Comparative genomics of protoploid Saccharomycetaceae

Mark Johnston
Washington University School of Medicine in St. Louis

Robert S. Fulton
Washington University School of Medicine in St. Louis

Patrick Minx
Washington University School of Medicine in St. Louis

Richard Wilson
Washington University School of Medicine in St. Louis

et al

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs

Recommended Citation
Johnston, Mark; Fulton, Robert S.; Minx, Patrick; Wilson, Richard; and et al, "Comparative genomics of protoploid Saccharomycetaceae." Genome Research.19, 1696-1709. (2009).
https://digitalcommons.wustl.edu/open_access_pubs/1938

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact engeszer@wustl.edu.
Comparative genomics of protoploid *Saccharomycetaceae*


*Genome Res.* 2009 19: 1696-1709 originally published online June 12, 2009

Access the most recent version at doi:10.1101/gr.091546.109

---

**Supplemental Material**

http://genome.cshlp.org/content/suppl/2009/09/02/gr.091546.109.DC1.html

**References**

This article cites 105 articles, 34 of which can be accessed free at:

http://genome.cshlp.org/content/19/10/1696.full.html#ref-list-1

**Open Access**

Freely available online through the *Genome Research* Open Access option.

**Creative Commons License**

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

**Email Alerting Service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](http://genome.cshlp.org/subscriptions).

---

To subscribe to *Genome Research* go to:

http://genome.cshlp.org/subscriptions

Copyright © 2009 by Cold Spring Harbor Laboratory Press
Comparative genomics of protoploid \textit{Saccharomycetaceae}

The GénoLevures Consortium\textsuperscript{1}

Our knowledge of yeast genomes remains largely dominated by the extensive studies on \textit{Saccharomyces cerevisiae} and the consequences of its ancestral duplication, leaving the evolution of the entire class of hemiascomycetes only partly explored. We concentrate here on five species of \textit{Saccharomycetaceae}, a large subdivision of hemiascomycetes, that we call “protoploid” because they diverged from the \textit{S. cerevisiae} lineage prior to its genome duplication. We determined the complete genome sequences of three of these species: \textit{Kluveromyces (Lachancea) thermotolerans} and \textit{Saccharomyces (Lachancea) kluyveri} (two members of the newly described \textit{Lachancea} clade), and \textit{Zygosaccharomyces rouxii}. We included in our comparisons the previously available sequences of \textit{Kluveromyces lactis} and \textit{Ashbya (Eremothecium) gossypii}. Despite their broad evolutionary range and significant individual variations in each lineage, the five protoploid \textit{Saccharomycetaceae} share a core repertoire of approximately 3300 protein families and a high degree of conserved synteny. Synteny blocks were used to define gene orthology and to infer ancestors. Far from representing minimal genomes without redundancy, the five protoploid yeasts contain numerous copies of paralogous genes, either dispersed or in tandem arrays, that, altogether, constitute a third of each genome. Ancient, conserved paralogs as well as novel, lineage-specific paralogs were identified.

[Supplemental material is available online at http://www.genome.org and at http://www.genolevures.org/. The sequence data for \textit{Zygosaccharomyces rouxii} and \textit{Kluveromyces thermotolerans} have been submitted to EMBL-Bank (http://www.ebi.ac.uk/emb/) under accession nos. CU928173–CU928176, CU928178, CU928179, CU928181 and CU928165–CU928171, and CU928180, respectively. \textit{Saccharomyces kluyveri} sequences were deposited to GenBank under accession no. AACE03000000.]

Yeasts have played a critical role in our understanding of molecular function and evolution in eukaryotes. Their small, compact genomes, their importance in a variety of fermentation processes, and the facility of manipulating them in the laboratory have led to the determination and analysis of the genome sequences of several yeast species (for review, see Dujon 2005, 2006; Kurtzman and Piskur 2006; Scannell et al. 2007a). Molecular phylogenies suggest that unicellular yeasts arose from ancestral fungal lineages several million years ago (Miranda et al. 2006). This subdivision has been intensely studied because it contains the pathogenic yeast \textit{Candida albicans} (Jones et al. 2004; Noble and Johnson 2007) and related \textit{Candida} species (Magee et al. 2008). It also contains \textit{Debaryomyces hansenii}, whose genome shows numerous tandem gene arrays (TGAs) that contribute to gene family expansion (Dujon et al. 2004). A third subdivision consists of yeasts of the "Saccharomyces complex" or \textit{Saccharomycetaceae}, the subdivision of hemiascomycetes with the most genome sequences available. This subdivision contains a large variety of species that, although sharing a number of common physiological and genomic properties, represent a broad phylogenetic range. They were recently classified into 14 distinct clades (Kurtzman 2003; Kurtzman and Robnett 2003) whose genomes remain unequally explored.

The importance of \textit{S. cerevisiae} as a model system (for review, see Barnett 2007; Replinsky et al. 2008) has resulted in most genomic studies focusing on closely related species leaving other clades of \textit{Saccharomycetaceae} relatively unexplored, with only few exceptions (Dietrich et al. 2004; Dujon et al. 2004; Kellis et al. 2004). Today, genome sequences are available (at various coverages) for \textit{Saccharomyces paradoxus}, \textit{Saccharomyces mikatae}, \textit{Saccharomyces kudriavzevi}, \textit{Saccharomyces bayanus}, \textit{Saccharomyces exigus}, \textit{Saccharomyces servazii}, and \textit{Saccharomyces castellii} (Bon et al. 2000a, b; Casaregola et al. 2000; Clifton et al. 2003, 2006; Kellis et al. 2003), and several strains of \textit{S. cerevisiae} have been entirely sequenced (Wei et al. 2007; Borneman et al. 2008; http://www.broad.mit.edu/annotation/genome/saccharomyces_cerevisiae/).
This offers a unique view on the population structure and reproductive preferences of a group of yeasts playing an important role in fermentation processes, and a powerful tool for quantitative trait analyses (Demogines et al. 2008). Recently, wild populations of S. cerevisiae and S. paradoxus have also been characterized by hybridization to genome tiling oligonucleotide arrays and by partial genome sequencing (Liti et al. 2009; Schacherer and Kruglyak 2009).

The recognition of a whole-genome duplication in the ancestry of S. cerevisiae (Wolfe and Shields 1997) prompted many studies on the evolutionary and functional consequences of such events in yeasts (Wong et al. 2002; Byrne et al. 2006) and in other eukaryotes such as plants (Arabidopsis Genome Initiative 2000; Yu et al. 2005; Jaillon et al. 2007), fishes (Jaillon et al. 2004), or ciliates (Aury et al. 2006). The subsequent loss of redundant gene copies, which may lead to speciation bursts (Scannell et al. 2006), destroys original gene neighborhood relationships and reshapes the genetic maps of post-duplication species, leaving typical dual synteny patterns (Dietrich et al. 2004; Jaillon et al. 2004; Kellis et al. 2004). Inference of the original gene order has been attempted by analyzing duplicated genomes (Byrne and Wolfe 2005; Byrne et al. 2006; Scannell et al. 2007b; Conant and Wolfe 2008), but an authentic image of the original genome organization of this important group of yeasts is desirable.

We report the complete sequence and manual annotation of the genomes of three novel yeast species, Zygosaccharomyces rouxii, a member of the Zygosaccharomyces clade, Saccharomyces (Lanchancea) kluyveri, and Kluyveromyces (Lanchancea) thermotolerans, two members of the Lanchancea clade. We have compared these genomes with those of Kluyveromyces lactis and Ashbya (Eremothecium) gossypii (Dietrich et al. 2004; Dujon et al. 2004), providing a multispecies comparison among nonduplicated Saccharomyces. We discovered that, far from being minimal genomes, these species contain many paralogous genes, which are often highly diverged in sequence and represent approximately a third of their total genes. We also found that, despite sharing significant conservation of synteny and a common protein repertoire of approximately 3300 families, these species span a large evolutionary range as evidenced from their important sequence divergence.

