological processes for the ohnologs and paralogs of *S. cerevisiae*.

GO term	Cluster frequency	Background frequency	*P*-value
Ohnologs			
Biological regulation	21.9%	13.8%	5.2×10^{-13}
Regulation of biological process	18.0%	11.3%	3.9×10^{-10}
Regulation of cellular process	16.8%	10.5%	2.6×10^{-09}
External encapsulating structure organization and biogenesis	6.4%	2.8%	6.8×10^{-09}
Cell wall organization and biogenesis	6.4%	2.8%	6.8×10^{-09}
Protein amino acid phosphorylation	4.0%	1.4%	2.1×10^{-08}
Cellular polysaccharide biosynthetic process	2.0%	0.5%	4.1×10^{-08}
Polysaccharide biosynthetic process	2.0%	0.5%	9.9×10^{-08}
Carbohydrate biosynthetic process	2.7%	0.9%	9.3×10^{-07}
Cellular carbohydrate metabolic process	5.2%	2.3%	1.0×10^{-06}
Carbohydrate metabolic process	5.5%	2.5%	1.7×10^{-06}
Paralogs			
Transposition	32.8%	1.3%	9.7×10^{-109}
Transposition, RNA-mediated	32.8%	1.3%	9.7×10^{-109}
Carbohydrate transport	5.2%	0.5%	9.5×10^{-09}
Monosaccharide transport	4.0%	0.3%	4.0×10^{-07}
Hexose transport	4.0%	0.3%	4.0×10^{-07}
Thiamin and derivative metabolic process	3.2%	0.3%	4.0×10^{-05}
Thiamin biosynthetic process	2.8%	0.2%	2.0×10^{-4}
Thiamin and derivative biosynthetic process	2.8%	0.3%	3.1×10^{-4}
Thiamin metabolic process	2.8%	0.3%	3.1×10^{-4}
Telomere maintenance via recombination	2.8%	0.3%	4.8×10^{-4}
Amino acid catabolic process	3.6%	0.5%	1.0×10^{-3}
Cellular response to nitrogen levels	1.6%	0.1%	1.6×10^{-3}

Notes. Frequencies were calculated from 1100 ohnologs, 250 paralogs, and 7159 background genes. Only the twelve most significant results for each type of genes are shown.

Table 9: Correlations between the location of the converted regions and their position in the converted genes in pre- and post-WGD genomes.

Genome	Ohnolog		Paralog	
	R-value	Power	R-value	Power
Post WGD				
S. cerevisiae	−0.07	0.036	0.73*	n.a.
S. paradoxus	0.12	0.049	−0.19	0.072
S. mikatae	0.00	0.025	−0.19	0.076
S. kudriavzevii	−0.17	0.065	−0.09	0.043
S. bayanus	0.24	0.095	0.11	0.047
C. glabrata	0.00	0.025	0.06	0.034
S. castellii	0.17	0.066	−0.09	0.043
Pre WGD				
K. lactis	n.a.	n.a.	−0.32	0.14
D. hansenii	n.a.	n.a.	0.02	0.028
Y. lipolytica	n.a.	n.a.	0.14	0.055

The R-values indicate correlation values. Significant correlations (Spearman rank correlation test $P < 0.05$) are labeled with *. The power of each correlation test is provided except for *S. cerevisiae* paralogs, where the null hypothesis was rejected, and for ohnologs for which a power test could not be performed. n.a.: not applicable.

found on the same chromosomes (Table 2). A likely explanation for this difference is that paralogs often originate from unequal crossing over or replication slippage events whereas ohnologs originate from whole genome duplication events (page 250 of [31], [18], pages 199–202 of [32]). Since gene conversions tend to be more frequent between genes found on the same chromosomes than between genes located on different chromosomes (Table 3), this explains, in part, why gene conversions tend to be more frequent between paralogs than between ohnologs (Table 4). In fact, on average, when comparing gene conversion using total numbers, frequency calculated using the number of multigene family members, or frequency based on the number of gene comparisons, gene conversions are more frequent in the paralogs of pre- and post-WGD genes than in the ohnologs of the post-WGD genomes (Table 4).

The previous work on yeast, Drosophila, and humans has shown that intrachromosomal gene conversions are more frequent than interchromosomal gene conversions [4–6]. A possible explanation for the relatively high frequency of intrachromosomal conversions in *D. hansenii* (36%, Table 3) is that multiple tandem duplication events have been identified within this genome and, therefore, most paralogs are still located on the same chromosomes [10]. In contrast, in *K. lactis* and *Y. lipolytica*, gene conversions between intra- and interchromosomal paralogs are equally frequent (Table 3). The highly redundant *Y. lipolytica* genome has been shown to be undergone a high degree of map dispersion [10]. The low frequency of intrachromosomal conversions observed in this genome might therefore be the result of the dispersion of tandemly duplicated paralogs to other chromosomes. A similar phenomenon might be present in *K. lactis*. It is unlikely that these exceptions are due to mechanistic differences in the repair of double-stranded-breaks between pre- and post-WGD species because the majority of repair genes have been maintained throughout the evolution of the hemiascomycetes [33].

