Complete DNA Sequence of *Kuraishia capsulata* Illustrates Novel Genomic Features among Budding Yeasts (*Saccharomycotina*)

Lucia Morales[1], Benjamin Noel[2], Betina Porcel[2], Marina Marcet-Houben[3,4], Marie-Francoise Hullo[1], Christine Sacerdot[1,8], Fredj Tekaia[1], Véronique Leh-Louis[5], Laurence Despons[5], Varun Khanna[1], Jean-Marc Aury[2], Valérie Barbe[2], Arnaud Couloux[2], Karen Labadie[2], Eric Pelletier[2], Jean-Luc Souciet[6], Teun Boekhout[7], Toni Gabaldon[3,4], Patrick Wincker[2], and Bernard Dujon[1,*]

[1]Institut Pasteur, Unité de Génétique Moléculaire des Levures, CNRS UMR3525, Univ. P. M. Curie UFR927, Paris, France

[2]Commissariat à l'Energie Atomique, Institut de Génomique/Génoscope, Evry, France

[3]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Barcelona, Spain

[4]Universitat Pompeu Fabra (UPF), Barcelona, Spain

[5]Univ. de Strasbourg, CNRS UPR9002, Architecture et Réactivité de l'ARN, Institut de Biologie Moléculaire et Cellulaire du CNRS, Strasbourg, France

[6]Univ. de Strasbourg, CNRS UMR7156, Institut de Botanique, Strasbourg, France

[7]CBS Fungal Biodiversity Centre, KNAW, Utrecht, Netherlands

[8]Present address: Institut de Biologie (IBENS), Ecole Normale Supérieure, 46 rue d'Ulm, Paris, France

*Corresponding author: E-mail: bdujon@pasteur.fr.

Accepted: November 27, 2013

Data deposition: Assembled sequences as well as reads have been deposited at European Nucleotide Archive under the accession CBUD020000001–CBUD020000057 (contigs) and HG793125–HG793131 (scaffolds).

Abstract

The numerous yeast genome sequences presently available provide a rich source of information for functional as well as evolutionary genomics but unequally cover the large phylogenetic diversity of extant yeasts. We present here the complete sequence of the nuclear genome of the haploid-type strain of *Kuraishia capsulata* (CBS1993^T), a nitrate-assimilating *Saccharomycetales* of uncertain taxonomy, isolated from tunnels of insect larvae underneath coniferous barks and characterized by its copious production of extracellular polysaccharides. The sequence is composed of seven scaffolds, one per chromosome, totaling 11.4 Mb and containing 6,029 protein-coding genes, ~13.5% of which being interrupted by introns. This GC-rich yeast genome (45.7%) appears phylogenetically related with the few other nitrate-assimilating yeasts sequenced so far, *Ogataea polymorpha*, *O. parapolymorpha*, *and Dekkera bruxellensis*, with which it shares a very reduced number of tRNA genes, a novel tRNA sparing strategy, and a common nitrate assimilation cluster, three specific features to this group of yeasts. Centromeres were recognized in GC-poor troughs of each scaffold. The strain bears *MAT alpha* genes at a single *MAT* locus and presents a significant degree of conservation with *Saccharomyces cerevisiae* genes, suggesting that it can perform sexual cycles in nature, although genes involved in meiosis were not all recognized. The complete absence of conservation of synteny between *K. capsulata* and any other yeast genome described so far, including the three other nitrate-assimilating species, validates the interest of this species for long-range evolutionary genomic studies among *Saccharomycotina* yeasts.

Key words: phylogeny, centromere, synteny, noncoding RNA, introgression.

Introduction

During the last decade, evolutionary genomics in yeasts has progressed at a rapid pace based on the comparative analysis of a growing number of species and strain isolates (Dujon et al. 2004; Butler et al. 2009; Liti et al. 2009; Souciet et al. 2009; reviewed in Dujon 2010). But the taxon sampling has remained biased by the considerable attractiveness of *Saccharomyces cerevisiae*, a favored model for experimental studies also used for bread leavening and alcoholic fermentations, and by the preference given to agents of major biotechnological applications or species responsible for human infections. Compared with the broad evolutionary range covered by the multiple phylogenetic lineages of yeasts (Kurtzman et al. 2011), attention was primarily focused on members of the *Saccharomycetaceae* family and, to a lesser extent, to species of the broadly defined "CTG group" comprising yeasts using a slightly modified genetic code (reviewed in Dujon 2010). Few other genome sequences are also available but concern a sprinkled set of species of other phylogenetic branches of the *Saccharomycotina* subphylum (budding yeasts) of *Ascomycota*, such as *Yarrowia lipolytica* (Dujon et al. 2004), *Komagataella pastoris* (De Schutter et al. 2009; Mattanovich et al. 2009; Küberl et al. 2011), *Dekkera bruxellensis* (Curtin et al. 2012; Piškur et al. 2012), or *Ogataea polymorpha* (Ramezani-Rad et al. 2003). Thus, nearly half of all sequenced yeast species today belong to the *Saccharomycetaceae* family and, accordingly, their genomes share a common global architecture with that of *S. cerevisiae* whether they have evolved from the ancestral whole-genome duplication that occurred within this family. The few non-*Saccharomycotina* yeasts sequenced so far are not sufficient to correct this bias. For the basal *Ascomycota* subphylum *Taphrinomycotina* (fission yeasts), genome analysis has been limited so far to the *Schizosaccharomyces* species (Wood et al. 2002; Rhind et al. 2011) and to *Saitoella complicata* (Nishida et al. 2011), whereas for the basidiomycetous yeasts attention has been mostly devoted to strains pathogenic for humans (Loftus et al. 2005; Xu et al. 2007; D'Souza et al. 2011; Gillece et al. 2011) or showing some biotechnological interest (Konishi et al. 2013; Morita et al. 2013). There is, therefore, a need to explore other branches of yeasts in order to gain a more comprehensive and balanced picture of the actual diversity and evolution of their genomes.

To address this question, we have recently undertaken a large sequencing project (Dikaryome) targeting representatives of the unexplored or poorly studied phylogenetic branches from both ascomycetous and basidiomycetous yeasts. As part of this project, we report here the results on one member of the poorly studied nitrate-assimilating yeasts, *Kuraishia capsulata* (Wickerham) (Yamada et al. 1994). This species, previously designated *Hansenula capsulata* (Wickerham 1952) or *Pichia capsulata* (Kurtzman 1984), is usually recovered from frass or tunnels of insect larvae underneath the bark of certain conifers (Kurtzman et al. 2011). It belongs to a genus phylogenetically broadly related to the genera *Citeromyces* and *Nakasawaea* for which no genome data are presently available (Kurtzman et al. 2011) and is even more distantly related to methanol-assimilating yeasts such as *O. polymorpha* and *O. parapolymorpha* (Suh and Zhou 2010; Kurtzman 2011). In *K. capsulata*, conjugation usually precedes ascus formation and ascospores (usually 1–2 per ascus) are hat-shaped. It has been reported that conjugation occurs between a cell and its own bud or the bud of another cell, suggesting homothallism. Another important aspect concerning the physiology of *K. capsulata* is the copious production of extracellular polysaccharides that causes the cell to adhere to bark beetles (Wickerham 1970) and have a drag-reducing activity in vitro (Petersen et al. 1990). Among them, phosphomannan was shown to reduce dental biofilm development (Shimotoyodome et al. 2006). *Kuraishia capsulata* was also the source of a new L-arabino-furanosidase activity (Yanai and Sato 2000) and used for enantioselective chemical reactions (Patel et al. 2004). It is reported as efficient for phosphate assimilation (Breus et al. 2012).

