DOI: 10.1534/genetics.105.048900

After the Duplication: Gene Loss and Adaptation in Saccharomyces Genomes

Paul F. Cliften,*,[1] Robert S. Fulton,[†] Richard K. Wilson*,[†] and Mark Johnston*,[2]

**Department of Genetics and [†]Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri 63110*

Manuscript received July 29, 2005
Accepted for publication November 8, 2005

ABSTRACT

The ancient duplication of the *Saccharomyces cerevisiae* genome and subsequent massive loss of duplicated genes is apparent when it is compared to the genomes of related species that diverged before the duplication event. To learn more about the evolutionary effects of the duplication event, we compared the *S. cerevisiae* genome to other Saccharomyces genomes. We demonstrate that the whole genome duplication occurred before *S. castellii* diverged from *S. cerevisiae*. In addition to more accurately dating the duplication event, this finding allowed us to study the effects of the duplication on two separate lineages. Analyses of the duplication regions of the genomes indicate that most of the duplicated genes (~85%) were lost before the speciation. Only a small amount of paralogous gene loss (4–6%) occurred after speciation. On the other hand, *S. castellii* appears to have lost several hundred genes that were not retained as duplicated paralogs. These losses could be related to genomic rearrangements that reduced the number of chromosomes from 16 to 9. In addition to *S. castellii*, other Saccharomyces *sensu lato* species likely diverged from *S. cerevisiae* after the duplication. A thorough analysis of these species will likely reveal other important outcomes of the whole genome duplication.

GENE redundancy is common. It is produced by duplication of individual genes, by duplication of large chromosomal segments (segmental duplication), by duplication of entire chromosomes (aneuploidy), and by duplication of whole genomes. Gene duplications play a major role in evolution by providing paralogous genes that can acquire specialized functions over time (Ohno 1970). Although rare, whole genome duplications have played a major role in the evolution of species. For instance, whole genome duplications are postulated to have had a major impact on the vertebrate lineage (Ohno 1970, 1998). The whole genome duplication in the Saccharomyces lineage is thought to have shaped the fermentative lifestyle of these yeasts (Wolfe and Shields 1997; Piskur 2001).

Remnants of the whole genome duplication of *Saccharomyces cerevisiae* are apparent in its genome sequence (Wolfe and Shields 1997). There are 52 "probable" and 32 "possible" blocks of duplicated genes that include ~500 duplicated gene pairs spanning at least 70% of the genome (Seoighe and Wolfe 1999; Wong *et al.* 2002). The 16 centromeric regions map into eight duplicated pairs (Wong *et al.* 2002). Three lines of evidence suggest that these duplicated blocks of genes arose from a whole genome duplication event rather than by successive segmental duplications (Wolfe and Shields 1997). First, most of the duplicated blocks have the same orientation with respect to the telomere, a situation that would not be expected for segmental duplications. Second, if the duplications are due to successive segmental duplications, several triplicated blocks would be expected to have occurred on the basis of Poisson probability, but none are found. Finally, the order of genes in relatives of *S. cerevisiae* that did not undergo a whole genome duplication, such as *Kluyveromyces lactis*, *K. waltii*, and *Ashbya gossypii*, is what would be expected before a genome duplication event (Keogh *et al.* 1998; Dietrich *et al.* 2004; Kellis *et al.* 2004).

To learn more about the evolutionary consequences of the whole genome duplication, we investigated the fate of the duplicated genes during evolution of Saccharomyces species. We analyzed the genome sequence of species from each of the three major Saccharomyces subgroups (*sensu stricto*, *sensu lato*, and petite-negative). The genome of *S. bayanus*, a member of the *sensu stricto* group of Saccharomyces species, is highly similar to that of *S. cerevisiae* and has a high degree of synteny, indicating that it speciated after the genome duplication. *S. castellii*, a member of the *sensu lato* group of Saccharomyces species that are more distantly related to *S. cerevisiae*, also contains a duplicated genome similar to that seen in *S. cerevisiae*. The fate of many of the duplicated genes is different in these two Saccharomyces species, providing a view of genome evolution after a genome duplication.

[1]*Present address:* Department of Biology, Utah State University, 5305 Old Main Hill, Logan, UT 84322.

[2]*Corresponding author:* Department of Genetics, Campus Box 8232, Washington University Medical School, 660 S. Euclid Ave., St. Louis, MO 63110. E-mail: mj@genetics.wustl.edu

MATERIALS AND METHODS

Strains and sequences: The yeast species whose genome sequences we determined were previously described (Cliften *et al.* 2003). The genomes analyzed in this report are those of *S. bayanus* (623-6c), *S. castellii* (NRRL Y-12630), and *S. kluyveri* (NRRL Y-12651). The draft sequence assemblies are available at GenBank (project accession nos.: *S. bayanus*, AACG02000000; *S. castellii*, AACF00000000; and *S. kluyveri*, AACE02000000) and from the Saccharomyces Genome Database (http://yeastgenome.org).

The sequencing strategy was described previously (Cliften *et al.* 2003). Briefly, 3–4× shotgun sequence data were generated for each of the three species. The reads were assembled with Phrap (http://www.phrap.org). Autofinish (Gordon *et al.* 2001), a semiautomated sequence prefinishing tool, was used to identify clones that spanned sequence gaps between linked contigs and to design primers for filling the gaps.

The *S. kluyveri* assembly used in this work has undergone two additional rounds of prefinishing since the previously reported assembly. The assembly consists of 79,312 sequencing reads that assembled into 1344 contigs (*vs.* 2446 in the previous assembly). The total length of the assembly is 11 Mb, with an average contig length of 8.2 kb. The estimated sequence coverage is 3.5× with 95% of the bases having a Phred quality score of ≥40 (estimated error rate of <1/10,000) and 99% of the bases having a Phred score of ≥20 (estimated error rate of <1/100).

The *S. bayanus* assembly used in this analysis consists of 78,780 sequence reads generated by us and 146,796 reads generated at the Broad Institute (Kellis *et al.* 2003). The statistics of the Phrap assembly are: 11.7 Mb in total, 8.6-fold sequence coverage with 99.0 and 99.8% of the bases being P40 and P20, respectively, 678 contigs (335 of them >1 kb), and average contig length 17.3 kb.