These five yeast species represent four distinct clades of Saccharomycetales, respectively, designated as Zygosaccharomyces, Lachancea, Kluyveromyces, and Eremothecium (Fig. 1A). Given their complex phylogenetic relationship within the Saccharomycetales, we designate them collectively as “protoploid” solely to distinguish them from the “duplicated” yeasts (clades 1–6) (Kurtzman 2003). Other protoploid genomes, such as Kluyveromyces waltii (Kellis et al. 2004) and Kluyveromyces marxianus (Llrente et al. 2000a), were not considered since our comparisons required complete genome sequences with a single contig per chromosome.

The five yeast species studied here show diverse biological and metabolic properties. Z. rouxii is an osmo- and halotolerant species used in some fermentation processes (Solieri and Giudici 2007). It is able to grow on high concentrations of salt and/or sugar (Jansen et al. 2003) and is often considered as a food-spoiling agent. The type strain (CBS732) is haploid, but some wild isolates of Z. rouxii are diploid (Solieri et al. 2008). K. thermotolerans is usually associated with fruits or with insects feeding on plants, but one isolate was recently obtained from a wine re-fermentation process (Vilела-Moura et al. 2008). S. kluyveri is often used in protein production because it wastes less glucose by aerobic fermentation than does S. cerevisiae, and hence is better at biomass production (Møller et al. 2002, 2004). It has been developed as a model organism to study a variety of biological processes such as pyrimidine degradation (Beck et al. 2008), metabolic flux (Blank et al. 2005), or fatty acid desaturases (Oura and Kajiwara 2008). S. kluyveri appears widespread in the environment, with strains isolated from insect guts, soil, or trees in North America, Europe, and India. K. lactis and A. gossypii are, respectively, a haploid, lactose-utilizing yeast (Dujon et al. 2004), and a haploid phytopathogen able to form filaments (Dietrich et al. 2004).

Overall genome anatomy, composition, and duplications
We annotated the three novel yeast genomes sequenced here and the previously sequenced genome of K. lactis (see Methods and http://www.genolevures.org/). Annotations of A. gossypii were taken from AGD (http://agd.vital-it.ch/index.html). Table 2 lists major features and properties of these five genomes. For comparison, the genomes of S. cerevisiae (Goffeau et al. 1996) and Candida glabrata

Table 1. Sequencing and assembly of new yeast genomes

<table>
<thead>
<tr>
<th>Species</th>
<th>Strain</th>
<th>No. of chromosomes</th>
<th>Genome ploidy</th>
<th>Shotgun coverage</th>
<th>No. of final contigs</th>
<th>Final assembly size (nuc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z. rouxii</td>
<td>CBS 732*</td>
<td>7</td>
<td>n</td>
<td>11.1×</td>
<td>7</td>
<td>9,764,635</td>
</tr>
<tr>
<td>K. thermotolerans</td>
<td>CBS 6340*</td>
<td>8</td>
<td>2n</td>
<td>12.3×</td>
<td>8</td>
<td>10,392,862</td>
</tr>
<tr>
<td>S. kluyveri</td>
<td>CBS 3082</td>
<td>8</td>
<td>2n</td>
<td></td>
<td>8</td>
<td>11,345,820</td>
</tr>
</tbody>
</table>

All chromosomes were assembled as unique contigs from one end to the other, except for chromosome E of Z. rouxii and H of K. thermotolerans and S. kluyveri, each assembled as two contigs separated by rDNA repeats. Two tandem copies of rDNA repeat units, flanked by a series of Ns to constitute a total of 17 kb, were manually inserted into the final assembly of these chromosomes. Assembled contigs correspond to entire chromosome sequences with the exception of the telomeric repeats. Sequences of Z. rouxii and K. thermotolerans were deposited to EMBL (accession nos. CU928173, CU928174, CU928175, CU928176, CU928178, CU928179, CU928181; and CU928165, CU928166, CU928167, CU928168, CU928169, CU928170, CU928171, and CU928180, respectively). Sequences of S. kluyveri were deposited to GenBank under accession no. AAEO3000000. Assemblies from diploid strains did not reveal significant heterozygosity. The assembled sequences of S. kluyveri and K. thermotolerans contain the MATa sequence on chromosomes C and F, respectively. A MATalpha sequence has been recovered from the short, unassembled contigs of S. kluyveri (Payen et al. 2009). Other short, unassembled contigs of Z. rouxii correspond to mtDNA (data not shown) and to the pSRI plasmid (Araki et al. 1985). mtDNA of K. thermotolerans was previously published (Talla et al. 2005).

*Genome surveys of the same strains were previously published for Z. rouxii (de Montigny et al. 2000) and K. thermotolerans (Malpertuy et al. 2000a).
The Génolevures Consortium

Figure 1. Phylogeny and protein-coding repertoire of Saccharomycetae and outgroups. (A) The phylogenetic tree results from the alignment of 180 proteins (66,709 amino acids), selected from all universal one-member families having a homolog in S. pombe, using the MAFFT algorithm (Katoh et al. 2005), cleaned with Gblocks (Castresana 2000) before concatenation. Only families for which the ratio between the cleaned blocks and the initial alignment was higher than 75% were considered. The tree was constructed by maximum likelihood using PhyML (Guindon and Gascuel 2003) with a JTT substitution model corrected for heterogeneity among sites by a gamma-law distribution using four different categories of evolution rates. The proportion of sites that were invariable and the alpha-parameter of the gamma-law distribution were optimized according to the data. Branch length is indicated next to each branch. Clades number and designation for Saccharomycetae are according to Kurtzman (2003) and Kurtzman and Robnett (2003). Protoploid species are highlighted (bold blue names). (B) Figures represent the total number of protein families in each species or node (pan-proteome, defined as the sum of all families present in all species posterior to the node). Pie charts indicate the proportion of families classified as “ubiquitous” (core-proteome, common to all species posterior to the node), “species-specific” (present in only one of the species posterior to the node), or other combinations. In orange, the proportion of families shared by all five protoploid species. Note the similar pie charts for the node), “species-specific” (present in only one of the species posterior to the node), or other combinations.

(Dujon et al. 2004) were included as representative of post-genome duplication species of Saccharomycetae, while the genomes of Debaryomyces hansenii and Yarrowia lipolytica (Dujon et al. 2004) served as outgroups. The five protoploid genomes vary in size (from 8.7 Mb to 11.3 Mb) and are smaller than the two post-duplication genomes and the two outgroup genomes. Genome size reduction in Z. rouxii and A. gossypii is accompanied by an increased gene density (76.1% and 79.6% compared to 69.2% to 72.3% for the three other protoploid yeasts).

Two species, K. thermotolerans and A. gossypii, show a markedly higher GC content than other yeasts, Y. lipolytica excepted (Table 2). In every species, the nucleotide composition is generally uniform across the entire genome, despite local fluctuations in some cases (Supplemental Fig. 1). S. kluveri represents an astonishing exception to this rule. An ~1-Mb-long region of chromosome C displays a GC content 12% higher that the rest of the genome (25% higher for the third codon position of CDS). The remaining 250-kb portion of this chromosome is identical in nucleotide composition to the rest of the genome. This heterogeneity is studied in detail in a separate work (Payen et al. 2009). In brief, it is found that the GC-rich chromosomal segment shows conserved synteny with K. thermotolerans, is devoid of transposable elements, and replicates later than other chromosomes during the S phase.