We have determined the complete sequence of the genome of *K. capsulata*-type strain CBS1993^T. It is assembled into seven scaffolds, each corresponding to an identified chromosome, and reveals a total of 6,029 protein-coding genes among which a nitrate-assimilation gene cluster appears as the result of a chromosomal introgression from filamentous fungi. The lack of any closely related species among sequenced yeasts is evident from the absence of gene order conservation, as well as the novel signatures found in centromeres and in noncoding RNAs. Altogether, our results indicate that the *K. capsulata* genome sequence will provide interesting insights into the biology of an interesting, but yet poorly understood, yeast lineage.

Materials and Methods

Strain, Media, Cultivation, and Nucleic Acid Purification

Strain CBS1993^T was isolated from insect frass of a coniferous tree on the shore of Wabatongushi Lake in Ontario, Canada (Kurtzman et al. 2011). A subclone from the CBS stock was grown in YPD medium (yeast extract 10 g/l, Bacto-Peptone 10 g/l, D-glucose 20 g/l) at 30°C to late log phase. Total DNA extracted from this culture was submitted to bisbenzamide CsCl equilibrium gradient centrifugation, resulting in two well-separated bands that were independently purified (supplementary fig. S1, Supplementary Material online). The denser band corresponding to nuclear DNA was used for sequencing. The lighter band, that must correspond to mitochondrial DNA as judged from its low complexity on gel electrophoresis after *Eco*RI digestion, was not sequenced.

Total RNA was purified by acid phenol extraction from cells cultivated in either complete (YPD) or minimal (W0: yeast

nitrogen base 6.7 g/l, D-glucose 20 g/l) media and harvested at either early log or stationary phases. The quality (RNA Integrity Number > 6.5) and quantity of each RNA preparation were assessed using the Bioanalyzer Total RNA Nano Chip. Equimolar amounts of material from the four preparations were eventually mixed and a single cDNA library of polyadenylated mRNA molecules was constructed for sequencing using the Illumina TrueSeq RNA Sample Preparation Kit.

DNA Sequencing, Sequence Assembly, and Correction

Nuclear DNA was sequenced using Roche/454 GSFLX Titanium and Illumina GAIIx. Roche/454 8 kb mate-paired reads were assembled with Newbler (version MapAsmResearch-04/19/2010-patch-08/17/2010) into 73 original contigs (N50 = 371 kb) covering 97% of the initial 701,882 reads, and subsequently linked into 7 scaffolds (N50 = 1.68 Mb) totaling 11.4 Mb using mate-pairs and some manual editing according to Aury et al. (2008). Newbler output files (454Pairstatus.txt and 454Scaffolds.txt) were used to assess the scaffolding (checking both inconsistent links between contigs and placement of contigs inside gaps), and a new scaffold organization was finally proposed.

Illumina short-read sequences were used to lower the high error rate at homopolymer sites observed with pyrosequencing technology, as previously described (Aury et al. 2008). Reads were aligned on the *K. capsulata* genome assembly using SOAP (Li et al. 2008) with a seed size of 12 bp and maximal mismatch and gap size of 2 and 3 nt per read, respectively. Only uniquely mapped reads were retained and sequence differences were taken into consideration solely when 1) not located within the first or last 5 bp of each Illumina read, 2) quality of the considered, previous, and next positions was above 20, and 3) flanking sequences were not homopolymers (to avoid misalignment). Retained differences were piled up when located at the same position and the assembled sequence was eventually corrected if at least three reads detected the same difference and ≥70% of the reads overlapping that position agreed. Given our mapping parameters, several regions of the assembled sequence remained devoid of Illumina tags in the first cycle but became covered with Illumina tags after sequence correction and iteration of the process. The first cycle corrected 350 errors, the second 18, and we stopped error correction at the sixth cycle, when no new errors were found.

Sequence Annotation

Gene models of the *K. capsulata* genome were automatically built by GAZE (Howe et al. 2002) using the four resources described in supplementary methods, Supplementary Material online (protein sequence alignments, RNA sequencing, expressed sequence tags, and ab initio predictions), each affected with a different weight to reflect its reliability and accuracy. This procedure predicted a total of 6,029 protein-coding gene models with 1,063 putative spliceosomal introns in 947 of them. Examination of donor, acceptor, and branch point sequences validated 89% of the automatically predicted introns (see text). Genes for stable noncoding RNAs were identified separately (see supplementary material, Supplementary Material online, and text), and their coordinates were eventually merged with the results of the automated annotation. Conflicts were manually solved for optimal interpretation.

Phylome Reconstruction

The complete collection of phylogenetic trees of all predicted *K. capsulata* proteins was reconstructed as described in Huerta-Cepas et al. (2011). Briefly, a systematic homology search was conducted for each *K. capsulata* protein sequence (seed) against the proteomes of 15 other yeast species, filtering results at e-values <1e−05 and minimal overlaps with hit sequences at 50% of seed length. The 150 top-scoring matches for each seed protein were aligned in both forward and reverse orders (Landan and Graur 2007), using three programs: Muscle v 3.8.31 (Edgar 2004), MAFFT v6.814b (Katoh et al. 2005), and DIALIGN-TX (Subramanian et al. 2008). The six resulting alignments were then combined using M-COFFEE (T-Coffee v8.80) (Wallace et al. 2006) and trimmed using trimAl v1.3 (Capella-Gutiérrez et al. 2009) with a consistency cutoff of 0.1667 and a gap score cutoff of 0.1. Trees were then reconstructed by neighbor joining using BioNJ (Gascuel 1997) with different models (JTT, WAG, MtREV, VT, LG, Blosum62, CpREV, and DCMut). The two best models according to the Akaike information criterion (Akaike 1973) were then used to reconstruct two maximum likelihood trees using phyML (Guindon et al. 2010). In all cases, a discrete gamma-distribution model with four rate categories plus invariant positions was used, and the gamma parameter and the fraction of invariant positions were estimated from the data. A total of 5,052 trees were reconstructed. Trees, alignments, orthology, and paralogy predictions can be found in phylomeDB (http://phylomedb.org, last accessed December 20, 2013) with the PhylD code 181.

Species Tree Reconstruction

To reconstruct the species tree, we selected the 729 protein families in the phylome that displayed one-to-one orthologs in all species considered. These sequences were then concatenated into an alignment comprising 520,513 amino acids. The alignment was then trimmed using trimAl (Capella-Gutiérrez et al. 2009) with gap threshold and conservation parameters both set at 0.5, resulting in a final alignment of 381,280 positions. RAxML version 7.2.6 (Stamatakis et al. 2005) was then used to reconstruct the species tree. LG model (Le and Gascuel 2008) was selected, and a four-category GAMMA distribution was used. Bootstraps were calculated by creating 100 random sequences using SeqBoot from the phylip package. A tree was

Table 1
Numerical Overview of *Kuraishia capsulata* Genome Assembly

Scaffold	Size (bp)	No. of Gaps	Total Gap Size (bp)	GC %	No. of Protein-Coding Genes	Mean Intergenic Size (bp)	No. of tRNA Genes	No. of Other nc RNA Genes
1	2,374,218	20	7,484	45.7	1,266	445	12	19
2	1,753,033	7	1,237	45.6	959	415	10	12
3	1,678,882	8	2,499	45.5	881	530	16	9
4	1,639,918	3	617	45.7	871	504	12	9
5	1,417,141	9	8,670	45.7	738	497	15	6
6	1,378,385	10	3,653	45.6	745	492	9	3
7[a]	1,129,669	9	9,473	45.8	569	525	9	6
Total	11,371,246	66	33,633	—	6,029	—	83	64
Mean	—	—	—	45.7	—	487	—	—

[a]The rDNA locus at the extremity of scaffold 7 is not included in the size indicated.

then reconstructed for each sequence and the consensus tree was inferred using phylip (Retief 2000). A second species tree was reconstructed using a supertree approach based on a gene tree parsimony method that minimizes the number of inferred duplications in all the gene trees contained in the phylome (Wehe et al. 2008).