Identification and analysis of duplicated blocks: Sequence contigs of the Saccharomyces species' genomes were annotated on the basis of similarity to *S. cerevisiae* proteins detected by WU-BLASTX. We omitted ORFs labeled as dubious by the Saccharomyces Genome Database (SGD) that overlap other genes since any sequence similarity could be attributed to the overlapping genes. Matches to *S. cerevisiae* proteins with a *P*-value $<10^{-5}$ and a WU-BLAST score of at least 200 were considered significant. The annotated contigs were compared to contigs of the same species to identify contig pairs containing multiple paralogs. Duplicated *S. cerevisiae* genes that are part of identified duplicated blocks of genes were treated as identical to increase the sensitivity of detecting paralogous genes in the different genome assemblies. *S. castellii* duplicated blocks are listed in supplemental Table 1 at http://www.genetics.org/supplemental/. Identification and analysis of the duplicated blocks were carried out with *ad hoc* Perl scripts.

The 130 duplicated sequence blocks of *S. castellii* (from 65 duplicated pairs) were compared to the 104 duplicated blocks (from 52 duplicated pairs) of *S. cerevisiae* to identify orthologous blocks. *S. cerevisiae* blocks that contain at least three homologs of each *S. castellii* block were identified and compared. Some large *S. castellii* blocks spanned two or more *S. cerevisiae* duplication blocks. The matching blocks were compared to determine which block was the most similar to the *S. castellii* block on the basis of the number of orthologous genes shared between the blocks. Only one *S. cerevisiae* block was assigned as orthologous to each *S. castellii* block, but since the *S. castellii* assembly is fragmented, several *S. castellii* blocks could be orthologous to nonoverlapping regions of a *S. cerevisiae* duplication block. Supplemental Table 2 (http://www.genetics.org/supplemental/) lists each of the 108 unambiguous orthologous blocks that we identified between the two species.

Identification of genes not present in *S. castellii*: Annotated *S. castellii* contigs were compared to the complete list of *S. cerevisiae* genes obtained from the SGD. We took the complete list of *S. cerevisiae* genes not present in the automated annotations of *S. castellii* contigs (1852 genes) and determined whether they were annotated as Ty coding sequences, dubious ORFs, or duplicated genes unique to *S. cerevisiae* or if weakly homologous sequences were present in the genomes, but below our threshold for annotation.

Identification of genes not present in *S. cerevisiae*: We identified all ORFs of at least 100 codons that do not overlap ORFs that are similar to *S. cerevisiae* genes. The 532 identified ORFs were compared to a nonredundant set of proteins in GenBank using BLASTP and matches of *P*-values <1*e*-20 were considered significant.

Identification of centromeric sequences: Using Perl scripts we searched for centromere sequences in each species on the basis of known conserved sequence elements of *S. cerevisiae* centromeres. We searched for conserved DNA element (CDE)I (RTCACRTG), then for CDEII (an AT-rich sequence of at least 75 bp), and finally for CDEIII (TCCGA) (Olson 1991). The stringency of the search was also reduced by searching only for two of the three elements or by shortening the length of the conserved elements. Since *C. albicans* centromeres have a different structure than the simple *S. cerevisiae* centromeres, we searched for homologous sequences in *S. castellii*, using BLASTN. No significant matches were found.

Comparison of centromere-binding proteins: Orthologs of known *S. cerevisiae* centromere-binding proteins (Cbf1, Cbf2, Cep3, Cse4, Ctf13, Mif2, and Skp1) were identified in *S. kluyveri*, *S. castellii*, *K. waltii*, and *C. glabrata*. The orthologous protein-coding sequences were compared, using WU-BLAST (BLASTP) with a postsearch Smith and Waterman alignment option. The protein sequences were also aligned with CLUSTALW to produce gene tree information.

Comparison of intron positions: Sequence data from the SGD were used to determine the intron positions within the amino acid sequence of *S. cerevisiae* spliced genes. The genome sequences of the other Saccharomyces species were compared to *S. cerevisiae* spliced genes by TBLASTN. The Blast alignment output was parsed and compared to the location of the intron sites in *S. cerevisiae* proteins to look for intron loss or gain events or for changes in the location of the introns in the orthologs of *S. cerevisiae* spliced genes. Orthologous sequences with potential intronic differences were examined manually in ACEDB to identify possible splice signals such as 5′ and 3′ splice sites and intron branch junctions [GT(ATGT), YAG, and (T)ACTAAC respectively]. *S. kluyveri* and *S. castellii* genome sequences were also compared to *S. cerevisiae* proteins by TBLASTN to look for spliced genes in these species that are not spliced in *S. cerevisiae*. The output was parsed to show gaps within the protein-coding alignments of the translated homologous sequences that could be indicative of a spliced gene. Because of the genetic distance between these species, many breaks are present within the alignments, most of which are not due to introns. We manually inspected the parsed data and looked at interesting cases in more detail within ACEDB. However, because of the large number of gaps in the protein-coding alignments, these evaluations were not exhaustive. Therefore, other spliced genes are likely present in these genomes.

RESULTS

***S. castellii*, but not *S. kluyveri*, underwent a genome duplication:** To determine the extent of gene duplication in *S. kluyveri* (a petite-negative Saccharomyces

TABLE 1

The number of duplicated gene blocks in the *S. bayanus*, *S. castellii*, and *S. kluyveri* sequence assemblies containing *X* or more duplicated genes

Minimum no. of duplicated genes in block	No. of duplication blocks in a species		
	S. bayanus	*S. castellii*	*S. kluyveri*
10	1	1	0
6	15	17	0
5	23	32	0
4	40	46	0
3	71	65	0
2	163	143	23

species), *S. castellii* (a *sensu lato* species), and *S. bayanus* (a *sensu stricto* species), we evaluated the (incomplete) genome sequences of these species. The sequence contigs from each species were compared to each other to identify contig pairs that contain multiple paralogous protein-coding genes. We did not identify any *S. kluyveri* contig pairs with more than two similar genes (Table 1), suggesting that *S. kluyveri* has not undergone extensive segmental gene duplication. In contrast, 71 *S. bayanus* contig pairs and 65 *S. castellii* contig pairs contain at least three duplicated genes. The numbers of duplication blocks in the genomes of these species are similar to the 52 probable duplicated blocks identified in *S. cerevisiae*. Most of the extra duplication blocks result from similarity to the possible duplicated blocks in *S. cerevisiae* and from gaps in the (incomplete) sequences of the genomes of these species (inferred from sequence data linking sequence contigs into supercontigs). It should be noted that the 65 duplicated blocks we identified in *S. castellii* do not represent the full complement of duplicated regions of the genome, just as the 52 probable duplicated blocks in *S. cerevisiae* do not represent the full extent of its genomic duplication. The blocks represent merely regions where duplication is most evident.