The protoploid yeast species have six, seven, or eight chromosomes. A single centromere composed of three short elements (CDEI, CDEII, and CDEIII) was identified for each chromosome of Z. rouxii, K. thermotolerans, and S. kluveri (Supplemental Table 1; Supplemental Fig. 2B). Centromeres of K. lactis and A. gossypii were previously described (Dietrich et al. 2004; Dujon et al. 2004). Interestingly, the AT-rich CDE I spacers vary in length (Supplemental Fig. 2A). The correspondence between centromeres of the five protoploid Saccharomycetae genomes could be deduced from synteny conservation of flanking genes (Supplemental Table 2). In all likelihood, the protoploid ancestor was a species with eight chromosomes. The eight centromeres of K. thermotolerans and S. kluveri show a simple one-to-one congruence. Non-ambiguous correspondence with the former two species is also found for the seven centromeres of Z. rouxii and A. gossypii. The case of K. lactis is more complex with only six chromosomes and multiple rearrangements around several centromeres.

As anticipated from their phylogenetic position among the Saccharomycetae, the five yeast species studied here do not show traces of the whole-genome duplication that occurred in the ancestry of clades 1–6 (Kurtzman and Robnett 2003). Similarly, we can rule out the possible occurrence of another genome duplication in either of the four clades studied here since we could not find evidence of dual synteny in all pairwise map comparisons between our five species. Taking advantage of this fact, we have examined the possible presence of segmental duplications in these genomes (Supplemental Table 3). If one ignores known transposable elements and subtelomeric regions, only a few cases of segmental duplication are found in the genomes of K. thermotolerans, S. kluveri, and K. lactis, and none in Z. rouxii or A. gossypii. Such a paucity of segmental duplications is surprising in view of the high frequency ($10^{-17}$ per mitosis) of their spontaneous formation in S. cerevisiae (Payen et al. 2008), and contrasts with their abundance.

1698 Genome Research www.genome.org
in mammalian genomes (for review, see Samonte and Eichler 2002) (see Discussion).

**Genome content: Protein–coding genes and spliceosomal introns**

Protein-coding genes were identified as described in Methods. The protoploid genomes contain from 4715 (A. gossypii) to 5320 (S. kluveri) protein-coding genes (Table 2). This is 8%–18% less than *S. cerevisiae*, which, itself, contains ~10% less genes than the two species used here as outgroups, *D. hansenii* and *Y. lipolytica*. With approximately 5000 protein-coding genes on average, protoploid yeasts have the smallest known gene set among hemiascomycetes. Protein size distributions and global amino acid compositions are similar for the nine yeasts considered here (mean: 492 (six) amino acids; median: 410). Splice sites and branch-point sequence motifs are similar to those or *S. cerevisiae*, and intron lengths are shorter, except for *K. lactis*. Additional information on yeast spliceosomal introns can be found at http://genome.jouy.inra.fr/genolevures/.

**tRNA-coding genes**

The variety of tRNA species (identified by their anticodon), number of genes, and corresponding codon usage are given in Supplemental Table 4. The five protoploid yeasts follow the tRNA sparing rules previously described for other hemiascomycetes, enabling them to interpret the entire genetic code with only 43 or 44 tRNA species instead of the complete eukaryotic set of 46 (Mark et al. 2006). As expected for yeasts, a significant proportion (25%–35%) of tDNAs harbor an intron, all of which are short (except for the tDNA-Leu [CAG], where introns are 288, 150, and 134 nucleotides [nt] long for *K. lactis*, A. gossypii, and S. kluveri, respectively).

In each yeast, most species of tRNA molecules are encoded by more than one gene (up to 17 copies), and the different paralogous gene copies are identical, or almost identical, in sequence, that is, there exists only one tRNA species for each anticodon. Two exceptions are notable here: (1) in *Z. rouxii*, the two copies of tDNA-Val (TAC) differ from each other at 26 positions; (2) in *S. kluveri*, one of the three copies of tDNA-Glu (CIT) differs by 17 nt changes from the other two copies of tDNA-Glu (CIT). Interestingly, this tDNA, which is located on the GC-rich left arm of chromosome C, results from a silent T-to-C mutation in the anticodon but is otherwise identical in sequence to the 13 tDNA-Glus (TTC).

As in other yeasts, tDNAs are dispersed throughout the genomes. Distances between two successive tDNAs along chromosomes vary from a few hundred base pairs to >300 kb (median values 20–30 kb). Clusters of tDNAs are rare in yeasts and tandem arrays of identical tDNAs even rarer. Thus, it is worth mentioning the array of five tandemly repeated tDNA-Gls (CIT), each separated by ~1 kb, that lies on chromosome B of *K. thermotolerans*. As previously described for other hemiascomycetes (Marck et al. 2006; Acker et al. 2008), probable dicistronic tDNA pairs are also found in the yeast genomes we analyzed (Supplemental Table 5). In all cases, the intervals between the two successive tDNAs, that are always co-oriented, are very short and lack Pol III terminators. Some pairs, such as Arg (TCT)–Asp (GTC) or His (GTG)–Val (AAC), are common to several yeast species and probably result from ancestral fusions. Others appear specific to individual lineages.

**Genes for other noncoding RNA molecules**

A single tDNA locus is found in each protoploid species (Table 2). It is located within a chromosome arm, as in *S. cerevisiae*, not in a subtelomeric position as in some other yeasts. In each case, a gene encoding the SS RNA molecule is part of the tDNA repeat, in opposite orientation of the 35S transcript (precursor of 18S, 5.8S, and 26S rRNAs).

Genes for other major noncoding RNA molecules were annotated as described in Methods (Supplemental Table 6). The five spliceosomal RNAs (U1, U2, U4, U5, U6 snRNAs) are conserved in structure and size, and each is encoded by a single gene in the five yeast species. The same holds true for the U3 snoRNA, except for an interchromosomal duplication that created two identical genes in *Z. rouxii*. Similarly, the RNA moieties of the RNase P and SRP complexes are each encoded by a single gene in all yeast species, but their sequences and sizes differ owing to insertions of variable lengths around the conserved structural core. C/D snoRNA genes are highly....