Detection of Horizontally Acquired Genes

Two complementary methods were used to identify protein-coding genes of *K. capsulata* putatively acquired by recent transfer from prokaryotes. The first one (Marcet-Houben and Gabaldon 2010) is based on the phylome reconstruction (discussed earlier) and included systematic comparison of predicted *K. capsulata* proteins with those of 102 completely sequenced fungi (including yeasts), 95 other eukaryotes and 1,395 prokaryotes (downloaded from KEGG [Kanehisa et al. 2010] as of June 2011). Were only retained as putative candidates for horizontal acquisition genes having homologs in many prokaryotes (>30), less than 10 fungal species, and no other eukaryotes. The second method (Tekaia and Yeramian 2012) involved a multidimensional comparison of proteins predicted from *K. capsulata* and 42 other eukaryotic genomes (mostly fungi) out of which proteins specific to *K. capsulata* or common only to *K. capsulata* and neighboring species (*D. bruxellensis*, *O. polymorpha*, and *O. parapolymorpha*) were compared with a data set of 6,165,075 proteins from 1,942 complete prokaryotic genomes (downloaded from NCBI as of June 2012). Only genes having homologs in many prokaryotes were retained as putative candidates for horizontal acquisition.

Results

Global Genome Architecture and Structural Elements of Chromosomes

The nuclear genome sequence of *K. capsulata* is composed of seven scaffolds totaling 11.4 Mb (excluding rDNA repeats) and ranging in size from 1.1 to 2.4 Mb (table 1). Our final sequence assembly contains 66 gaps whose estimated sizes amount to less than 0.3% of the total genome size. The global GC content is 45.7%, a high figure for yeasts, and is homogeneous throughout the scaffolds with the sole exception of the centromeric regions (discussed later). The sequence-based genome size nicely coincides with its estimate from flow cytometry (11.8 ± 0.8 Mb) and pulsed-field gel electrophoresis (11.5 Mb for seven chromosomal bands), indicating that the strain CBS1993^T is haploid. No mitochondrial DNA sequence was found among large scaffolds in agreement with our method of DNA purification (see Materials and Methods).

Centromeres of *K. capsulata* represent a novel type of structure compared with other yeasts. No homologs of the short centromeres characteristic of the *Saccharomycetaceae* yeasts (Meraldi et al. 2006) were found by searching for CDI and CDEIII consensus sequences separated by AT-rich intervals of the CDEII size range. A similar homology search for the long (3–4 kb) centromeres of the yeasts of the CTG clade is not likely to be meaningful owing to their absence of consensus sequence (Sanyal et al. 2004; Ketel et al. 2009). For *Y. lipolytica*, it was observed that the functional centromeres, containing 17- to 21-bp imperfect palindromes (Yamane et al. 2008), coincide with regions of poor GC content, a property shared with *Clavispora lusitaniae* and *Scheffersomyces stipitis* that also exhibit one region per chromosome with a markedly reduced GC content (Lynch et al. 2010). This composition-bias method to identify putative centromeric region was successfully applied to *Millerozyma sorbitophila* (Leh-Louis et al. 2012). When searching for a similar composition bias in *K. capsulata* (see supplementary material, Supplementary Material online), we initially identified 14 regions, one to three per scaffold, ranging in size between 2 and 6 kb with a composition of 36–39% GC, significantly lower than the average genome value (45.7%). In five of those regions located in scaffolds 1, 2, 3, 5, and 7, we identified a common segment of 200 bp with conserved sequence motifs defined by the regular expression shown in figure 1. This regular