The duplicated blocks of *S. castellii* and of *S. cerevisiae* have a common origin: Since *S. bayanus* is so closely related to *S. cerevisiae* and exhibits such a high degree of synteny to the *S. cerevisiae* genome (Llorente *et al.* 2000), its duplicated blocks are undoubtedly derived from the same genome duplication event. We therefore focused on *S. castellii* and compared its 65 duplicated gene blocks to duplicated blocks in the *S. cerevisiae* genome to determine if they have a common origin. All of the 65 duplicated blocks of genes in *S. castellii* correspond to duplicated gene blocks in the *S. cerevisiae* genome. Conversely, all of the 52 *S. cerevisiae* blocks have a corresponding duplicated region of the *S. castellii* genome. Thus, there are no unique duplicated gene blocks in either genome. This supports the idea of a whole genome duplication, since new duplication blocks would be expected to have arisen after the divergence of these two species if they resulted from a series of segmental duplications. We conclude that the duplicated blocks have a common origin, despite the relatively large phylogenetic distance between *S. cerevisiae* and *S. castellii*. That is, *S. cerevisiae* and *S. castellii* speciated after the genome duplicated in the Saccharomyces lineage.

Orthology of *S. cerevisiae* and *S. castellii* duplication blocks: We compared the gene content and gene order of each duplicated block in *S. castellii* to its paralogous block and to the most similar of the 52 duplicated blocks in *S. cerevisiae*. Most of the *S. castellii* duplicated segments are more similar to their orthologous duplicated sequence block in *S. cerevisiae* than they are to their paralogous block in *S. castellii*, on the basis of the number of homologous genes in the blocks (see Figure 1). Of the 130 duplicated *S. castellii* blocks (65 pairs of duplicated blocks), 108 can be unambiguously assigned to an ortholog among the 52 *S. cerevisiae* duplication blocks. Of these 108 duplicated blocks, 84 are more similar to their orthologous block in *S. cerevisiae* than to the paralogous block in *S. castellii*, 18 blocks are as similar to their orthologous block in *S. cerevisiae* as they are to the paralogous block in *S. castellii*, and only 6 *S. castellii* blocks are more similar to their paralogous block than to the orthologous block in *S. cerevisiae*. Half of these 6 *S. castellii* blocks and all but 2 of the 18 blocks would be judged more similar to the orthologous block in *S. cerevisiae* if genes surrounding the *S. castellii* or *S. cerevisiae* block were also considered, but in each case either the *S. cerevisiae* block is much shorter than the *S. castellii* block or the two blocks only partially overlap. Thus, it is likely that all of the *S. castellii* duplicated segments are more similar to segments of the *S. cerevisiae* genome than to their paralogous block. Therefore, the majority of genome changes (gene loss) following the duplication must have occurred before these two species diverged (Figure 2).

Comparison of duplicated gene pairs: We identified 310 duplicated gene pairs in the 65 *S. castellii* duplicated blocks and an additional 239 duplicated gene pairs where at least one of the genes did not fall within the 65 duplicated blocks (compared to ~500 duplicated gene pairs in *S. cerevisiae*). Many of this latter set of 239 duplicated genes are likely to be separated from their duplication block simply because of gaps in the genome sequence assembly or because of genome rearrangements that occurred since the divergence of these species. Over half of these 549 duplicated gene pairs (319) are also duplicated in *S. cerevisiae*; 230 are uniquely duplicated in *S. castellii*. Similarly, there is no evidence of duplication in *S. castellii* for 153 of the duplicated gene pairs in *S. cerevisiae*. Thus, less than half of the duplicated gene pairs are species specific, supporting the idea that speciation occurred well after the whole genome duplication (see Figure 2).

By adding the number of duplicated genes that are present in the two species we can estimate the number

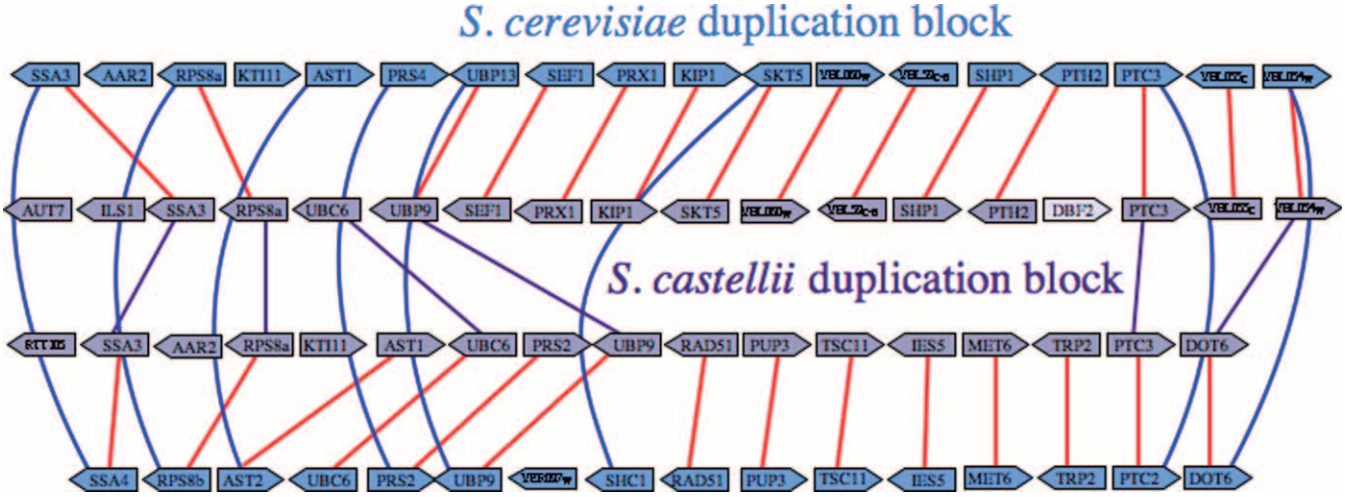

FIGURE 1.—Comparison of an othologous duplication block in *S. cerevisiae* and *S. castellii*. *S. castellii* genes (in purple) are labeled according to their top *S. cerevisiae* match by a BLASTX comparison to *S. cerevisiae* protein-coding genes. Paralogous duplicated gene pairs in the *S. castellii* block are connected by purple lines; *S. cerevisiae* genes are depicted in blue with blue lines connecting paralogous gene pairs. Red lines connect orthologous gene pairs between *S. cerevisiae* and *S. castellii*. Note that the *DBF2* homolog in the top *S. castellii* block appears to be translocated from another part of the genome.

of duplicated genes that remained before the speciation event. A minimum of ~700 (319 + 230 + 153 = 702) duplicated genes were still present prior to speciation. The maximum number of duplicated genes is more difficult to estimate since both lineages could have lost the same duplicated gene, but assuming a random loss of duplicated genes after speciation (*i.e.*, the probability of a duplicated gene being lost in both species is equal to the product of the probabilities of the duplicated gene being lost in either species), we estimate that 812 genes were duplicated before speciation. This again supports the conclusion that most of the gene loss after the whole genome duplication occurred before these two species diverged from one another.