The previous studies have demonstrated a negative correlation between gene conversion frequency and physical distance on the same chromosome [4, 7]. We also observed such a negative correlation in the genomes of *S. cerevisiae*, *C. glabrata,* and *D. hansenii* (see above). However, a lack of data (statistical power) prevented the detection of such a relationship in the paralogs of *K. lactis* and *Y. lipolytica* and the ohnologs of *S. cerevisiae* and *C. glabrata*. This correlation could result from the fact that the DNA repair mechanisms preferentially search for suitable repair templates close to the damaged gene. Since ohnologs are more often found on different chromosomes (Table 2), this would also explain why conversions are less frequent between ohnologs than between paralogs. On the other hand, our recent analyses of the human genome [6] has shown that, in the human genome, the negative correlation between gene conversion frequency and physical distance is simply the result of the fact that most duplicated genes are found next to one another. Thus the negative correlation we observed in some yeast species might also disappear if we normalized our data to take into account the fact that most paralogs are located next to one another on the same chromosome [10].

Sequence similarity requirements for ectopic conversions and the amount of negative selection are very similar between pre- and post-WGD paralogs. Several pieces of information support these conclusions. The fact that the frequency (Table 4), length (Table 5), and flanking sequence similarities (Table 6) of gene conversion of the paralogs within pre- and post-WGD species are similar indicates that mechanistic similarities are present between these genomes. In addition, the fact that the mean Ka/Ks values for the paralog families of pre- and post-WGD species are alike (Table 7) suggests that their genes are under similar selective pressures and have similar gene conversion constraints. This suggests that, despite the different ecological niches of the yeast species, these paralogs evolve in similar ways.

Surprisingly, the sequence of similarity flanking conversions between post-WGD ohnologs is always lower than that flanking post-WGD paralogs (Table 6). This is likely due to the fact that ohnologs are much older than paralogs (i.e., they have larger Ks values; Table 7), which gave time to accumulate more substitutions, and are under more selective constrains (i.e., they have larger Ka/Ks ratios; Table 7). Stronger selective constraints are expected to select against conversions which would homogenize ohnologs because such homogenization would erase the functional differences that each member of a pair of ohnologs has acquired during evolution. As mentioned above, the different function each member of a pair of ohnologs has acquired (neofunctionalization) also likely explains why different yeast genomes have so many common ohnologs (Table 1; [20]). Conversely, one of the effects of repeated gene conversion due to less negative selective pressure on paralogs is that the sequence of similarity between them will increase. Thus, the observation that ectopic gene conversions occur more frequently between

paralogs than ohnologs (Table 4) might not only be due to the fact that ohnologs are more often found on different chromosomes (Table 2) but also due to ohnologs being under stronger selective constraints than paralogs (Table 7). These stronger selective constraints are due to the fact that ohnologs are involved in essential processes (regulation, essential biosynthetic processes and metabolic processes) whereas paralogs are involved in nonessential processes (transposition, transport and nonessential biosynthetic processes; Table 8). This is similar to the situation within genes where gene conversions have been shown to be less frequent in more functionally important regions [34, 35].

The previous studies on *S. cerevisiae* have found that gene conversions are biased toward the 3′ end of converted genes. This has been attributed to ectopic gene conversion via cDNA intermediates [4]. Our results confirm that conversions are biased toward the 3′-end of genes within the *S. cerevisiae* paralog dataset [4, Table 9]. The fact that no significant bias was detected within any other species is likely a result of the low statistical power due to the small amount of data available for each of these species (Table 9). This low statistical power for the distribution of gene conversions other than those between *S. cerevisiae* paralogs likely reflects the facts that whereas there were 110 conversions between *S. cerevisiae* paralogs, there were only between 8 and 52 gene conversions between the paralogs of the other nine yeast species (Table 4). They were also only between 2 and 14 gene conversions between the ohnologs of the 7 post-WGD species. These low numbers of gene conversion are therefore not sufficient to ascertain whether their distribution is significantly biased.