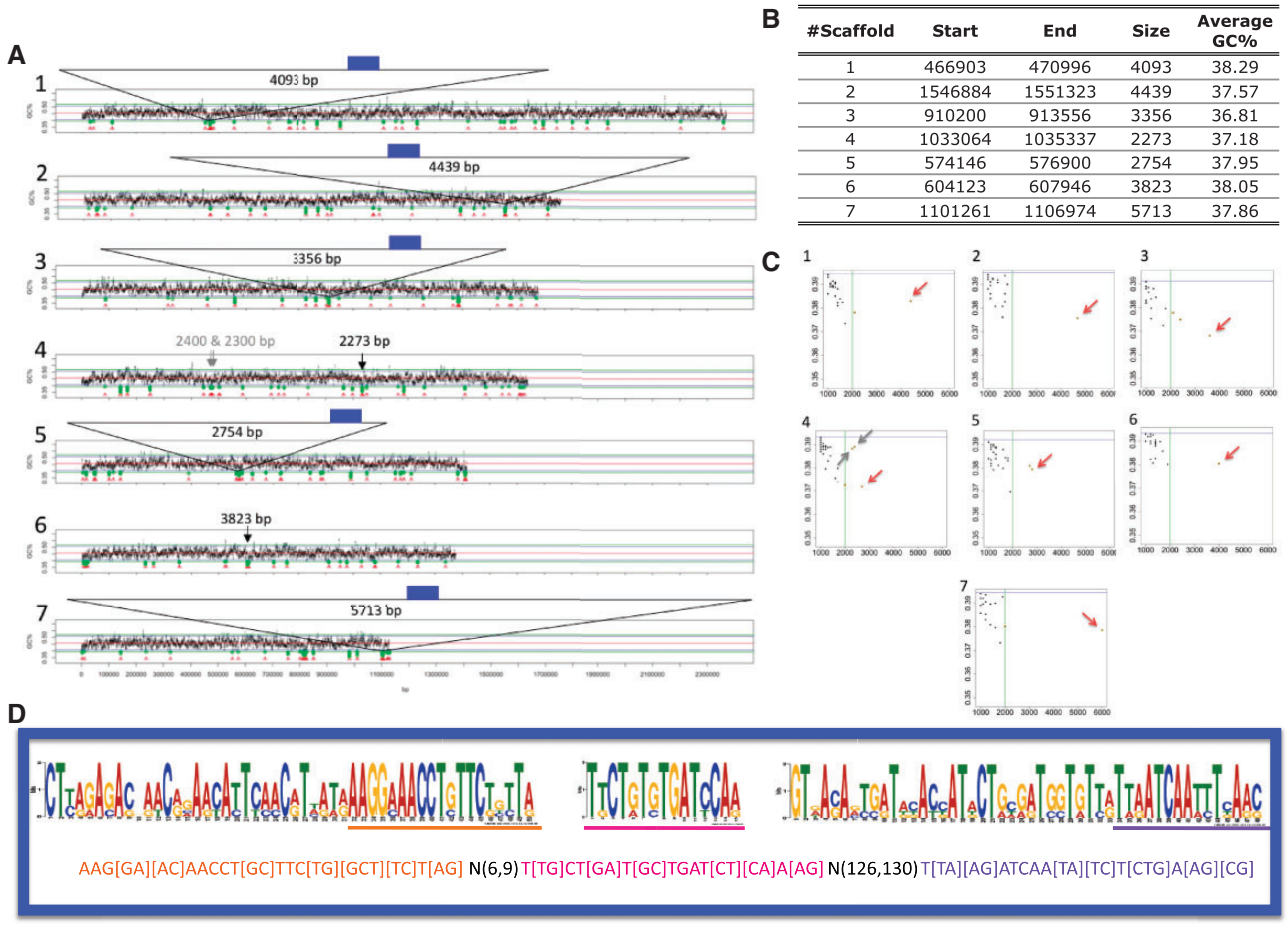

Fig. 1.—Centromeres. (*A*) Compositional variation along the seven scaffolds (as described in Materials and Methods). The nine candidates for centromeric regions are indicated by arrows or by a blue block showing the localization of the conserved motif. For scaffold 4, the chosen candidate among the three is depicted in black, while the other two are in gray (see text). (*B*) A table with the coordinates of the centromeric regions that overlap with the 3C data (see text and supplementary material, Supplementary Material online). The coordinates shown are of the GC-poor troughs trimmed at the edges to avoid overlap with annotated coding sequences. For scaffold 2, a small gene of 237 nt (KUCAT00002123001) predicted by the automatic annotation is located in the middle of the region and completely overlaps. All the other regions do not overlap with predicted coding genes. (*C*) The GC% (ordinates) relative to size (abscissae) of all the GC-poor troughs identified per scaffold. Arrows point to the regions that are proposed as putative centromeric regions by the composition analysis (see text). Note that for the three candidates for scaffold 4 the one chosen is pointed with a red arrow. (*D*) MEME logo of the conserved motif found in the centromeric regions of five scaffolds and the regular expression used to locate them in the genome. The blue rectangles of panel (*A*) correspond to this motif.

expression identified no other sequences in the entire *K. capsulata* genome, making it an attractive candidate for a conserved element of active centromeres, despite its absence from scaffolds 4 and 6. Thus, a total of nine putative centromeres were finally proposed for the seven scaffolds: the five GC-poor regions containing the conserved sequence motif for scaffolds 1, 2, 3, 5, and 7 and the unique GC-poor region of scaffold 6 represent likely centromeres for the six corresponding chromosomes, while three possible candidates remain for scaffold 4. This in silico prediction was validated experimentally by the analysis of data from chromosomal conformation capture experiments (3C) as described in supplementary material, Supplementary Material online. All predictions were confirmed for scaffolds 1, 2, 3, 5, and 7, and one of our three prediction was confirmed for scaffold 4 (fig. 1*B*). Note, however, that the proposed centromere for scaffold 7 lies next to the rDNA locus, an intriguing position given the opposite location of the centromeric pole and the nucleolus in interphasic nuclei of *S. cerevisiae* (reviewed in Zimmer and Fabre 2011).

While telomeric sequence repeats could be characterized in detail in *S. cerevisiae* and a few other yeast species by the combination of experimental procedures with sequence analyses (Louis 1995), shotgun sequence assembly is poorly adapted to identify these sequences owing to the repetitive nature of subtelomeric regions. Accordingly, no short repeats

of telomerase-generated sequences could be found in any of our 14 scaffold ends, raising the possibility that our sequence assembly does not reach actual chromosome ends. Comparing, however, the last 30 kb at all scaffold ends, we observe a frequent occurrence of large gene families (discussed later), a characteristic feature of subtelomeric regions in other yeasts (Fairhead and Dujon 2006). Among them are genes encoding proteins with a major facilitator superfamily domain likely corresponding to sugar transporters, proteins with a phytanoyl-CoA dioxygenase InterPro domain, proteins with amino acid/polyamine transporter and amino acid permease InterPro domains, and proteins with the Zn(2)-C6 fungal-type DNA-binding InterPro domain and the DUF3468 domain of unknown function.

Stable Noncoding RNA Genes

rRNA Genes

The sequence at the right end of scaffold 7 contains the genes for the 18S, 5.8S, and 23S rRNAs, forming a complete rDNA repeat unit, plus one gene for the 5S RNA in parallel orientation. Excess sequencing coverage of this region, compared with other parts of the genome, indicates the existence of ~30 copies of the rDNA–5S RNA repeat unit at this locus. No other rDNA sequences were identified in the seven scaffolds, indicating that *K. capsulata* has a single rDNA locus in a subtelomeric location. Interestingly, two additional loci (positions 41,541–41,661 of scaffold 4 and 645,264–645,384 of scaffold 7) contain copies of 5S RNA genes (according to sequence coverage, three gene copies are likely present at the scaffold 4 locus and one on scaffold 7). The existence of these additional but dispersed copies of 5S RNA genes places *K. capsulata* as an interesting intermediate between *Y. lipolytica* and related species where all 5S RNA genes are dispersed in the genome (like in most eukaryotes) and the other *Saccharomycotina* yeasts sequenced so far where one gene of 5S RNA (occasionally two) lies in reverse orientation between each rDNA repeat unit.

tRNA Genes

In *Saccharomycotina* genomes, tRNA genes vary in number from over 500 in *Y. lipolytica* to only 131 for *C. albicans* to encode an almost constant set of 42–46 species of tRNA molecules (Marck et al. 2006; Man and Pilpel 2007; Souciet et al. 2009). The genome of *K. capsulata* shows an even lower gene number with a total of only 83 tRNA genes (supplementary table S1, Supplementary Material online). This unusually low figure results from the reduction of the number of identical paralogs, preserving a set of 44 different tRNA species to interpret the genetic code. The most frequently redundant tRNA species in other yeasts (Phe-tRNA$^{[GAA]}$, Leu-tRNA$^{[CAA]}$, Ile-tRNA$^{[AAU]}$, Val-tRNA$^{[AAC]}$, Ser-tRNA$^{[AGA]}$, Pro-tRNA$^{[UGG]}$, Thr-tRNA$^{[AGU]}$, Ala-tRNA$^{[AGC]}$, Tyr-tRNA$^{[GUA]}$, His-tRNA$^{[GUG]}$, Gln-tRNA$^{[UUG]}$, Asn-tRNA$^{[GUU]}$, Lys-tRNA$^{[CUU]}$, Glu-tRNA$^{[UUC]}$, Arg-tRNA$^{[UCU]}$, and Gly-tRNA$^{[GCC]}$, Marck et al. 2006) are encoded by only two or three genes in *K. capsulata*, and only the Asp-tRNA$^{[GUC]}$ is encoded by four genes. Note that similarly reduced numbers of tRNA genes are observed for *O. polymorpha*, *O. parapolymorpha*, and *D. bruxellensis*, (supplementary table S2, Supplementary Material online), supporting a closer relationship of these three species with *K. capsulata* (discussed later), than is the case for *Ko. pastoris* (123 or 124 tRNA genes).

The tRNA genes are uniformly distributed throughout the seven scaffolds with an average distance of ~109 kb (median ~60 kb) as expected from their putative role in global nuclear organization (Iwasaki and Noma 2012). Only seven pairs of successive tRNA genes are separated from each other by less than 1 kb and, in each case, the two members of the pair correspond to distinct species, ruling out tandem gene duplication. Among them, the pair Ile-tRNA$^{[AAU]}$–Gln-tRNA$^{[UUG]}$ on scaffold 7 is a good candidate for a transcriptional fusion as observed in other yeasts (Dujon et al. 2004; Souciet et al. 2009) as the interval between the two genes is only 4 nt. Note that free-standing gene copies identical in sequence to each of the two fused tRNA genes are also present in the genome. This fusion is rare among all yeasts sequenced so far. It has been observed only in *Naumovozyma castelli* and in *Y. lipolytica*, two very distantly related yeasts (Marck et al. 2006). In contrast, the Arg-tRNA$^{[UCU]}$–Asp-tRNA$^{[GUC]}$ fusion common to most *Saccharomycetaceae* is not found in *K. capsulata*.