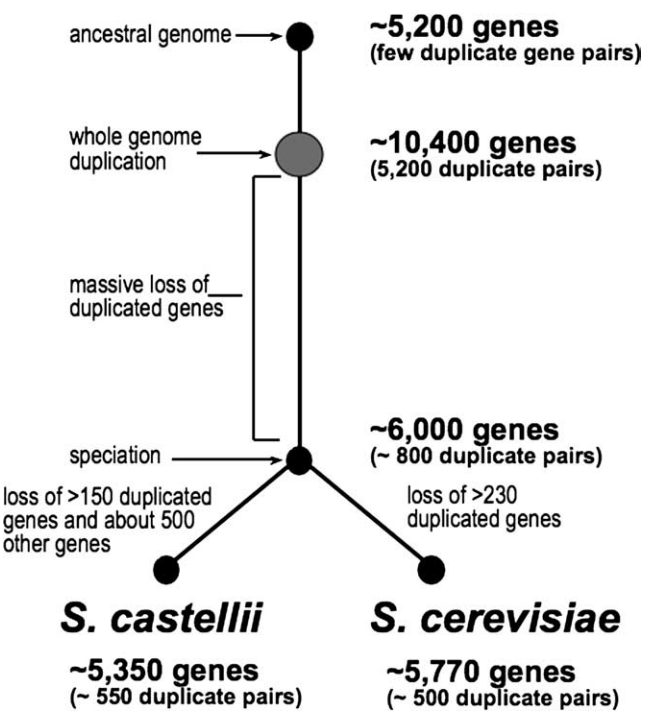

FIGURE 2.—Summary of evolutionary events that led to *S. castellii* and *S. cerevisiae*.

Table 2 shows functional classes of genes in which multiple duplicated gene pairs are more abundant in *S. castellii* than in *S. cerevisiae*. These are of interest because they may reflect pathways where new gene functions have arisen in *S. castellii*. For instance, *S. castellii* contains four genes encoding G1 cyclins compared to three in *S. cerevisiae* (for review see BREEDEN 2003). The extra gene originated from a duplication of *CLN3* in *S. castellii*. In *S. cerevisiae*, *CLN3* is located in a duplicated block, but is not duplicated. One notable set of duplicated genes in *S. castellii* is the *GAL* genes, encoding enzymes for galactose utilization (For review see HITTINGER *et al.* 2004). In *S. cerevisiae*, *GAL1* and *GAL3* are paralogs derived from the whole genome duplication. *GAL1* encodes a galactokinase; its paralog *GAL3* encodes a protein that binds ATP and galactose but whose prime role is to interact with Gal80 and relieve inhibition of the Gal4 transcription factor by Gal80 in the presence of galactose (PENG and HOPPER 2002). In *S. castellii*, *GAL1* is duplicated, but both copies are more similar to *S. cerevisiae GAL1* than to *GAL3* (see Figure 3). *S. castellii* also contains a duplication of *GAL7* (galactose-1-phosphate uridyl transferase). One of the duplicated blocks containing *GAL7* in *S. castellii* is missing *GAL10*. The *GAL1* and *GAL7* genes in this cassette are most similar to their *S. cerevisiae* homologs in the *GAL1–GAL10–GAL7* cassette and are flanked by *SNQ2* in both species, suggesting that the *GAL1–GAL7* genes in this *S. castellii* duplication block are orthologous to the genes in the *S. cerevisiae GAL1–GAL10–GAL7* gene cluster.

TABLE 2

Large classes of genes where more duplicated pairs are found in *S. castellii* than in *S. cerevisiae*

Gene class	Function	Duplicated genes
CDC	Cell division cycle	CDC14, CDC19[a], CDC25[a], CDC34, CDC48, CDC50[a], CDC55
CLN	G1 and B type cyclin	CLN2[a], CLN3, CLB2[a], CLB3[a], CLB5[a]
ECM	Extracellular matrix	ECM4, ECM18[a], ECM21[a], ECM33[a]
ERV	ER vesicle	ERV1, ERV2, ERV25, ERV29, ERV41, ERV46
GAL	Galactose utilization	GAL1[a], GAL4, GAL7, GAL11, GAL80, GAL83[a]
GRX	Glutathione reductase	GRX1[a], GRX3[a], GRX5
PCL	Pho85 cyclin	PCL1, PCL2[a], PCL5, PCL6[a], PCL8[a]
PHO	Phosphate regulation	PHO84, PHO87[a], PHO88
SEC	Secretory	SEC4, SEC9, SEC12, SEC24[a]
STB	Sin3 binding	STB3, STB5, STB6[a]
VPS	Vacuolar protein sorting	VPS5[a], VPS35, VPS62[a], VPS64[a], VPS73, VPS74

[a] Genes that are also duplicated in *S. cerevisiae*.

Duplicated copies of additional galactose utilization genes have also been retained in *S. castellii* (but not in *S. cerevisiae*), including *GAL4* and *GAL80*, which encode transcriptional regulators, and *GAL11*, which encodes a component of the mediator complex that interacts with RNA polymerase II and general transcription factors. The fact that *S. castellii* has more *GAL* genes suggests that its use of galactose may be more highly regulated than it is in *S. cerevisiae* or that it may have gained the ability to process other galactose-like molecules. These possibilities warrant further experimental analysis.

Another notable class of duplicated genes in *S. castellii* is the endoplasmic reticulum vesicle (*ERV*) gene set, which encodes proteins involved in ER-to-Golgi protein transport that are localized to COPII-coated vesicles. *S. cerevisiae* contains 6 *ERV* genes, 2 of which (*ERV14* and *ERV15*) are a duplicate pair. *S. castellii* encodes 11 *ERV* genes, 10 of which are duplicated pairs, suggesting that *S. castellii* is more versatile than *S. cerevisiae* with regard to the processes of membrane fusion, vesicle formation, and delivery of specific cargo proteins to vesicles in which the Erv proteins are involved.

There are few gene classes for which duplicated gene pairs are more prevalent in *S. cerevisiae* than in *S. castellii*. One interesting case is succinate dehydrogenase (encoded by *SDH1*, *SDH2*, *SDH3*, and *SDH4*), a multisubunit enzyme that couples succinate oxidation to the transfer of electrons to ubiquinone. In *S. cerevisiae*, *SDH1*, *SDH3*, and *SDH4* are duplicated, but *SDH2* is not. In *S. castellii*, the opposite is true: *SDH2* is duplicated, but *SDH1*, *SDH3*, and *SDH4* are not.