---

**Table 2. Summary of annotated features in yeast genomes of interest**

<table>
<thead>
<tr>
<th>Species</th>
<th>Strain</th>
<th>No. of chromosomes</th>
<th>Genome size (Mb)</th>
<th>Average G+C content (%)</th>
<th>Total no. of CDS</th>
<th>Genome coverage CDS per 10 kb</th>
<th>Gene density (no. of genes per 100 kb)</th>
<th>Average C/G in CDS (%)</th>
<th>Average CDS size (codons)</th>
<th>Total tRNA genes</th>
<th>Total snoRNA genes</th>
<th>Total snoRNA genes</th>
<th>No. of rRNA clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. cerevisiae</em></td>
<td>S288c</td>
<td>16</td>
<td>12.1</td>
<td>38.3</td>
<td>5769</td>
<td>70.0</td>
<td>4.8</td>
<td>40.3</td>
<td>488</td>
<td>274</td>
<td>6</td>
<td>77</td>
<td>1 (1)</td>
</tr>
<tr>
<td><em>C. glabrata</em></td>
<td>CB5138</td>
<td>13</td>
<td>12.3</td>
<td>38.8</td>
<td>5204</td>
<td>64.2</td>
<td>4.2</td>
<td>41.7</td>
<td>507</td>
<td>207</td>
<td>6</td>
<td>ND</td>
<td>2 (5)</td>
</tr>
<tr>
<td><em>Z. rouxii</em></td>
<td>CBS732</td>
<td>7</td>
<td>9.8</td>
<td>39.1</td>
<td>4992</td>
<td>76.1</td>
<td>5.1</td>
<td>40.2</td>
<td>492</td>
<td>272</td>
<td>5</td>
<td>44</td>
<td>1 (1)</td>
</tr>
<tr>
<td><em>K. thermotolerans</em></td>
<td>CBS6340</td>
<td>8</td>
<td>10.4</td>
<td>47.3</td>
<td>5094</td>
<td>72.3</td>
<td>4.9</td>
<td>49.2</td>
<td>492</td>
<td>299</td>
<td>5</td>
<td>53</td>
<td>1 (1)</td>
</tr>
<tr>
<td><em>S. kluveri</em></td>
<td>CBS3082</td>
<td>8</td>
<td>11.3</td>
<td>41.5</td>
<td>5320</td>
<td>69.6</td>
<td>4.7</td>
<td>43.1</td>
<td>497</td>
<td>257</td>
<td>5</td>
<td>43</td>
<td>1 (1)</td>
</tr>
<tr>
<td><em>K. lactis</em></td>
<td>CBS2359</td>
<td>6</td>
<td>10.7</td>
<td>38.8</td>
<td>5076</td>
<td>69.2</td>
<td>4.8</td>
<td>40.5</td>
<td>485</td>
<td>163</td>
<td>5</td>
<td>43</td>
<td>1 (1)</td>
</tr>
<tr>
<td><em>A. gossypii</em></td>
<td>ATCC10895</td>
<td>7</td>
<td>8.7</td>
<td>52.0</td>
<td>4715</td>
<td>79.6</td>
<td>5.4</td>
<td>52.0</td>
<td>491</td>
<td>190</td>
<td>5</td>
<td>79</td>
<td>1 (1)</td>
</tr>
<tr>
<td><em>D. hansenii</em></td>
<td>CBS767</td>
<td>7</td>
<td>12.2</td>
<td>36.3</td>
<td>6395</td>
<td>74.2</td>
<td>5.2</td>
<td>38.0</td>
<td>479</td>
<td>200</td>
<td>5</td>
<td>ND</td>
<td>3 (1)</td>
</tr>
<tr>
<td><em>Y. lipolytica</em></td>
<td>CBS7504</td>
<td>6</td>
<td>20.5</td>
<td>49.0</td>
<td>6580</td>
<td>46.0</td>
<td>3.1</td>
<td>33.8</td>
<td>489</td>
<td>510</td>
<td>6</td>
<td>ND</td>
<td>6 (5)</td>
</tr>
</tbody>
</table>

Data from *S. cerevisiae* were taken from SGD (http://www.yeastgenome.org/); *C. glabrata*, *D. hansenii*, and *Y. lipolytica* from Génolevures (http://www.genolevures.org/); and *A. gossypii* from AGD (http://agd.vital-it.ch/index.html). Annotations for *Z. rouxii*, *K. thermotolerans*, *S. kluveri*, and *K. lactis* are part of this work and are available from the Génolevures online database (http://www.genolevures.org/). ND, Not determined.

See Supplemental Tables 4 and 5 for details.

rDNA clusters may be internal to chromosome arms (I) or in subtelomeric (S) locations.
conserved. (H/ACA snRNAs were not systematically investigated because of their variable size and poor sequence conservation.) The five polycistrionic clusters, which encode 17 snRNA molecules, are conserved. Also conserved are the five snoRNA genes embedded in introns of protein-coding genes. Finally, nearly all of the 21 monocistronic C/D snoRNA genes were unambiguously identified in each yeast, including snR52 transcribed by RNA polymerase III. The three absent cases (snR4 in K. lactis and S. kluveri, snR50 in K. thermotolerans, and snR62 in Z. rouxii) were missed, probably owing to extensive sequence divergence and lack of synteny. We did not precisely define telomerase RNAs (beside the two previously annotated ones) (McEachern and Blackburn 1995; Chappell and Lundblad 2004), owing to their extensive sequence variation.

**Transposable elements**

Very few transposable elements are present in the five yeast genomes we analyzed (Supplemental Table 7). Z. rouxii and A. gossypii have a few degenerate Ty3-like elements and no other class I elements (retro-elements). K. thermotolerans, S. kluveri, and K. lactis have several solo-LTRs of Ty1-like elements, but only S. kluveri has intact copies of that retro-element. In K. thermotolerans, there are only two degenerated Ty1 copies, and both are located in subtelomeric regions, as previously observed for K. lactis (Fairhead and Dujon 2006).

Our greatest surprise came from the presence of a novel class II element in the protoploid yeast genomes. Such elements have only been found so far in C. albicans (Goodwin et al. 2001, 2007) and Y. lipolytica (Neuvéglise et al. 2005). Now, we found sequences similar to the hAT DNA transposon of plants and fungi (Rubin et al. 2001) in K. thermotolerans, S. kluveri, K. lactis, and A. gossypii (Supplemental Table 7). Four full-length elements of this family (called Rover) were recognized in total. All carry a single CDS (between 631 and 867 codons), possess Terminal Inverted Repeats (TIR), and create 8-bp target site duplications. The presence of this element in four yeasts only, suggests an invasive transfer (possibly in the common ancestor of clades 10, 11, and 12 of Saccharomycetales, as the element is not found in Z. rouxii). Judging from the limited number of Rover sequences or their remnants, this element had a limited evolutionary success.

**Protein families and functional repertoire**

**Overall description of protein families and core protein repertoire**

The predicted proteomes of the five protoploid yeast species (see Methods) were classified into families based on all-to-all sequence comparisons and consensus clustering, as previously described (Nikolski and Sherman 2007). Proteomes of S. cerevisiae, C. glabrata, D. hansenii, and Y. lipolytica were included in the comparisons, bringing the total to 48,889 yeast proteins. Among these, 98% could be partitioned into 7927 protein families (Table 3). The remaining 1015 proteins (2%) could not be unambiguously classified using our clustering parameters. The list of protein families and their content can be found at http://www.genolevures.org/. A third of the protein families are common to all nine yeasts. In each species, they are represented either by a single gene (1689 families) or by several paralogous genes (902 families). Another 45% of protein families are species-specific, and the remaining 22% are represented in various subsets of species. We examined the distribution of the three classes of families in the different yeast species and at each evolutionary node according to the phylogenetic tree (Fig. 1B). As expected, most species-specific families originate from the two outgroup species. Yet about a quarter of the 5074 families found in all Saccharomycetales is made up of species-specific proteins. The five protoploid proteomes are quite homogeneous, with only small proportions of species-specific families. They share additional protein families in addition to the ones common to all nine species, bringing the total of their common families to 3295. This “core protein repertoire” represents 81%-88% of the protein families present in each of the five protoploid species.

As expected, most families of this core repertoire are represented in S. cerevisiae and C. glabrata, making it a characteristic feature of Saccharomycetales whose functions are worth elucidating.

**Functional categorization of the core protein repertoire**

Because few functional data are available for yeasts other than S. cerevisiae, we restricted our analysis of the core repertoire to families having, among them, at least one S. cerevisiae representative with a functional annotation. A total of 4097 distinct S. cerevisiae proteins with an associated GO-term for biological processes were retrieved from GO resources at http://www.yeastgenome.org/. For each of the 32 informative GO-terms recovered, the proportion of families belonging to the core repertoire was computed (Supplemental Fig. 3). Genes involved in ribosome biogenesis, protein translation or modification, transport, RNA metabolism, or cellular respiration are highly represented among the core repertoire families. In contrast, processes such as sporulation, meiosis, or conjugation are more frequently based on species- or lineage-specific genes.