More than 20% of all tRNA genes (18/83, corresponding to 10 distinct tRNA species) have introns (supplementary table S2, Supplementary Material online), a proportion commonly observed in other yeasts (Marck et al. 2006), but the presence/absence of introns in the different tRNA species is only partially conserved. Note that the proportion of split tRNA genes is much lower in *O. polymorpha*, *O. parapolymorpha*, and *D. bruxellensis*. The tRNA genes of *K. capsulata* are very GC-rich (69.5 % on average) with compositions reaching over 75% for Asn-, Thr-, and most Ser-tRNAs.

As in other yeasts, *K. capsulata* tRNA species reading the Val, Ser (UCN), Pro, Thr, and Ala four-codon families share the eukaryotic signature with an A at position 34 (first position of anticodon) to read codons ending with U or C instead of a G as in bacteria. The same is true for the Leu (CUN) codons (fig. 2), as in *Y. lipolytica* and other fungi, while yeasts of the *Saccharomycetaceae* family use the bacterial tRNA sparing strategy (with a G at position 34), and yeasts of the CTG group have Ser-tRNA reading the CUG codon. Interestingly, for the Arg (CGN) codons, however, *K. capsulata* shares the bacterial sparing strategy common to the *Saccharomycetacae* and the CTG yeasts with only two tRNAs bearing A_{34} and C_{34}, respectively, instead of A_{34} and U_{34} as in *Y. lipolytica* or three tRNAs as in other eukaryotes (fig. 2). The same is true for *O. polymorpha*, *O. parapolymorpha*, and *D. bruxellensis*, while *Ko. pastoris* conserved the eukaryotic type of tRNA sparing

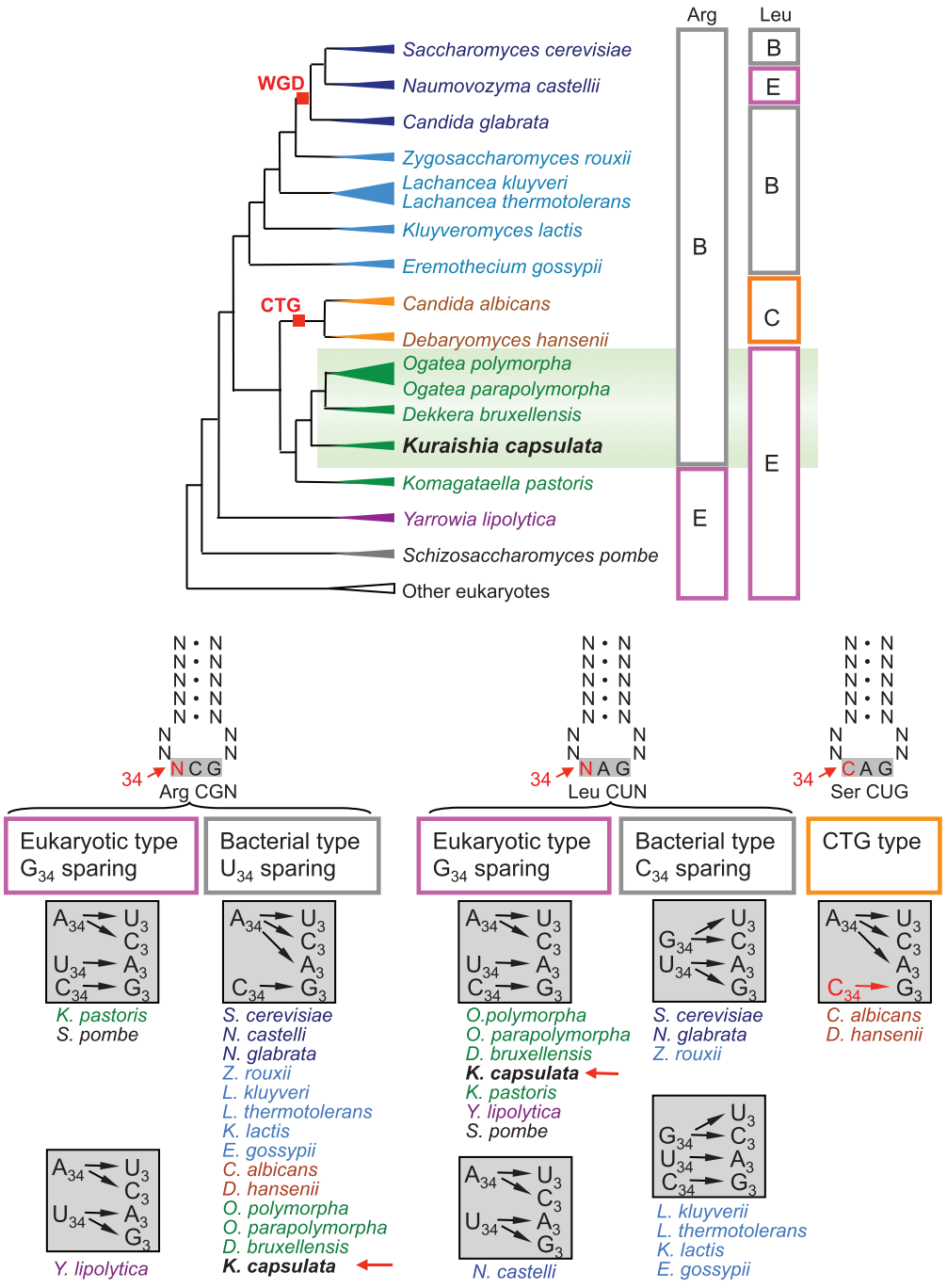

Fig. 2.—Arg (CGN) and Leu (CUN) tRNA sparing strategies identify *Kuraishia capsulata* and related species as an evolutionary intermediate within *Saccharomycotina*. Bottom: Sparing strategies for the Arg (CGN) and Leu (CUN) codon families as deduced from the set of tRNA genes identified in genomes of the various yeast species listed (blue: *Saccharomycetaceae*, orange: CTG clade, green: *K. capsulata* and related lineages, purple: *Y. lipolytica*, black: *Taphrinomycotina*). Top: summary of sparing strategies related to yeast phylogeny. Tree topology is deduced from figure 4 and Dujon (2010). Note the intermediary position of the *Kuraishia* and related yeasts with respect to other clades.

strategy for these codons. This places *K. capsulata* and the three related yeasts whose sequences are presently available at an interesting intermediate position between the more extensively studied yeast lineages and confirms *Ko. pastoris* as an outgroup.

Other Stable ncRNA Genes

Genes for other ncRNAs of *K. capsulata* were identified by their structural homology with those of other *Saccharomycotina* (supplementary methods, Supplementary

Material online). The five spliceosomal snRNAs (U1, U2, U4, U5, and U6) are each encoded by a single gene, as well as the U3 snoRNA and the RNA moieties of the RNAse P, MRP, and SRP complexes (supplementary table S3, Supplementary Material online). *Kuraishia capsulata* SRP RNA is composed of 281 residues (supplementary fig. S2, Supplementary Material online), which is a regular size for eukaryotes (Rosenblad et al. 2009). Compared with *S. cerevisiae* and related species, *K. capsulata* SRP RNA lacks helices 9, 11, and 12, and helices 7 and 10 are truncated. All other conserved motifs are present in the *K. capsulata* SRP RNA. *Kuraishia capsulata* RNase P RNA is very similar to its ortholog in *S. cerevisiae* (except for a possible insertion of 6 nt in helix P1), and all conserved motifs of RNase P molecules are present (supplementary fig. S3, Supplementary Material online). In contrast, *K. capsulata* MRP RNA significantly differs from that of *S. cerevisiae* (supplementary fig. S4, Supplementary Material online). Two large extra helices are found in the specificity domain 2 (between helices ymP6 and ymP7). This is reminiscent of the MRP RNA of several *Pezizomycotina* species (Piccinelli et al. 2005). No helix eP15 or k-turn motif could be found. Other conserved motifs are present.