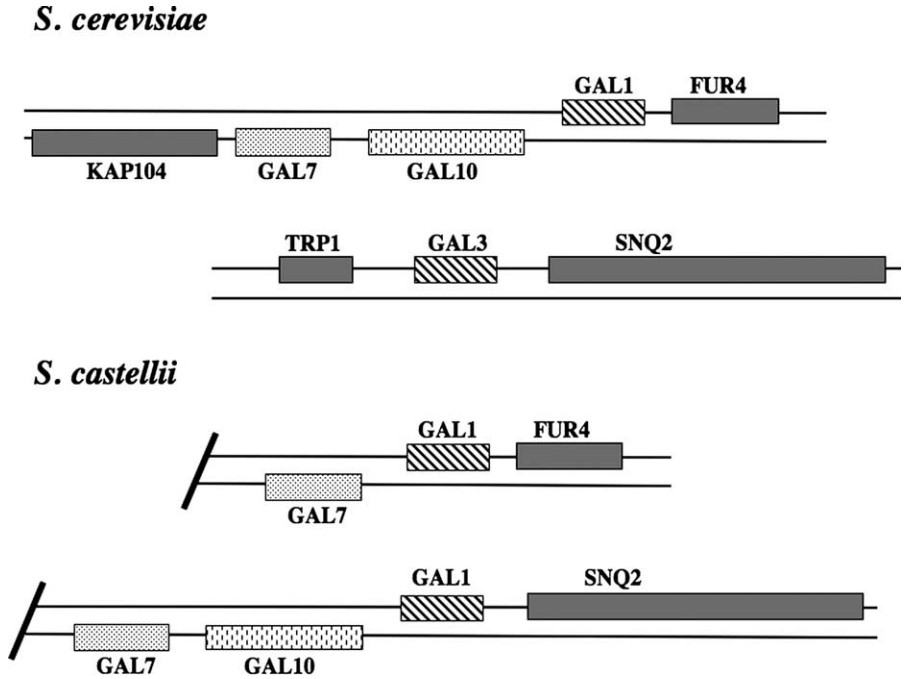

FIGURE 3.—Diagram of the *GAL1* duplication region in *S. cerevisiae* and *S. castellii*.

TABLE 3

S. cerevisiae* gene families or gene groups underrepresented in *S. castellii

	Missing gene families or functional groups	
AAD	Aryl alcohol dehydrogenase	*AAD10*, *AAD14*, *AAD15*, *AAD3*, *AAD4*, *AAD6*
MAL	Maltose utilization	*MAL11*, *MAL12*, *MAL13*, *MAL31*, *MAL32*, *MAL33*
PAU	Seripauperin	*PAU1*, *PAU2*, *PAU3*, *PAU4*, *PAU5*, *PAU6*, *PAU7*
COS	Conserved subtelomeric sequence	*COS1*, *COS10*, *COS12*, *COS2*, *COS3*, *COS4*, *COS5*, *COS6*, *COS7*, *COS8*, *COS9*
	Underrepresented gene families or functional groups	
ASP	Aspartate serine protease	*ASP3-1*, *ASP-2*, *ASP3-3*, *ASP3-4*
BIO	Biotin synthase	*BIO3*, *BIO4*, *BIO5*
BNA	Nicotinic acid biosynthesis	*BNA1*, *BNA2*, *BNA4*, *BNA5*, *BNA6*
BUD	Bud site selection	*BUD19*, *BUD25*, *BUD26*, *BUD28*
ECM	Extracellular matrix	*ECM12*, *ECM19*, *ECM23*, *ECM34*
	Iron transport	*FET5*, *FIT1*, *FIT2*, *FIT3*, *FRE1*, *FRE3*, *FRE4*, *FRE5*, *FRE6*, *FRE7*, *FTH1*
FYV	Killer toxin resistance	*FYV1*, *FYV12*, *FYV13*, *FYV15*, *FYV2*, *FYV3*, *FYV5*
KRE	Killer toxin resistance	*KRE20*, *KRE21*, *KRE22*, *KRE23*, *KRE24*, *KRE25*, *KRE26*, *KRE34*
PHO	Phosphate metabolism	*PHO11*, *PHO12*, *PHO3*, *PHO4*, *PHO5*, *PHO89*
PRM	Pheromone-regulated membrane protein	*PRM10*, *PRM3*, *PRM5*, *PRM7*, *PRM9*
SPO	Sporulation	*SPO12*, *SPO13*, *SPO19*, *SPO20*, *SPO21*, *SPO73*, *SPO74*, *SPO77*
THI	Thiamine biosynthesis	*THI11*, *THI12*, *THI13*, *THI21*, *THI22*
VPS	Vacuolar protein sorting	*VPS61*, *VPS63*, *VPS65*, *VPS68*, *VPS69*

Ribosomal protein-coding genes are the largest class of duplicated genes in both *S. cerevisiae* and *S. castellii*. *S. cerevisiae* contains 57 pairs of duplicated ribosomal genes, 52 of which lie in duplicated blocks of the genome. We identified 54 duplicated ribosomal protein gene pairs in *S. castellii*, 5 of which are not duplicated in *S. cerevisiae* (*RPL25*, *RPL29*, *RPL32*, *RPL39*, and *RPS5*). *S. cerevisiae*, on the other hand, may contain as many as 8 duplicated ribosomal gene pairs that are singletons in *S. castellii* (*RPL1*, *RPL4*, *RPL7 RPL8*, *RPL11*, *RPL15*, *RPL35*, and *RPL41*).

Genes not present in *S. castellii*: Since the *S. castellii* genome is smaller than that of *S. cerevisiae*, we compared the gene content of the two species to determine if certain gene classes or metabolic functions are absent in the *S. castellii* genome. We identified 1852 *S. cerevisiae* genes that appear to be absent in *S. castellii*, but 792 of them are classified as "dubious" ORFs by the SGD and 84 are related to Ty retrotransposon sequences. Of the remaining 976 genes that appear to be missing in *S. castellii*, 153 can be explained by genes duplicated in *S. cerevisiae* that are single copy in *S. castellii*. Approximately 250 have weak similarity to sequences in *S. castellii* and could represent rapidly diverging genes or pseudogenes. Another set of ~230 genes are members of gene families and have similarity to other proteins in *S. castellii*, so in these cases it appears that *S. castellii* has fewer members of these gene families. This leaves 340 *S. cerevisiae* genes that have no orthologs in the *S. castellii* sequence. *S. castellii* contains an additional 230 genes derived from the genome duplication that were not retained in *S. cerevisiae*. *S. castellii* may contain a few additional genes that are not in *S. cerevisiae* (see below), but overall *S. castellii* seems to have fewer genes.