**Paralogous genes, genome redundancy, and tandem arrays**

**Paralogous genes**

As noted above, a large number of protein families (1479) are represented by more than one protein in a given yeast species, revealing paralogs derived from ancestral duplications. Figure 2A

Table 3. Numerical distribution of protein families and corresponding numbers of protein-coding genes

<table>
<thead>
<tr>
<th>Families present in all nine yeast species</th>
<th>Families present in a subset of species (2-8)</th>
<th>Families present in only one species</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Families with ≤1 gene per species</td>
<td>1689</td>
<td>1416*</td>
<td>3343</td>
</tr>
<tr>
<td>Families with &gt;1 gene per species</td>
<td>15,201</td>
<td>7,640</td>
<td>3,343</td>
</tr>
<tr>
<td></td>
<td>902</td>
<td>362*</td>
<td>215</td>
</tr>
<tr>
<td></td>
<td>17,518</td>
<td>3,378</td>
<td>794</td>
</tr>
<tr>
<td>Total</td>
<td>2591</td>
<td>1778</td>
<td>3558</td>
</tr>
<tr>
<td></td>
<td>32,719</td>
<td>11,018</td>
<td>4137</td>
</tr>
</tbody>
</table>

Protein families were computed from the nine yeast species as explained in Methods, and were classified according to their presence in all species (column 1), a subset of species (column 2), or only one species (column 3). In each case, a family may be represented by one (line 1) or by several genes (line 2). For each category, the table indicates the total number of distinct protein families (bold) and the corresponding number of protein-coding genes. The complete list of families and corresponding genes can be found at http://www.genolevures.org/.

*Including 99 families (607 genes) absent from S. cerevisiae.
Evolution of protoploid yeast genomes

Figure 2. Genome redundancy. (A) Shown for each yeast species is the total number of protein families distributed according to their size (nb of paralogs per family; families of one member are not shown). The proportion (in percent) of the number of CDS belonging to multigene families to the total number of CDS is indicated in brackets next to the species abbreviation. (B) Compared genome redundancies for all nine yeast genomes. (SACE) S. cerevisiae; (CAGL) C. glabrata; (ZYRO) Z. rouxii; (KLTH) K. thermotolerans; (SAKL) S. kluyveri; (KLLA) K. lactis; (ERGO) A. gossypii; (DEHA) D. hansenii; (YALI) Y. lipolytica. Global genome redundancy is calculated as the ratio of total number of protein-coding genes in a genome vs. the total number of protein families in the same species. Raw redundancy counts all gene copies within tandem gene arrays (TGA); net redundancy considers only one gene-equivalent per TGA.

shows that the distributions of protein family sizes for the five protoploid species are globally similar. With 300–350 families of two paralogous genes per species, 95–115 families of three paralogs, and a few larger families, a total of 30%–34% of protein-coding genes are members of multigene families in the protoploid genomes (compared to 44% for S. cerevisiae). Thus, ancestral duplications of various sorts (see below) contribute to a third of the genomes of Saccharomyces cerevisiae, while the whole-genome duplication in the ancestry of S. cerevisiae contributes to only an additional 10% of paralogs. We also compared the global genome redundancy of all nine yeast species (Fig. 2B) (note that redundancy only refers here to sequence similarity, not to function). The five protoploid yeasts are similar in this regard (slightly lower for A. gossypii and different from other yeasts. The higher genome redundancy of S. cerevisiae and C. glabrata reflects their ancestral whole-genome duplication. In contrast, the higher genome redundancy of D. hansenii and Y. lipolytica results from the expansion of specific gene families (if the 104 multigene families specific to Y. lipolytica were ignored, then its net redundancy would drop to the same figure as D. hansenii). In both cases, the total gene number increases without a corresponding increase in the protein repertoire.

Ignoring TGAs (see below), paralogs appear randomly dispersed in the genomes of protoploid yeasts (Supplemental Table 8). This dispersion is compatible with interchromosomal as well as intrachromosomal duplication events that reshuffled the genomes over a long time. Global distributions of sequence identities between paralogous proteins are remarkably similar for the species studied (Fig. 3A). A major mode is observed at ~27% amino acid identity, which corresponds to highly diverged proteins (probably very ancient gene duplications). Higher identity (~35%–75%), representing more recent duplications or stronger functional constraints, is observed with slowly decreasing frequency. Note the relative excess of highly similar paralogs (>90% amino acid identity), representing even more recent gene duplications or stringent functional constraints (slightly more important in K. lactis and K. thermotolerans than for the three other species).

Tandem gene arrays (TGAs)

The total number of TGAs, with either no (most cases) or a few intervening genes between the successive paralogous copies, ranges from 31 for A. gossypii to 51 for S. kluyveri, most of which consist of only two or three genes copies (Supplemental Table 9). The few TGAs of four and five genes may correspond to functional adaptations. Some TGAs consist of a gene and a pseudogene, indicating functional inactivation of a gene copy after the duplication. Interestingly, paralogs in TGAs are generally less diverged in sequence than dispersed paralogs (Fig. 3A), in agreement with the idea that they represent more recent gene duplication events, possibly of adaptive type. Accordingly, several TGAs appear as species-specific, as if the tandem duplications occurred independently in the various phylogenetic branches (Supplemental Table 10). They encompass a large variety of functions and, for the most part, protein families of limited sizes (one to three members per species). Other TGAs are conserved within the protoploid yeasts. Many consist of large multigene families (with more than 10 members per species). Using synteny conservation as an indication of common ancestry, a minimum of 18 TGAs were formed prior to species divergence and conserved (despite copy number variation; Supplemental Table 10). For example, genes encoding...
the TAF14 subunit of the polymerase II transcriptional machinery, protein kinases involved in mitotic exit network, and B-type cyclins comprise two-copy TGA genes conserved in all five protoploid yeasts (the last one is duplicated in S. cerevisiae and C. glabrata). TGA expansion and conservation among all hemiascomycetous yeasts will be published elsewhere (L. Desbons, P. Baret, L. Frangeul, V. Leh-Louis, P. Durrens, and J.L. Souciet, in prep.).

**Conservation of dispersed paralogous gene copies**

We also examined the conservation of dispersed paralogs to pinpoint events of gene duplication or loss during the evolution of the five protoploid yeasts and to try to evaluate their functional consequences. Many families of two members (204) or three members (52) are conserved among the five species studied. Considering, as a conservative underestimate, only pairs of paralogs for which orthology of both members is clearly demonstrated for all five species by synteny conservation, a total of 114 cases could be retrieved (Supplemental Table 11). They are involved in a variety of functions, as deduced from their S. cerevisiae homologs, and have diverse degrees of sequence conservation. In addition to the genes encoding histones H4 and H2B, and mitochondrial, cytoplasmic, or peroxysomal protein isoforms that were expected, we found genes encoding a variety of protein kinases, translational elongation factors, thioredoxins, a nuclear pore component, an ATPase of the AAA family, the 1,3-beta-D-glucan synthase, factors involved in fatty acid metabolism, and specific components of the proteasome and ubiquitination pathway. Some conserved duplicated genes are of unknown function. Among the conserved paralogous proteins having undergone the greatest sequence variation (~20% of amino acid identity) are a component of the SSU processome containing the U3 snoRNA, a ubiquitin-specific protease, and a protein involved in structural maintenance of chromosomes.

**Variation of dispersed paralogous gene copies**

At the other extreme, events of gene duplication or loss that occurred during the evolution of protoploid yeasts are illustrated by several species-specific families of paralogs. Given the diversity of the resulting situations, we have examined three simple cases. If one considers (Supplemental Table 12) genes present in one copy in four species but in families of two or more paralogs in the fifth, the largest amplification is represented by five paralogs (including one tandem) in S. kluyveri that encode proteins with weak similarity to the Amn1 proteins of S. cerevisiae involved in daughter cell separation and chromosome stability. Other series of two or three paralogs are involved in a variety of functions. If one considers (Supplemental Table 13) genes present in families of two or more paralogs in one species but absent from all others, the most spectacular cases are two families of, respectively, 20 and 10 members observed in Z. rouxii, the first one related to the COS/DUP family of mostly subtelomeric genes in S. cerevisiae. Similarly, there are two families of seven members in K. lactis, two families of five members, and one family of four members in K. thermotolerans. Among this species-specific category, families of two or three paralogs are also found in all species, but, unfortunately, almost no functional inference is possible owing to lack of homologs. Finally, there are a variety of gene family expansions in some species that are probably related to functional adaptation (Supplemental Table 14). For example, K. lactis has 12 copies (compared to zero to four for other species) of a gene similar to the Mch2 protein of S. cerevisiae that may be involved in importing monocarboxylic acids. Similarly, S. kluyveri has 12 copies (compared to three to six for other species) of a gene encoding proteins weakly similar to cell surface proteins, and Z. rouxii has 10 copies (compared to zero to four for other species) of genes potentially encoding oxydoreductases involved in the formation of chiral alcohols.