Fifty-six snoRNAs (39 C/D, 17 H/ACA) could be predicted in the sequence of *K. capsulata* (supplementary table S3, Supplementary Material online). This is ~75% of the number of snoRNA genes recorded in *S. cerevisiae* (http://people.biochem.umass.edu/sfournier/fournierlab/snornadb/mastertable.php, last accessed December 20, 2013) and represents a comparable proportion to what was found before for other *Saccharomycotina* yeasts (Cruz and Westhof 2011). As in *S. cerevisiae*, snR38, snR44, snR191, U18, and U24 are encoded within spliceosomal introns in *K. capsulata*. The same is probably true for snR54 and snR58 as each map between two annotated CDS displaying similarity with, respectively, the 5′ and 3′ parts of a same homolog, suggesting overlooked introns in the automated annotation (discussed later). We did not find the homologs to the intronic snR59 and snR89 genes of *S. cerevisiae*. Polycistronic clusters of snoRNA genes are also conserved in *K. capsulata*, except that snR72 could not be found in cluster 1, and snR53 and snR128 could not be accurately predicted preventing us to identify clusters 4 and 5, respectively.

Protein-Coding Genes, Spliceosomal Introns, and Protein Families

A total of 6,029 protein-coding genes were predicted from the sequence, based on automated annotation (see supplementary methods, Supplementary Material online). Coding sequences cover 73.5% of the total genome size (excluding rDNA), a figure in good agreement with other *Sacchromycotina* yeasts (Dujon 2010). A total of 1,063 spliceosomal introns were initially predicted to interrupt the reading frames of 947 protein-coding genes (847 genes have a single intron). Intron sequences were examined to identify junction motifs and branch points as in other yeasts (supplementary methods, Supplementary Material online). Note that nine automatically predicted introns have incomplete internal sequences (presence of Ns), preventing us to examine the presence of branch points, but from their sizes, donor, and acceptor sites, only one of them is a likely candidate for an actual intron. Canonical branch points were found within the distal one-third of their sequence for 973 of the remaining 1,054 putative introns (supplementary table S4, Supplementary Material online), leaving out only 81 dubious introns. The most frequent branch point, UACUAAC, and its shorter derivatives, •ACUAAC and ••CUAAC, are very typical for yeast introns. Conserved canonical sequences typical for yeasts could also be recognized at donor and acceptor sites (supplementary table S4, Supplementary Material online). Combining these data, a conservative estimate indicates the existence of 901 actual introns inserted within the reading frames of 812 protein-coding genes. *Kuraishia capsulata* appears, therefore, as an intron-rich yeast with 13.5% of its protein-coding genes interrupted by introns (71 genes have two introns, 6 genes have three introns, and 2 genes have four introns, a distribution similar to other *Saccharomycotina* yeasts). With their short size (from 46 to 519 nt, median of 62) spliceosomal introns of *K. capsulata* are also consistent with other *Saccharomycotina* yeasts.

Many predicted proteins of *K. capsulata* could be classified into major functional categories as previously defined (Tatusov et al. 1997), and the distribution of the gene numbers within categories could be compared with those of *O. polymorpha* and *D. bruxellensis*, its two most closely related species (discussed later), as well as with *Ko. pastoris*, their immediate outgroup. Functional categories of *K. capsulata*, *O. polymorpha*, and *D. bruxellensis* show similar distributions, while for *Ko. pastoris* categories related to secretion and RNA processing are underrepresented and categories involved in transcription, translation, and replication are overrepresented (fig. 3). *Kuraishia capsulata* genome comprises 641 genes related to carbohydrate metabolism, including the alpha-L-arabinofuranosidase (Yanai and Sato 2000). Future investigation of this complement may help to understand the mechanisms underlying the copious production of extracellular polysaccharides by this yeast, including the synthesis of the most abundant of them, phosphomannan, for which the responsible enzyme remains to be identified (Brettauer et al. 1969).

Phylogenomic Analysis

To gain insight into the evolutionary history of *K. capsulata*, we reconstructed the evolutionary histories of all of its protein-coding genes using the PhylomeDB pipeline (see Materials and Methods). The resulting 5,052 gene trees were automatically scanned to predict orthology and paralogy relationships (Gabaldón 2008) and to date duplication events (Huerta-

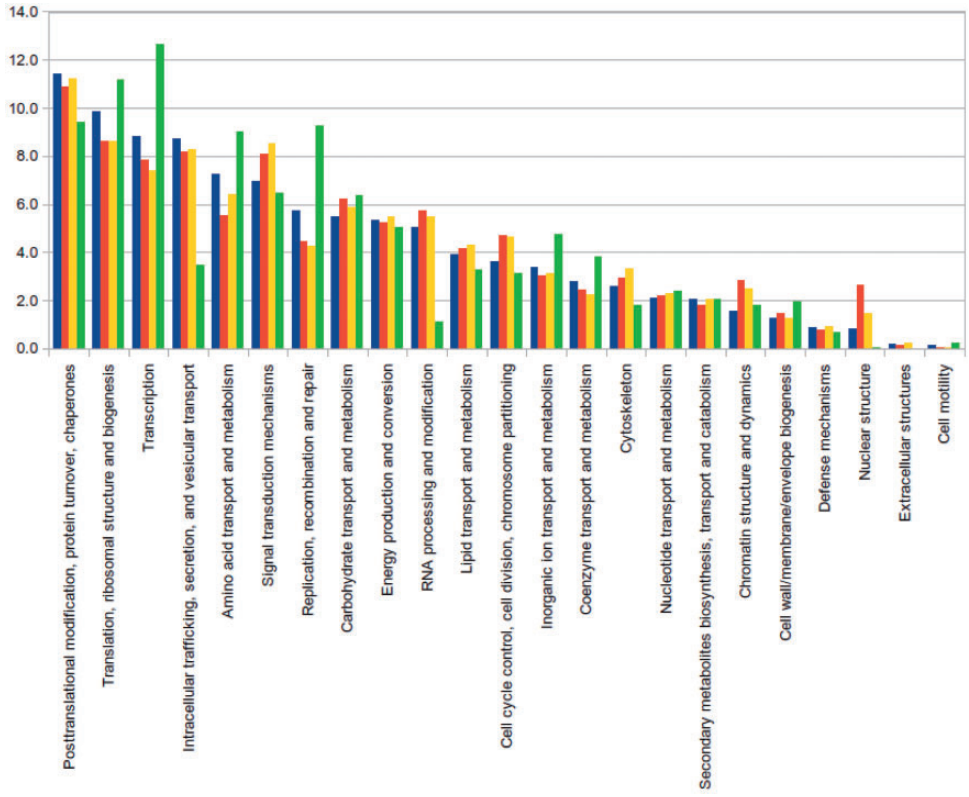

Fig. 3.—Functional categories of predicted proteins. Percentages (ordinate) of protein-coding genes classified in each of the functional categories described by Tatusov et al. (1997) in the four species: *Kuraishia capsulata* (blue), *Dekkera bruxellensis* (orange), *Ogataea polymorpha* (yellow), and *Komagatella pastoris* (green). Note the difference of the last species against the homogeneity of the first three.