Not surprisingly, the largest blocks of missing genes are located in the telomeric and subtelomeric regions of *S. cerevisiae* chromosomes. In these regions of the genome, it is not uncommon to find 5–10 consecutive genes that are missing in *S. castellii*. It is unlikely that all the missing telomeric genes are due to a cloning bias against telomeric DNA sequences or to our inability to assemble those sequences of *S. castellii* into contigs, because many *S. cerevisiae* telomeric genes are absent in the (complete) genome sequences of the related yeasts *A. gossypii* and *K. lactis*. Although telomeric repeat sequences are not common in the *S. castellii* sequence assembly, we identified 10 copies of the subtelomeric repeat *YRF* (for review see Louis and Haber 1992). Many of the *S. cerevisiae* telomeric genes may be absent in these diverged species or may have diverged beyond recognition, since the telomeric regions of the genome show a higher rate of sequence divergence (Kellis *et al.* 2003).

The *S. cerevisiae* telomeric gene families missing in the *S. castellii* genome (Table 3) include genes for biotin and thiamine synthesis and for maltose utilization, which explains known physiological deficiencies of *S. castellii* (Barnett *et al.* 2000). *S. castellii* also has a reduced number of aryl-alcohol dehydrogenase (*AAD*) genes, ferric iron uptake (*FIT*) genes, genes encoding ferric reductases (*FRE*), and *COS* and *PAU* (seripauperin) genes of unknown function.

Missing nontelomeric genes include three of four *BNA* genes, required for the synthesis of nicotinic acid from tryptophan, multiple *FYV* and *KRE* genes

implicated in resistance to killer toxin, and half of the 10 *PRM* genes that encode pheromone-regulated transmembrane proteins required for membrane fusion during mating (see Table 3). *S. castellii* is also missing several alkaline phosphatase genes (*PHO3*, *PHO5*, *PHO11*, and *PHO12*) and *PHO4*, encoding a transcriptional regulator of *PHO5*, and *PHO89*, a Na^+/Pi cotransporter gene (for a review of phosphate metabolism see Lenburg and O'Shea 1996). This is surprising, because duplicate copies of several genes encoding phosphate transporters, *PHO84*, *PHO88*, and the *PHO87/PHO90* duplicate pair (the only *PHO* gene duplication in *S. cerevisiae*), are present in the *S. castellii* genome. Thus, there appear to be distinct differences in the way these two species utilize phosphate and regulate expression of genes for this process.

***S. castellii* genes not present in *S.cerevisiae*:** Despite its smaller genome size, there are a number of *S. castellii* genes that are not found in *S. cerevisiae*. We identified 532 ORFs (of at least 100 codons) in this category, most of which are unlikely to be genes [303 (57%) are <150 codons], but there is some reason to believe that at least 41 ORFs encode functional proteins since they match known or hypothetical proteins of other organisms in GenBank (Table 4). Two gene families—one related to quinone reductase, the other related to pirin, a highly conserved nuclear protein of unknown function that is found in animals, plants, fungi, and bacteria—account for eight of the matches to GenBank proteins. Remarkably, no copies of pirin are found in the *sensu stricto* Saccharomyces species; one copy is found in *S. kluyveri*.

A total of 147 of the 532 ORFs unique to *S. castellii* are similar to other hypothetical ORFs in *S. castellii*. These include a group of subtelomeric repeats (~16) that are often found near Y′ elements. There are 18 additional *S. castellii* gene families (of three or more genes) with no similarity to proteins in GenBank. One of the largest of these gene families, consisting of at least 10 members, encodes putative proteins of ~800 amino acids. One homolog of this family is encoded in the *S. kluyveri* genome. Another six-member gene family encodes hypothetical proteins between 650 and 700 amino acids in length.

***S.castellii* chromosomes and centromeres:** Although *S. castellii* and *S. cerevisiae* appear to have diverged well after the same whole genome duplication event, *S. castellii* contains only about half as many chromosomes as *S. cerevisiae* (9 compared to 16) (Petersen *et al.* 1999). To investigate the fate of the duplicated centromeres in *S. castellii*, we searched the *S. castellii* genome for centromere sequences. First we searched for instances of CDEI (RTCACRTG), then for CDEII, an AT-rich sequence of at least 75 bp, and finally for CDEIII (TCCGA). In this way, we were able to identify all 7 centromeres in *S. kluyveri*, 7 of 8 centromeres in *K. waltii*, and 9 of 13 centromeres in *C. glabrata*, but we found no centromere sequences in *S. castellii*, even after reducing the stringency of the search. In an attempt to identify the centromeres by synteny we identified the *S. castellii* homologs of *S. cerevisiae* genes that flank centromeres. Five of the gene pairs flanking *S. cerevisiae* centromeres are also adjacent to one another in *S. castellii*, but no centromeric sequences are apparent between them. In fact, the intergenic regions between the *S. castellii* orthologs of *S. cerevisiae* genes that flank centromeres are strikingly short: an average of 313 bp, compared to an average of 1342 bp in *S. cerevisiae* and 1416 bp in *S. kluyveri*. This suggests that centromere sequence and location in *S. castellii* have diverged significantly from that of *S. cerevisiae*, which is surprising given the conservation of centromeres in the species more distantly related to *S. cerevisiae*.

In light of this, it is notable that the *S. castellii* centromere-binding proteins seem to be diverging more rapidly than orthologous sequences in the other related yeast species. For instance, of seven known centromere-binding proteins (Cbf1, Cbf2, Cep3, Cse4, Ctf13, Mif2, and Skp1) only two *S. castellii* proteins (Ctf13 and Skp1) are more similar to their *S. cerevisiae* orthologs than the *S. cerevisiae* proteins are to their orthologs from the more distantly related *S. kluyveri*. In three cases (*CBF1*, *CBF2*, and *CSE4*), the centromere-binding proteins encoded in the *K. waltii* genome are more similar to their *S. cerevisiae* orthologs than are the *S. castellii* orthologs.