**Orthologous genes and synteny conservation**

**Definition of orthologs**

The proper identification of orthologous genes is a major challenge (Kuzniar et al. 2008) because accumulation of gene-loss and duplication events tends to blur the recognition of true orthologs among the set of remaining homologs. Reciprocal best hits are often used to assign orthologs among a set of homologs (Rivera et al. 1998), but true orthologs are not necessarily those with the most similar sequences among all homologs (Lynch and Conery 2000). The present set of yeast species offered us a chance to define orthologous genes based on conservation of their genetic location. We started from a series of homologous genes identified from the corresponding protein families (see above), and examined their flanking chromosomal regions for the presence of other homologs (see Methods). In this way, we defined subsets of orthologous genes from each inspected family of homologs. The procedure was reiterated exhaustively until all protein-coding genes of all yeast species had been examined, defining SONS (Subset of Orthologs by Neighborhood and Similarity). In total, we identified 3896 SONS out of 4393 families (80% of families present in more than one species). A total of 27,926 genes (64% of the genes) are assigned to SONS, that is, belong to groups of orthologs confirmed by a shared neighborhood. Most of the remaining genes belong to Y. lipolytica and D. hansenii as expected from the very low conservation of synteny between those species and other yeasts. The complete list of genes in SONS can be found at http://www.genolevures.org/.

**Conservation of orthologs**

A total of 1609 SONS (41%) are common to all yeasts (i.e., they comprise a series of 1:1:1:1:1 orthologs confirmed by synteny). For the rest, some genes are missing in some species owing to gene loss during evolution. Using all alignments of orthologous proteins (as defined from cognate SONS), we computed the distributions of sequence identities for all pairwise comparisons of our five yeasts (Fig. 3B). Strikingly, their mean or median values (all distributions are monomodal, as expected for orthologs) underline the large evolutionary distances between protoploid species of Saccharomycetaceae. Even species belonging to the same clade, such as S. kluyveri and K. thermotolerans, show an average amino acid identity as low as 58.2%. Other pairwise comparisons show even lower similarity (48%–53%), consistent with their proposed phylogeny (see Fig. 1). Thus, even a homogeneous phylogenetic group such as the Saccharomycetaceae spans a very broad evolutionary range.

**Synteny conservation**

Using gene orthologs (as defined from SONS), we computed the maximal synteny conservation between the five protoploid yeast genomes as described in Methods. Pairwise comparisons between the five species (Supplemental Fig. 4) show the significant amount of map reshuffling that occurred during the evolution of this group of species, leaving numerous and relatively short synteny blocks scattered throughout these genomes. However, recognizable synteny blocks cover significant proportions of each genome (~80%–90% of all protein-coding genes). The distributions of block size (14–26 genes) and number (180–291) vary only slightly between pairs of species, except for the S. kluyveri vs. K. thermotolerans...
comparison that shows larger (mean size 59 genes) and fewer (84) synteny blocks. An example of synteny block conservation is illustrated by Supplemental Figure 5. Note that, despite extensive map reshuffling, the five protoploid species clearly form a homogeneous group: When a more distantly related yeast such as D. hansenii is compared to any of them, the size of synteny blocks and their coverage of the genome diminish considerably (Supplemental Fig. 6). Distributions of size of synteny blocks in all pairwise species comparisons (Fig. 4) are consistent with the phylogenetic relationships between species. Interestingly, the frequency distribution of synteny blocks according to their size fits a power law rather than the expected exponential law implied by the random breakage model of Nadeau and Taylor (1984).

Our definition of synteny blocks allows for the presence of “intervening” genes in them (see Methods). The number of such genes is roughly proportional to block sizes (Supplemental Fig. 6). Figures are similar for Z. rouxii, K. lactis, and A. gossypii (approximately seven intervening genes for 100 orthologs in syntenic blocks), but are significantly higher for S. kluyveri (approximately 12). An intermediate figure is observed for K. thermotolerans (approximately nine).

Relationship between sequence divergence and synteny conservation

We examined the relationship between the rates of sequence divergence and chromosomal rearrangements during the evolution of hemiascomycetes (Fig. 5). There is a nonlinear relationship between average amino acid identity of orthologous proteins and the proportion of genes remaining in synteny. The five protoploid species are grouped in a sector of the curve in which sequence identity decreases rapidly while synteny conservation decreases slowly. Pairwise comparisons of the same species with those of the outgroup (D. hansenii or Y. lipolytica) extrapolate well with results of the most distant species among protoploids (Z. rouxii vs. K. lactis or A. gossypii). The nonlinear form of this relationship indicates that significant sequence divergence occurs before chromosomal maps become extensively rearranged. At larger evolutionary distances, chromosome reshuffling catches up to protein-sequence divergence, which becomes limited by saturation and functional constraints (see Discussion).

Genome rearrangements from inferred ancestors

The slow pace of genome rearrangements enabled us to infer ancestral events from contemporary genomes based on conservation of synteny blocks. Such an inference is possible because, in protoploid yeasts, the probability of independent formation of such blocks in different lineages is likely to remain negligible compared to their inheritance from a common ancestor. Each genome was factored into a sequence of ordered synteny blocks common to all genomes, and super-blocks were computed using the method developed by G Jean, DJ Sherman, and M Nikolski (in prep.). From this, we computed median genomes for the five protoploid genomes, as well as rearrangement trees containing intermediate
ancestral candidates for different branches (Goëffon et al. 2008). Figure 6 shows pairwise rearrangement distances and a rearrangement tree. The rearrangement tree separates Z. rouxii from the other four species, in agreement with the phylogenetic tree based on sequence alignments (see Fig. 1). Again, in agreement with phylogeny, S. kluveri and K. thermotolerans are the closest in terms of rearrangements. However, in the rearrangement tree, A. gossypii and K. lactis are as diverged from each other as Z. rouxii is from the ancestor of all five species. This underlines the highly dynamic nature of yeast genomes, even within a related group like the Saccharomycetales. (The difference in rearrangement distance between the binary rearrangement tree and a simple star topology rooted at the median is only 10 events [3.5%].)

**Discussion**

With the complete genome sequencing and analysis of three novel yeast species, Z. rouxii, K. thermotolerans, and S. kluveri, we have been able to perform the first comparative genomic study involving four distinct clades (and five species) of Saccharomycetales, that all separated from S. cerevisiae before its ancestral genome duplication. In two previous studies, yeasts of these clades, namely, A. gossypii (Dietrich et al. 2004) and K. waltii (Kells et al. 2004), were individually compared to S. cerevisiae but not between themselves. In another study (Dujon et al. 2004), K. lactis was involved in multispecies comparative genomics with four other hemiascomycetes, but none was a protoploid Saccharomycetales. The present study was facilitated by the development of new methods that may be of general interest for other comparative genomic studies between species having sufficient conservation of synteny (see Methods). Among them is the method we used to define gene orthology (MI. Seret and PV Baret, in prep.). Compared to YGOB, which implies manual curation of data (Byrne and Wolfe 2005), and to SYNERGY, which requires a priori definition of parameters (Wapinski et al. 2007), our method (IONS) simply subdivides precomputed sets of homologs (based on sequence similarity) using gene neighborhood evidence in a reiterative process with gradually decreasing neighborhood sizes until series of homologs with only one gene per species are obtained.