Cepas and Gabaldón 2011). Results are available online at Phylome DB. About 64% of all predicted proteins of *K. capsulata* are each encoded by a single gene, while the rest are encoded by families of 2–85 paralogous genes (see supplementary fig. S8, Supplementary Material online). This leads to a relatively high global redundancy (1.37) for this genome compared with protoploid members of the *Saccharomycetaceae* family, but similar to what is observed in *Debaryomyces hansenii* or *Y. lipolytica* (Souciet et al. 2009). In total, 531 gene families were identified, of which 300 are composed of only two members and 105 of three members (see supplementary table S10, Supplementary Material online). Among the few large families ($\geq$10 members), some correspond to species-specific expansions. The three largest families of 85, 54, and 48 members, respectively, distributed throughout the genome, encode proteins with a ser/thr protein kinase domain, sugar transporters, and proteins with WD40 repeats. Few of these genes, however, were duplicated specifically in *K. capsulata* as evidenced by the low duplication rate found in this genome (0.15 duplication per gene, fig. 4). Some of the lineage-specific duplications have occurred in tandem, leading to 58 gene arrays. Most of the tandem repeats are pairs, although eight triples and two quadruples were also observed (supplementary table S5, Supplementary Material online). We identified 10 *K. capsulata*-specific gene expansions formed by families of 5–16 (in)paralogs (supplementary table S6, Supplementary Material online). Note that three expanded families encode proteins annotated as transporters and three others as oxidoreductases. The remaining families, which account for the large majority, are composed of dispersed genes, corresponding to ancient duplications and genome rearrangements.

Two complementary methods were used to assess the phylogenetic position of *K. capsulata*: 1) construction of maximum likelihood tree from a concatenated alignment comprising 729 proteins whose genes are single copy orthologs in all species considered and 2) reconstruction of a supertree from the 5,052 individual gene trees (discussed earlier) by the gene tree parsimony approach implemented in duptree. The two methods resulted in the same topology (fig. 4) in which *K. capsulata* groups next to the *D. bruxellensis* and *O. polymorpha* common ancestor. *Komagataella pastoris* represents the closer outgroup to these three genomes. The four species form a broad group that separated from the CTG clade after the separation of their common ancestor from the *Sacchromycetaceae* yeasts. This topology is congruent, for

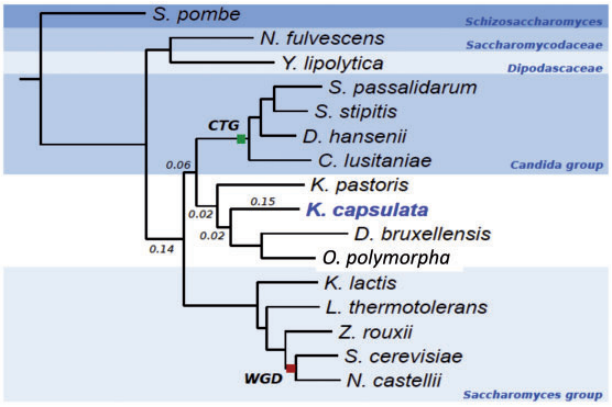

Fig. 4.—Phylogenetic classification of *Kuraishia capsulata*. The figure represents the phylogeny of *K. capsulata* relative to 14 other *Saccharomycotina* species (*Saccharomyces cerevisiae*, *Naumovozyma castelli*, *Zygosaccharomyces rouxii*, *Lachancea thermotolerans*, *Kluyveromyces lactis*, *Ogataea polymorpha*, *Dekkera bruxellensis*, *Komagatella pastoris*, *Clavispora lusitaniae*, *Debaryomyces hansenii*, *Scheffersomyces stipitis*, *Spathaspora passalidarum*, *Yarrowia lipolytica*, and *Nadsonia fulvescens*) and the *Taphrinomycotina* yeast *Schizosaccharomyces pombe* used as outgroup. See text for the phylogenetic reconstruction methods used. All bootstrap values (not shown) were 100. Numbers on the branches represent the duplication frequency (duplications per gene) found at a given branch in the lineages leading to *K. capsulata*. Major evolutionary events are marked on the tree by red and green squares (whole genome duplication [WGD] and alteration of the genetic code [CTG], respectively).

the shared species, with previous global fungal phylogenies (Fitzpatrick et al. 2006; Marcet-Houben and Gabaldón 2009) and with above results on noncoding RNA genes. It validates a posteriori the choice of *K. capsulata* for genome analysis.

The genome of *K. capsulata* shows very little conservation of synteny with *D. bruxellensis*, *O. polymorpha*, and *O. parapolymorpha*, its three closer relatives whose complete genome sequences are available (supplementary fig. S5, Supplementary Material online). Average sizes of maximal unique matches obtained by MUMmer PROmer (supplementary material, Supplementary Material online) are similar to average gene sizes indicating absence of synteny. Comparison of *K. capsulata* against itself also indicates the absence of segmental duplications. By comparison, *O. polymorpha* and *O. parapolymorpha* have kept an almost complete synteny (supplementary fig. S6, Supplementary Material online) in agreement with the fact that the two species are very close to each other (Kurtzman 2011). However, they both share only very short blocks of conserved synteny with *D. bruxellensis*, clearly indicating the broad evolutionary range covered by this group of yeasts.

Despite this fact, *K. capsulata* and related species mentioned above form a homogeneous group of yeasts clearly distinguished from other sequenced yeasts when systematic pairwise comparisons of predicted proteins are submitted to correspondence analysis (supplementary fig. S7, Supplementary Material online). Note the presence of a tight cluster including *K. capsulata* close to the center of the first two dimensions of the analysis (representing together 40% of the available information). This cluster is well separated from the two other tight clusters corresponding, respectively, to the 14 *Saccharomycetaceae* species and to the 8 CTG species included in the analysis. Interestingly, the *K. capsulata* cluster is relatively close to the position occupied by *Y. lipolytica* in this analysis, in agreement with observations made on noncoding RNA and number of spliceosomal introns (discussed earlier). In agreement with established phylogenies, *Taphrinomycotina*, basidiomycetous yeasts, and filamentous fungi occupy clearly different positions in this analysis from the three major groups of *Saccharomycotina* yeasts now recognized: the *Saccharomycetaceae*, the CTG, and the group illustrated by *K. capsulata* and related yeasts.

Nitrate Assimilation Pathway

Unlike *S. cerevisiae* and many other *Saccharomycotina* species that are unable to use nitrate as the sole nitrogen source, *K. capsulata* is a nitrate-assimilating yeast (Kurtzman et al. 2011). In plants and filamentous fungi, nitrate assimilation requires three major biochemical activities ensuring uptake of nitrate and the subsequent reduction of nitrate and nitrite (Crawford and Aarst 1993). In *Emericella nidulans*, a filamentous fungus of the *Pezizomycotina* subphylum of *Ascomycota*, three genes control these activities, respectively, *crnA*, *niaD*, and *niiA* forming a cluster (Johnstone et al. 1990). A similar cluster was found in *O. polymorpha* (Brito et al. 1996; Siverio 2002), whereas almost all other yeasts sequenced so far only carry two dispersed copies of the *niaD* gene alone. The annotated sequence of *K. capsulata* reveals orthologs of the three genes *crnA*, *niiA*, and *niaD* (KUCAT00000680001, KUCAT00000681001, and KUCAT00000684001, respectively), forming a ~12.5 kb cluster on scaffold 1 (coordinates 1,241,508–1,254,021) that also contains two additional genes (KUCAT00000682001 and KUCAT00000683001) encoding putative transcription factors (fig. 5). *Dekkera bruxellensis* also contains the same gene cluster, but not *Ko. pastoris*. In addition to this cluster, the genomes of *K. capsulata* and the two related yeasts also contain two dispersed copies of the *niaD* gene that group phylogetically with the *niaD* genes of the other *Saccharomycotina* yeasts and some *Pezizomycotina* species (fig. 5), suggesting an ancient duplication of *niaD* in fungi, conserved in all yeasts despite their lack of the nitrate assimilation cluster. In contrast, the phylogenetic proximity of the *niaD* gene present in the nitrate assimilation cluster in *K. capsulata* and the two related species with the *niaD* genes of the nitrate assimilation clusters of the *Pezizomycotina* species strongly suggests that a transfer of a chromosomal segment bearing this gene cluster occurred in the common ancestor of *K. capsulata*, *O. polymorpha*, and *D. bruxellensis* after its separation from *Ko. pastoris*, creating a