***S. castellii* contains fewer introns than do *S. cerevisiae* and *S.kluyveri*:** *S. cerevisiae* contains relatively few introns. Evidence suggests that introns and spliceosomal components have been lost in the *S. cerevisiae* lineage (Rymond and Rosbash 1992). A paucity of introns has also been observed in other hemiascomycetous yeast (Bon *et al.* 2003), suggesting that the loss of introns is characteristic of this lineage. To learn more about the evolutionary fate of introns in the Saccharomyces yeasts, we looked for intron loss events with respect to the *S. cerevisiae* genome. As expected, we did not find any differences in intron number or location between *S. bayanus* and *S. cerevisiae*, nor did we find evidence of lost introns in *S. kluyveri* [one intron loss event has previously been reported for *S. kluyveri* (Bon *et al.* 2003), but there was not sufficient evidence in our assembly to confirm this]. However, we identified two extra introns in *S. kluyveri* genes that are orthologs of *S. cerevisiae* spliced genes (*YKL002w* and *YPL109c*). In contrast to what we found in *S. bayanus* and *S. kluyveri*, we identified 22 *S. castellii* genes that appear to have lost introns (see Table 5). Since these introns are in the other Saccharomyces species, the losses are specific to the *S. castellii* lineage. We also looked for introns in genes that are not spliced in *S. cerevisiae*. We found good evidence for 13 additional spliced genes in *S. kluyveri* and 2 additional spliced genes in *S. castellii*. We discovered another 8 genes that appear to be spliced in *S. kluyveri*, but where the sequence evidence alone was not conclusive. Thus, additional spliced genes are likely to be present in these genomes. In summary, *S. castellii* has fewer spliced genes

TABLE 4

S. castellii* genes not present in *S. cerevisiae

Gene name	Length	*P*-value	Annotation
Scast_Contig718.G	602 aa	4.2*e*-52	Hypothetical protein (2)
Scast_Contig667.A	674 aa	3.2*e*-92	Hypothetical WRY family protein (3)
Scast_Contig638.A	279 aa	3.7*e*-28	Hypothetical protein (2)
Scast_Contig640.C	288 aa	1.5*e*-70	Hypothetical protein[a]
Scast_Contig693.B	739 aa	4.1*e*-109	Hypothetical protein (2)
Scast_Contig652.B	345 aa	2.0*e*-107	Hypothetical protein[a]
Scast_Contig688.D	271 aa	1.2*e*-21	Hypothetical protein (2)
Scast_Contig702.J	399 aa	5.1*e*-70	Hypothetical protein (6)
Scast_Contig552.B	696 aa	4.8*e*-90	Hypothetical WRY family protein (3)
Scast_Contig627.C	363 aa	6.3*e*-104	Hypothetical protein[a]
Scast_Contig519.A	311 aa	7.3*e*-29	Hypothetical protein (1)
Scast_Contig703.A	281 aa	1.0*e*-36	Hypothetical protein (1)
Scast_Contig642.F	168 aa	8.2*e*-24	Hypothetical protein (1)
Scast_Contig704.C	441 aa	6.8*e*-123	Hypothetical protein[a]
Scast_Contig700.B	519 aa	8.9*e*-66	Hypothetical protein (5)
Scast_Contig707.B	330 aa	4.4*e*-42	Hypothetical protein (2)
Scast_Contig676.A	107 aa	7.1*e*-41	Hypothetical protein (7)
Scast_Contig656.D	382 aa	8.1*e*-88	Hypothetical protein[a]
Scast_Contig557.A	333 aa	2.0*e*-43	Hypothetical protein[a]
Scast_Contig623.A	579 aa	4.9*e*-39	Hypothetical protein (2)
Scast_Contig716.E	222 aa	2.1*e*-48	Hypothetical protein (6)
Scast_Contig642.E	288 aa	2.9*e*-51	Hypothetical protein (6)
Scast_Contig576.A	149 aa	8.5*e*-22	Hypothetical protein (1)
Scast_Contig654.C	634 aa	2.9*e*-257	Hypothetical protein[a]
Scast_Contig718.B	156 aa	3.4*e*-25	Hypothetical protein (2)
Scast_Contig668.F	172 aa	1.6*e*-36	Hypothetical protein (9)
Scast_Contig537.A	652 aa	1.1*e*-90	Hypothetical WRY family protein (3)
Scast_Contig672.D	289 aa	8.2*e*-40	Hypothetical protein (1)
Scast_Contig499.A	696 aa	4.0*e*-95	Hypothetical WRY family protein (3)
Scast_Contig473.A	650 aa	7.6*e*-91	Hypothetical WRY family protein (3)
Scast_Contig686.A	390 aa	1.2*e*-30	Hypothetical protein (1)
Scast_Contig629.B	595 aa	1.8*e*-128	Hypothetical protein[a]
		Pirin homologs[a]	
Scast_Contig674.A	361 aa	1.0*e*-105	
Scast_Contig380.A	311 aa	1.9*e*-111	
		Quinone reductases[a]	
Scast_Contig620.A	110 aa	4.6*e*-30	
Scast_Contig620.C	173 aa	6.9*e*-43	
Scast_Contig620.D	269 aa	3.2*e*-77	
Scast_Contig712.C	255 aa	5.3*e*-68	
Scast_Contig712.D	255 aa	8.9*e*-66	
Scast_Contig714.A	263 aa	4.0*e*-70	
Scast_Contig719.D	263 aa	4.4*e*-73	

The number of fungal species that contain homologous genes is shown in parentheses next to the annotation.
[a] Homologous genes that are found outside of the fungal kingdom.

than does *S. cerevisiae* (and the other *sensu stricto* species), which has fewer spliced genes than *S. kluyveri*.

DISCUSSION

The whole genome duplication event that occurred in the evolution of the Saccharomyces genus preceded the divergence of *S. cerevisiae* and *S. castellii*, but occurred after *S. kluyveri* diverged from the other Saccharomyces species (see Figure 2). Perhaps this is not surprising given the likelihood that *S. kluyveri* is a Kluyveromycete (Johnston *et al.* 1988; Llorente *et al.* 2000) and the fact that *S. kluyveri* has roughly half as many chromosomes as the *sensu stricto* Saccharomycetes (Petersen *et al.* 1999) [although *S. castellii* also has a low chromosome number (eight to nine) and a small genome size, despite having clearly undergone the genome duplication event].

TABLE 5

A summary of observed intronic differences among orthologous Saccharomyces genes

S. castellii genes missing introns that are present in orthologous *S. cerevisiae* genes
YAL001C TFC3
YAL030W SNC1
YBR078W ECM33
YCL002C YCL002C
YDR129C SAC6
YDR318W MCM21
YGL232W TAN1
YGR029W ERV1
YIL133C RPL16A
YJL001W PRE3
YJR094W-A RPL43B
YKL002W DID4
YKL081W TEF4
YKR004C ECM9
YLR306W UBC12
YMR033W ARP9
YMR125W STO1
YMR201C RAD14
YMR225C MRPL44
YNL069C RPL16B
YOL127W RPL25
YPR043W RPL43A
S. castellii genes containing an intron not present in orthologous *S. cerevisiae* genes
YDR276C PMP3
YOR351C MEK1
S. kluyveri genes containing an intron not present in orthologous *S. cerevisiae* genes
YBL058W SHP1
YDL084W SUB2
YDR377W ATP17
YGR243W YGR243W
YKL001C MET14
YKL154W SRP102
YLR066W SPC3
YML036W YML036W
YMR071C YMR071C
YMR074C YMR074C
YOL108C INO4
YOR327C SNC2
YOR351C MEK1

S. castellii's position as the earliest branching species in the Saccharomyces phylogeny suggests that the other Saccharomyces species also contain duplicated genomes. This is consistent with chromosomal numbers for the other species [16 for *sensu stricto* species and for *S. exiguus*, 12–14 for *S. servazzii* and *S. unisporus* (but only 8–9 for *S. dairenensis*, the closest known relative of *S. castellii*)] (Petersen *et al.* 1999) and with analysis of genomic survey sequences of several Saccharomycete species (Wong *et al.* 2002).