The suggestion that the 3′-end bias of the gene conversions between *S. cerevisiae* paralogs is due to ectopic gene conversions with cDNA intermediates is consistent with the low number of introns present in this species as well as their 5′-position bias [4, 36, 37]. The genome of this species contains only 286 introns, and most of these introns are located at the 5′-end of the genes in which they are present [37]. This contrasts with the 139,418 introns found in the human genome and with the absence of intron position bias in human genes [37]. The model proposed by Fink to explain both the paucity and 5′-position bias of *S. cerevisiae* introns posits that incomplete cDNA molecules can recombine with their genomic copies leading to both intron loss and a 5′-position bias of the remaining introns [36, 37]. This model was later supported by the experimental demonstration that cDNA molecule can recombine with their genomic copy [9]. Since the genomes of *C. glabrata*, *D. hansenii*, *K. lactis,* and *Y. lipolytica* all have few introns and that their introns have a 5′-position bias [38], one would also expect to observe a 3′-end bias for their gene conversions if they often occur with cDNA copies. As discussed above, the fact that we did not observe such a bias in these four species could be due to the low statistical power of our tests. Alternatively, it could reflect recombination differences between *S. cerevisiae* and these four species.

In summary, our results show that the number and mean size of multigene families composed of paralogous sequences are not significantly different between pre- and post-WDG species (Table 1, Figure 2), that paralogs are more often found on the same chromosomes than ohnologs (Table 2), that gene conversions tend to be more frequent between genes found on the same chromosomes than between genes located on different chromosomes (Table 3), that gene conversions tend to be more frequent between paralogs than between ohnologs (Table 4), that the frequency (Table 4), length (Table 5), and flanking sequence similarities (Table 6) of the gene conversions between the paralogs of pre- and post-WGD species are similar, that there is a positive correlation between the length of gene conversions and flanking similarity in all converted genes (Figure 3), that ohnologs are under stronger selective constraints than paralogs (Table 7), that these stronger selective constraints are due to the fact that ohnologs are involved in essential processes whereas paralogs are involved in nonessential processes (Table 8), and that conversions are biased toward the 3′-end of the *S. cerevisiae* paralogs (Table 9). In the future, since it has recently been shown that the expression levels of duplicated genes influence their rate of sequence divergence [39], it would be interesting to test whether the increased ectopic gene conversion frequency we observed in *C. glabrata*, *D. hansenii*, and *K. lactis* (Table 3) is due to conversions between highly expressed genes.

Acknowledgments

The authors thank the two anonymous referees for their useful and constructive comments on a previous version of this paper. This work was supported by a Discovery Grant from the Natural Science and Engineering Research Council of Canada to G. Drouin.

References

[1] Y. Aylon and M. Kupiec, "DSB repair: the yeast paradigm," *DNA Repair*, vol. 3, no. 8-9, pp. 797–815, 2004.

[2] V. M. Watt, C. J. Ingles, M. S. Urdea, and W. J. Rutter, "Homology requirements for recombination in *Escherichia coli*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 82, no. 14, pp. 4768–4722, 1985.

[3] P. Shen and H. V. Huang, "Homologous recombination in *Escherichia coli*: dependence on substrate length and homology," *Genetics*, vol. 112, no. 3, pp. 441–457, 1986.

[4] G. Drouin, "Characterization of the gene conversions between the multigene family members of the yeast genome," *Journal of Molecular Evolution*, vol. 55, no. 1, pp. 14–23, 2002.

[5] W. R. Engels, C. R. Preston, and D. M. Johnson-Schlitz, "Long-range *cis* preference in DNA homology search over the length of a Drosophila chromosome," *Science*, vol. 263, no. 5153, pp. 1623–1625, 1994.

[6] D. Benovoy and G. Drouin, "Ectopic gene conversions in the human genome," *Genomics*, vol. 93, no. 1, pp. 27–32, 2009.

[7] A. S. H. Goldman and M. Lichten, "The efficiency of meiotic recombination between dispersed sequences in *Saccharomyces cerevisiae* depends upon their chromosomal location," *Genetics*, vol. 144, no. 1, pp. 43–55, 1996.

[8] G. Achaz, E. Coissac, A. Viari, and P. Netter, "Analysis of intrachromosomal duplications in yeast *Saccharomyces*

cerevisiae: a possible model for their origin," *Molecular Biology and Evolution*, vol. 17, no. 8, pp. 1268–1275, 2000.

[9] L. K. Derr and J. N. Strathern, "A role for reverse transcripts in gene conversion," *Nature*, vol. 361, no. 6408, pp. 170–173, 1993.

[10] B. Dujon, D. Sherman, G. Fischer et al., "Genome evolution in yeasts," *Nature*, vol. 430, no. 6995, pp. 35–44, 2004.

[11] K. H. Wolfe and D. C. Shields, "Molecular evidence for an ancient duplication of the entire yeast genome," *Nature*, vol. 387, no. 6634, pp. 708–713, 1997.

[12] M. Kellis, B. W. Birren, and E. S. Lander, "Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*," *Nature*, vol. 428, no. 6983, pp. 617–624, 2004.

[13] K. Wolfe, "Evolutionary genomics: yeasts accelerate beyond BLAST," *Current Biology*, vol. 14, no. 10, pp. R392–R394, 2004.