Genomes of the five species we examined reveal common signatures that confirm the monophyletic origin of Saccharomycetales and distinguish them from representatives of other sub-divisions of hemiascomycetes, used here as outgroups for comparison. Among those signatures are short centromeres (Meraldi et al. 2006), triplication of mating cassettes (except for S. kluveri) (Butler et al. 2004), a very limited number of spliceosomal introns (Bon et al. 2003), usage of the universal genetic code (Miranda et al. 2006), and a single rDNA locus containing the genes for the SS RNA molecule (Acker et al. 2008). Additional discriminatory signatures are also found in their mitochondrial genomes: The genes encoding subunits of complex I of the respiratory chain are missing, while a gene for a variable ribosomal subunit and another one for an abnormal tRNA-Thr (UAG) altering the mitochondrial genetic code are present (http://www.ncbi.nlm.nih.gov/genomes/OrganelleResource.cgi?opt=organelle&taxid=4751).

The five protoploid genomes show only limited variation in the total number of protein-coding genes (~7% relative to the mean) and almost no variation for some noncoding RNA genes. But their chromosome number varies from six to eight, the number of tRNA genes varies from 163 to 272, and the average genome composition from ~39% to 52% GC, indicating important lineage-specific evolution that may be correlated with the different lifestyles and properties of these five species. Such diversity is not surprising given the large evolutionary distances separating the five species, as revealed by the significant sequence divergence among orthologous proteins. This was not known before. Even the two species classified within the same Lachancea clade, S. kluveri and K. thermotolerans, show average amino acid identity of only ~58%. This is much less than between the two most distant species of the Saccharomyces sensu stricto clade (Bon et al. 2000a; Cliften et al. 2003; Kells et al. 2003), which has nevertheless been more extensively studied, illustrating again previous emphasis on yeasts involved in biotechnological fermentation.

Evolutionary distances among protoploid yeasts are such that our sequence-based phylogeny (Fig. 1), which is in agreement with that of Fitzpatrick et al. (2006), is not entirely congruent with the phylogeny of Saccharomycetales published by Kurtzman (2003), where clade 12 (Eremothecium) is separated from clade 11 (Klyveromyces). In both cases, however, Z. rouxii branches separately from the four other yeasts and shares a common origin with the duplicated yeasts of the Saccharomyces and Nakasonezymes clades. Results presented here definitely demonstrate that Z. rouxii separated from them before the whole-genome duplication. Given the tree topology of Kurtzman (2003) and the fact that the Vanderwaltzoyna clade was recently demonstrated to have emerged after the duplication (Scannell et al. 2007b), Zygosaccharomyces represents the last clade having diverged from its relatives prior to the whole-genome duplication. Of all protoploid yeasts studied here, Z. rouxii is, therefore, the most closely related to the putative ancestral genome of S. cerevisiae.

Despite their long evolutionary branches, the five protoploid species we analyzed show a high degree of synteny conservation.
Even if chromosome rearrangement breakages are relatively numerous (a few hundred, except between S. kluyveri and K. thermotolerans), they leave relatively long synteny blocks of one or two dozen genes on average, and large proportions of genomes (>80%) are present in still recognizable synteny blocks in all pairwise comparisons. Such a range of synteny conservation was not available for yeasts before. The *Saccharomyces sensu stricto* yeasts share almost completely conserved synteny with few breakpoints (Fischer et al. 2000, 2001), and other yeasts are too distantly separated to retain much synteny (Fischer et al. 2006). The nonlinear relationship between sequence divergence and synteny conservation is compatible with the idea that genetic maps are more robust than DNA sequences over evolutionary periods corresponding to entire families or even orders of the Linnean taxonomical hierarchy. The effects of genome rearrangements become significant over longer evolutionary times when sequence changes saturate. Similar ideas based on much more limited sets of data were already discussed (Langkjaer et al. 2000; Llorente et al. 2000b; Malpertuy et al. 2000b). The frequency distribution of synteny blocks according to their size is similar to results obtained from the 12 *Drosophila* genomes and other insects (Zdobnov and Bork 2007), suggesting that chromosome rearrangements are not random in either of these two groups of eukaryotes despite very distinct genome organizations. Accordingly, genome rearrangement trees built from inferred ancestors are not entirely coincident with phylogenetic trees built on sequence conservation of gene products. Differences in the rate of genome rearrangements between yeast lineages have been reported (Fischer et al. 2006). Note that conserved synteny blocks often contain short internal gene deletions and/or single gene insertions (“intervening” genes). Many “intervening” genes have homologs at ectopic locations in either the same or other genomes and may have arisen by single gene transfer (possibly by retroposition) or by dispersed short segmental duplications similar to the few that were recognized. A few “intervening” genes correspond to horizontally transferred genes, as previously illustrated by the presence of a bacterial transposase-type of sequence in *A. gossypii* (Hall et al. 2005). Gene transfer from bacteria to yeasts is also documented in *S. kluveri* by the presence of six copies of a bacterial IS element (family IS607) on chromosomes B, G, and H (Rolland et al. 2009).

The genomes of protoploid yeasts contain remarkably few segmental duplications (except in subtelomeric locations), in contrast to genomes of higher eukaryotes, where they are a major source of genome rearrangements (Johnson et al. 2001; Samorte and Eichler 2002), and to the high frequency of their spontaneous formation in *S. cerevisiae* (Payen et al. 2008). The limited role of segmental duplications in the evolution of yeast genomes may be related to the important instability of duplicated copies of identical sequences. Compared to their paucity in recent duplications of nearly identical sequences, genomes of protoploid yeasts contain a surprisingly high number of paralogous genes: about a third of their gene content. Thus, we can say that there is no such thing as “nonduplicated” yeasts. Families of paralogs issue from ancient duplication events that leave diverged gene copies at dispersed genomic locations and from TGAs, in which gene copies are globally less diverged. Although some of the dispersed duplications appear ancestral and conserved in all protoploid species (some are also conserved in outgroup yeasts, suggesting they may be ancestral to all hemiascomycetes), we recognized a significant number of lineage-specific events of gene duplication and loss, some of which resulted in significant gene family expansions that may be correlated with specific biological properties of each yeast. But the remarkable point is that, despite such dynamics, the diverse protoploid species have similar levels of global genome redundancy (the figure for *A. gossypii* is slightly smaller, in agreement with a reductive and adaptive evolution [Dietrich et al. 2004]). The net redundancy, ~1.25 on average, reflects an equilibrium between gene duplication and loss over long evolutionary periods, when this equilibrium is not affected by accidental events such as whole-genome duplication or adaptive events such as gene family expansions.

The genome sequences of the *Saccharomycetaceae* we described, which offer a more balanced representation of the different clades, provide the framework to understand the origin of the common properties and individual variations of these yeasts. Their core genetic repertoire (core proteome) consists of approximately 3300 protein families, within a pan-proteome of approximately 5000 families for all *Saccharomycetaceae*. The fact that the pan repertoire dramatically increases when outgroup species are considered underlines the major evolutionary gaps between the different subdivisions of hemiascomycetes and the need for additional genomic studies of this interesting fungal group.

**Methods**

**Genome sequencing**

Complete genome sequences of *K. thermotolerans* (CBS6340) and *Z. rouxii* (CBS732; determined at Genoscope) and of *S. kluveri* (CBS3082; determined at the Washington University Genome Sequencing Center) were automatically assembled from shotgun Sanger sequencing reads using ARACHNE (Jaffe et al. 2003) and finished by dedicated sequencing for gaps and low-quality regions.

**Genome annotation**

**CDS**

Gene models for protein-coding genes were constructed using GeneMark (Borodovsky and Mclnirch 1993) trained with a conservative set of coding and noncoding sequences for each of the four genomes: *Z. rouxii*, *K. thermotolerans*, *S. kluveri*, and *K. lactis*. All open reading frames (ORFs) greater than 80 codons were examined. Overlapping conflicts were solved by best GeneMark prediction and/or existence of BLASTP alignments with the UniProtKB database. Gene models were manually examined and simultaneously annotated for the four species after automated identification of homologs (D. Sherman, T. Martin, and P. Durrens, in prep.).