Comparison of the *S. castellii* and *S. cerevisiae* genome sequences reveals the fate of genes after a whole genome duplication (see Figure 2). The majority of gene loss in this case appears to have occurred before the speciation of the two yeasts. This view is supported by the similarity of gene order between the duplication blocks of each species and by the small number of duplicated gene pairs that are present in each species. We estimate that only ~800 duplicated genes pairs were present at the time of this event (assuming that the loss of duplicated gene pairs was random after speciation). Almost 40% of these are present in both species. Several other notable changes occurred after speciation: genome rearrangements reduced the number of chromosomes from 16 to 8 or 9 in *S. castellii*, and *S. castellii* lost a considerable number of genes (400–500) that *S. cerevisiae* retained. Perhaps these gene losses occurred during the chromosomal rearrangements that condensed the number of chromosomes in *S. castellii*.

The Saccharomyces species provide an opportunity to investigate the consequences of genome duplication on the several evolutionarily stable clades that resulted from this event. *S. cerevisiae* retained ~11% of the duplicated genes, a number that has been postulated to be a normal outcome of whole genome duplication (Lynch and Conery 2000; Wagner 2001). However, *S. exiguus* may have retained a greater number of duplicated genes, since it seems to have 16 chromosomes and a genome size estimated to be 5 Mb larger than that of *S. cerevisiae* (Petersen *et al.* 1999). On the other end of the spectrum, *C. glabrata* exhibits only a small fraction of the genetic redundancy found in *S. cerevisiae* (Dujon *et al.* 2004). Our comparison of *S. cerevisiae* and *S. castellii* indicates that there are many important biological differences between these two duplicated species. The phylogenetic distance between these two species and other duplicated *sensu lato* species suggests that other important evolutionary changes are present in the other members of this group. One important implication of this situation is that new gene functions may be unique to a species. In this respect, species like *S. kluyveri* that possess less genetic redundancy may be a good models for studying basic cellular functions that are present in a wide range of eukaryotic cells.

LITERATURE CITED

Barnett, J. A., R. W. Payne and D. Yarrow, 2000 *Yeasts: Characteristics and Identification.* Cambridge University Press, Cambridge, UK/New York.

Bon, E., S. Casaregola, G. Blandin, B. Llorente, C. Neuveglise *et al.*, 2003 Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. Nucleic Acids Res. **31:** 1121–1135.

Breeden, L. L., 2003 Periodic transcription: a cycle within a cycle. Curr. Biol. **13:** R31–R38.

Cliften, P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton *et al.*, 2003 Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science **301:** 71–76.

Dietrich, F. S., S. Voegeli, S. Brachat, A. Lerch, K. Gates *et al.*, 2004 The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science **304:** 304–307.

Dujon, B., D. Sherman, G. Fischer, P. Durrens, S. Casaregola *et al.*, 2004 Genome evolution in yeasts. Nature **430:** 35–44.

Gordon, D., C. Desmarais and P. Green, 2001 Automated finishing with autofinish. Genome Res. **11:** 614–625.

Hittinger, C. T., A. Rokas and S. B. Carroll, 2004 Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. Proc. Natl. Acad. Sci. USA **101:** 14144–14149.

Johnston, J. R., C. R. Contopoulou and R. K. Mortimer, 1988 Karyotyping of yeast strains of several genera by field inversion gel electrophoresis. Yeast **4:** 191–198.

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

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

Keogh, R. S., C. Seoighe and K. H. Wolfe, 1998 Evolution of gene order and chromosome number in Saccharomyces, Kluyveromyces and related fungi. Yeast **14:** 443–457.

Lenburg, M. E., and E. K. O'Shea, 1996 Signaling phosphate starvation. Trends Biochem. Sci. **21:** 383–387.

Llorente, B., A. Malpertuy, C. Neuveglise, J. de Montigny, M. Aigle *et al.*, 2000 Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with Saccharomyces cerevisiae. FEBS Lett. **487:** 101–112.

Louis, E. J., and J. E. Haber, 1992 The structure and evolution of subtelomeric Y′ repeats in *Saccharomyces cerevisiae*. Genetics **131:** 559–574.

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

Ohno, S., 1970 *Evolution by Gene Duplication*. Allen & Unwin, London/Springer-Verlag, New York.

Ohno, S., 1998 The notion of the Cambrian pananimalia genome and a genomic difference that separated vertebrates from invertebrates. Prog. Mol. Subcell. Biol. **21:** 97–117.

Olson, M. V., 1991 Genome structure and organization in Saccharomyces cerevisiae, pp. 1–39 in *The Molecular and Cellular Biology of the Yeast Saccharomyces*, edited by J. R. Broach, J. R. Pringle and E. W. Jones. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Peng, G., and J. E. Hopper, 2002 Gene activation by interaction of an inhibitor with a cytoplasmic signaling protein. Proc. Natl. Acad. Sci. USA **99:** 8548–8553.

Petersen, R. F., T. Nilsson-Tillgren and J. Piskur, 1999 Karyotypes of Saccharomyces sensu lato species. Int. J. Syst. Bacteriol. **49**(4): 1925–1931.

Piskur, J., 2001 Origin of the duplicated regions in the yeast genomes. Trends Genet. **17:** 302–303.

Rymond, B. C., and M. Rosbash, 1992 Yeast pre-mRNA splicing, pp. 143–192 in *The Molecular and Cellular Biology of the Yeast Saccharomyces: Gene Expression*, edited by E. W. Jones, J. R. Pringle and J. R. Broach. Cold Spring Harbor Laboratory Press, Plainview, NY.

Seoighe, C., and K. H. Wolfe, 1999 Updated map of duplicated regions in the yeast genome. Gene **238:** 253–261.

Wagner, A., 2001 Birth and death of duplicated genes in completely sequenced eukaryotes. Trends Genet **17:** 237–239.

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

Wong, S., G. Butler and K. H. Wolfe, 2002 Gene order evolution and paleopolyploidy in hemiascomycete yeasts. Proc. Natl. Acad. Sci. USA **99:** 9272–9277.

Communicating editor: B. J. Andrews