[14] D. R. Scannell, K. P. Byrne, J. L. Gordon, S. Wong, and K. H. Wolfe, "Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts," *Nature*, vol. 440, no. 7082, pp. 341–345, 2006.

[15] A. Goffeau, G. Barrell, H. Bussey et al., "Life with 6000 genes," *Science*, vol. 274, no. 5287, pp. 546–567, 1996.

[16] P. Cliften, P. Sudarsanam, A. Desikan et al., "Finding functional features in Saccharomyces genomes by phylogenetic footprinting," *Science*, vol. 301, no. 5629, pp. 71–76, 2003.

[17] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander, "Sequencing and comparison of yeast species to identify genes and regulatory elements," *Nature*, vol. 423, no. 6937, pp. 241–254, 2003.

[18] K. H. Wolfe, "Yesterday's polyploids and the mystery of diploidization," *Nature Reviews Genetics*, vol. 2, no. 5, pp. 333–341, 2001.

[19] A. van Hoof, "Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication," *Genetics*, vol. 171, no. 4, pp. 1455–1461, 2005.

[20] S. Wong and K. H. Wolfe, "Duplication of genes and genomes in yeasts," in *Comparative Genomics*, P. Sunnerhagen and J. Piskur, Eds., vol. 15, pp. 78–99, Springer, Heidelberg, Germany, 2005.

[21] K. P. Byrne and K. H. Wolfe, "The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species," *Genome Research*, vol. 15, no. 10, pp. 1456–1461, 2005.

[22] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[23] S. Sawyer, GENECONV molecular biology computer program, 1999, http://www.math.wustl.edu/~sawyer/geneconv.

[24] D. Posada and K. A. Crandall, "Evaluation of methods for detecting recombination from DNA sequences: computer simulations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13757–13762, 2001.

[25] E. Erdfelder, F. Faul, and A. Buchner, "GPOWER: a general power analysis program," *Behavior Research Methods, Instruments, & Computers*, vol. 28, no. 1, pp. 1–11, 1996.

[26] Z. Yang, "PAML: a program package for phylogenetic analysis by maximum likelihood," *CABIOS*, vol. 13, no. 5, pp. 555–556, 1997.

[27] Z. Yang and R. Nielsen, "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models," *Molecular Biology and Evolution*, vol. 17, no. 1, pp. 32–43, 2000.

[28] E. L. Hong, R. Balakrishnan, Q. Dong et al., "Gene Ontology annotations at SGD: new data sources and annotation methods," *Nucleic Acids Research*, vol. 36, no. 1, pp. D577–D581, 2008.

[29] W. P. Rice, "Analysing tables of statistical tests," *Evolution*, vol. 43, no. 1, pp. 223–225, 1989.

[30] G. Fischer, E. P. Rocha, F. Brunet, M. Vergassola, and B. Dujon, "Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages," *PLoS Genetics*, vol. 2, no. 3, article e32, 2006.

[31] D. Graur and W.-H. Li, *Fundamentals of Molecular Evolution*, Sinauer Associates, Sunderland, Mass, USA, 2nd edition, 2000.

[32] M. Lynch, *The Origins of Genome Architecture*, Sinauer Associates, Sunderland, Mass, USA, 2007.

[33] G.-F. Richard, A. Kerrest, I. Lafontaine, and B. Dujon, "Comparative genomics of hemiascomycete yeasts: genes involved in DNA replication, repair, and recombination," *Molecular Biology and Evolution*, vol. 22, no. 4, pp. 1011–1023, 2005.

[34] Z. Zhao, D. Hewett-Emmett, and W.-H. Li, "Frequent gene conversion between human red and green opsin genes," *Journal of Molecular Evolution*, vol. 46, no. 4, pp. 494–496, 1998.

[35] J. P. Noonan, J. Grimwood, J. Schmutz, M. Dickson, and R. M. Myers, "Gene conversion and the evolution of protocadherin gene cluster diversity," *Genome Research*, vol. 14, no. 3, pp. 354–366, 2004.

[36] G. R. Fink, "Pseudogenes in yeast?" *Cell*, vol. 49, no. 1, pp. 5–6, 1987.

[37] T. Mourier and D. C. Jeffares, "Eukaryotic intron loss," *Science*, vol. 300, no. 5624, p. 1393, 2003.

[38] D.-K. Niu, W.-R. Hou, and S.-W. Li, "mRNA-mediated intron losses: evidence from extraordinarily large exons," *Molecular Biology and Evolution*, vol. 22, no. 6, pp. 1475–1481, 2005.

[39] S. Pyne, S. Skiena, and B. Futcher, "Copy correction and concerted evolution in the conservation of yeast genes," *Genetics*, vol. 170, no. 4, pp. 1501–1513, 2005.