**Spliceosomal introns**

Consensus splice sites and branch point sequences were defined for each yeast species from introns of ribosomal protein-coding genes (Bon et al. 2003). All combinations of the three motifs separated by appropriate distances (also defined for each yeast species) were examined for the possible formation of CDS greater than 80 codons after splicing. Note that introns in 5’-UTRs were, therefore, not systematically predicted. Overlapping intron predictions were manually curated.

**Genes for noncoding RNAs**

Genes for tRNAs were identified as described in Marck et al. (2006). Ribosomal DNA repeats were identified by comparison to
other yeast genomes (Chindamporn et al. 1993; Rustchenko and Sherman 1994) and mapped to chromosomes using previous karyotype data (Sor and Fukuhara 1989; Neuvéglise et al. 2000; Sychrova et al. 2000). Genes for C/D snoRNAs and snRNAs were annotated by a combination of sequence similarity search with S. cerevisiae, secondary structure, and synteny conservation. H/ACA snoRNAs were not systematically identified.

**Mobile genetic elements**

Transposable elements were detected by TBLASTN on chromosome sequences using known fungal transposable elements as queries, and manually curated. Identified elements were then used to detect partial or degenerate elements and solo LTRs.

**Protein families**

Four complementary distance matrices were computed between the predicted translation products for the 47,874 protein-coding genes of the seven yeasts annotated by Génolevures (C. glabrata, Z. rouxii, K. thermotolerans, S. kluveri, K. lactis, D. Hansenii, and Y. lipolytica) (http://www.genolevures.org/), plus S. cerevisiae (http://www.yeastgenome.org/) and A. gossypii (http://agd.vital-it.ch/), combining BLAST and Smith-Waterman alignments with and without filtering for homeomorphy (Wu et al. 2004). Symmetric matrices derived from amino acid identities were constructed and submitted to MCL clustering (Enright et al. 2002) with a range of inflation parameters. These competing partitions were reconciled using the consensus method of Nikolski and Sherman (2007) and manually curated using literature search and systematic comparisons with COG (Tatusov et al. 2003) and PIRSF (Wu et al. 2004) classifications.

**Orthologous genes**

SONS were defined by the combination of two dimensions: sequence similarity of gene products (defined from families; see above), and conservation of gene neighborhood along chromosomes (Supplemental Fig. 7). Two genes of different yeast species whose translation products belong to the same family (homologs by similarity) will be members of the same SONS if they share at least one pair of neighbors that are also homologous to each other by similarity. The process is reiterated for all possible hetero-specific pairwise comparisons of homologs. At an initial step, neighboring is examined for 15 genes on each side of the two homologs considered, giving rise to $\binom{2 + 14}{2}$ pairwise comparisons. Homologs that do not share a pair of homologous neighbors are separated in two distinct SONS. The process is reiterated (reducing the neighborhood size in order to decrease the probability of spurious connections) until each SONS contains $\geq 1$ gene for each yeast species. Homologs in such a case are considered as orthologs confirmed by synteny. Details of the method and comparisons with other methods to infer orthology will be published separately (ML Seret and PV Baret, in prep.).

**Syntenic block construction**

Syntenic blocks between two yeast genomes were constructed from the physical adjacency of orthologous genes (defined by SONS, above) along chromosomes (deduced from sequence-based map), using two parameters: $A$, the minimum number of ortholog pairs forming anchor points; and $I$, the maximum number of non-orthologous genes (“intervening”) between two successive anchor points. ($A$ and $I$ were set, respectively, to 5 and 10, following Fischer et al. [2006]). Note that TGAs (see text) were considered as equivalent to a single gene.

**List of participants and affiliations**

**Overall coordination**

Jean-Luc Souciet, Bernard Dujon, Claude Gaillardin, Mark Johnston, Philippe V. Baret, Paul Cliften, David J. Sherman, Jean Weissenbach, and Eric Westhoff

**Genome sequence and assembly**

Patrick Wincker, Claire Jubin, Julie Poulain, Valérie Barbe, Béatrice Séguers, François Artiguenave, Véronique Anthouard, Benoit Vacherie, Marie-Eve Val, Robert S. Fulton, Patrick Minx, and Richard Wilson

**Analysis and annotation**


**Acknowledgments**

We thank E. Beyne and H. Soueidan for help in programming and data management, Y. Boulard for help in tRNA gene search, M. Santos for sharing unpublished results on tRNA genes, A.N.

---

2Université de Strasbourg, CNRS UMR7156, F-67000 Strasbourg, France.
3Institut Pasteur, CNRS URA2171, University Pierre et Marie Curie, Paris 6 (UFR929, F-75724, Paris-CEDEX15, France.
4AgroParisTech, CNRS UMR2885, INRA UMR1238, Microbiologie et Génétique Moléculaire, F-78850 Thiverval-Grignon, France.
5Washington University School of Medicine, Department of Genetics and Genome Sequencing Center, St. Louis, Missouri 63108, USA.
6University of California—Denver, Anschutz Medical Campus, Mail Stop 8101, P.O. Box 6511, Aurora, Colorado 80045, USA.
7Washington University School of Medicine, Department of Genetics and Genome Sequencing Center, St. Louis, Missouri 63108, USA.
8Washington University School of Medicine, Department of Genetics and Genome Sequencing Center, St. Louis, Missouri 63108, USA.
9Université Louis Pasteur, Architecture et Réactivité de l’ARN, Institut de Biologie moléculaire et cellulaire du CNRS, F-67084 Strasbourg, France.
10Washington University School of Medicine, Department of Genetics and Genome Sequencing Center, St. Louis, Missouri 63108, USA.
11Institut de Biologie et de Technologies de Saclay (iBiTec-S), CEA, F-91191 Gif-sur-Yvette CEDEX, France.
12Université de Lyon 1, CNRS, UMR5240 Microbiologie, Adaptation et Pathogénie, INSA de Lyon, Villeurbanne, F-69621 Villeurbanne, France.
13Université de Lyon 1, CNRS, UMR5800, LaBRI INRA Bordeaux Sud-Ouest (MAGNÔME) F-33405 Talence, France.
14CEA, DSV, IG, Génoscope; CNRS UMR 8030; Université d’Evry Val d’Essonne, F-91057 Évry, France.
15Université Louis Pasteur, Architecture et Réactivité de l’ARN, Institut de Biologie moléculaire et cellulaire du CNRS, F-67084 Strasbourg, France.
16Université de Lyon 1, CNRS, UMR5240 Microbiologie, Adaptation et Pathogénie, INSA de Lyon, Villeurbanne, F-69621 Villeurbanne, France.
17Institut Pasteur, Platform Intégration et Analyse génomique, F-75015, Paris, France.
18Institut de Physiologie AS CR, Department of Membrane Transport, Vídeonska 1083, 14220 Prague 4, Czech Republic.
19Université de la Méditerranée, Laboratoire de Chimie Bactérienne, CNRS-UPR9043, F-13402 Marseille CEDEX 20, France.
20Present address: Department of Biochemistry and Molecular Genetics, University of Colorado—Denver, Anschutz Medical Campus, Mail Stop 8101, P.O. Box 6511, Aurora, Colorado 80045, USA.
Evolution of protoploid yeast genomes


References


All information including authors and titles are taken from the given text and organized in a logical manner. The references are formatted according to the given citation style, ensuring proper attribution and adherence to academic standards. The text is free of any bias, and the extracted data accurately reflect the content of the original document.
The Génolevures Consortium


1708 Genome Research www.genome.org
Evolution of protoploid yeast genomes


Received January 23, 2009; accepted in revised form April 20, 2009.