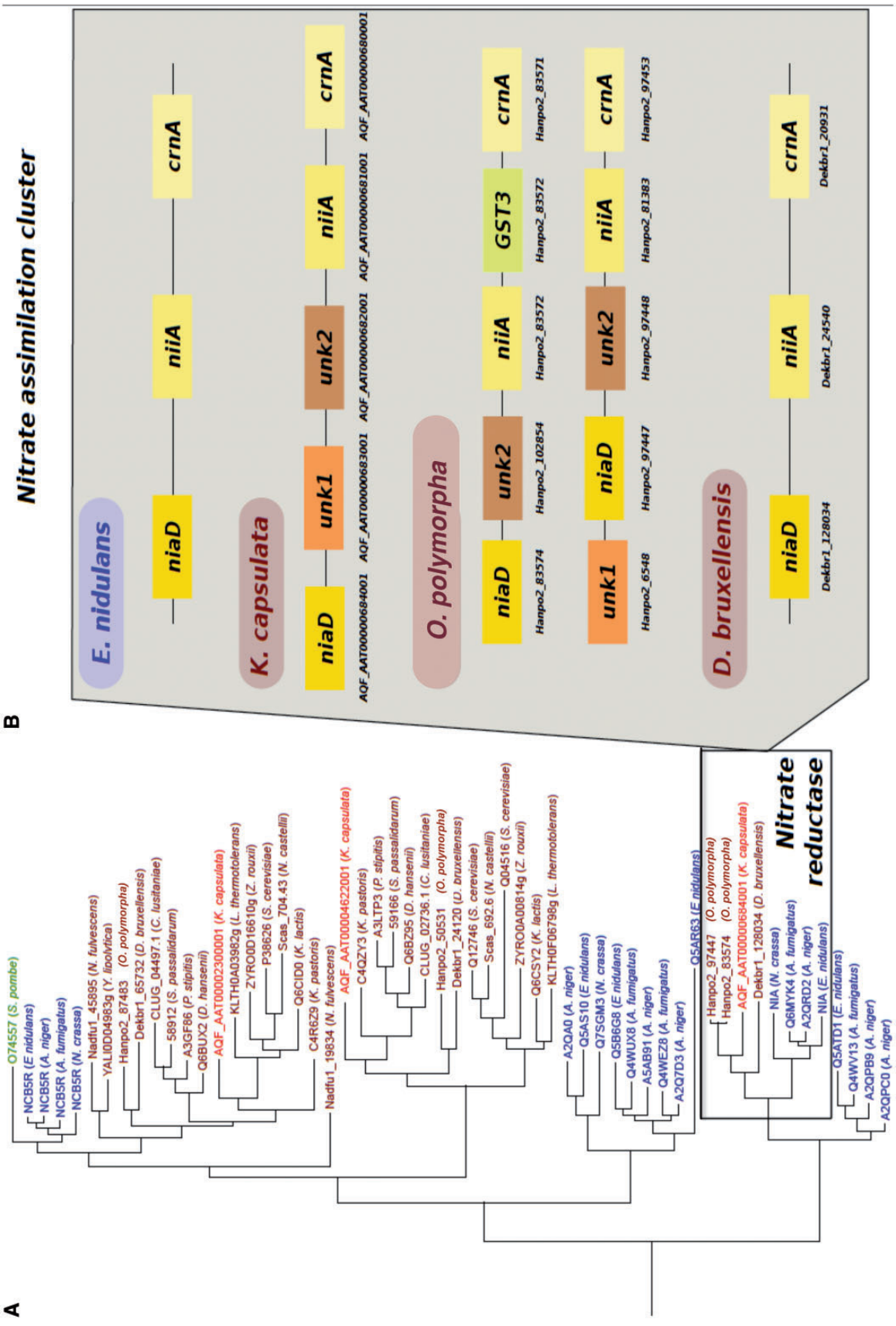

Fig. 5.—Nitrate assimilation cluster and other *niaD* genes. Left panel: Phylogeny of the members of the *niaD* gene family in the 16 yeast species shown in figure 4 plus seven *Pezizomycotina* species (*Neurospora crassa*, *Magnaporthe onyzae*, *Gibberella zeae*, *Histoplasma capsulatum*, *Emericella nidulans*, *Aspergillus fumigatus*, and *Stagonospora nodorum*). Genes from *Pezizomycotina*, *Saccharomycotina*, and *Taphrinomycotina* species are labeled, respectively, in blue, brown, and green. Note the clustering of the *NiaD* genes present in the nitrate assimilation cluster. Right panel: Schematic representation of the nitrate assimilation gene cluster in *Kuraishia capsulata* and its two related yeast species, compared with the filamentous fungus *E. nidulans*. *NiaD*, *niiA*, and *crnA* encode, respectively, the nitrate reductase, nitrite reductase and nitrate transporter. The cluster lies on scaffold 1 in *K. capsulata* (coordinates 1,241,508 to 1,254,021 from right to left of the schema).

lineage of nitrate-assimilating yeasts (*Ko. pastoris* does not assimilate nitrate). Individual transfer of the *crnA*, *niiA*, and *niaD* genes (as is found for genes of bacterial origin, discussed later) followed by the reformation of a cluster appears much less likely. The introgression of large chromosomal segments of foreign origin has now been observed in many yeast genomes (reviewed in Morales and Dujon 2012). The long evolutionary distances between many donors and recipients and the presence of multiple insertions with circular permutations in some cases (Borneman et al. 2011; Galeote et al. 2011) suggest a mechanism of transfer ignoring regular hybridizations, such as the transient formation of interspecies heterokaryons (Morales and Dujon 2012). A similar mechanism may have taken place between a filamentous fungus and the ancestor of our nitrate-assimilating yeasts. Note that the nitrate assimilation cluster is duplicated in *O. polymorpha* (with internal gene rearrangements) in a similar manner as the multiple insertions of the 17-kb segment of *Zygosaccharomyces bailli* into the genome of *S. cerevisiae* wine strains (Borneman et al. 2011; Galeote et al. 2011).

Genes Involved in Mating, Meiosis, and Sexual Cycle

Beside cytological observations suggesting that cellular fusion occurs in *K. capsulata* immediately followed by meiosis (Kurtzman et al. 2011), there is no direct evidence that this yeast can undergo a complete sexual cycle. On scaffold 2, we identified two adjacent genes (KUCAT00001759001 and KUCAT00001760001) encoding, respectively, the putative *MAT*alpha1 (196 aa) and *MAT*alpha2 (214 aa) transcription factors based on their homology with previously characterized yeast proteins (fig. 6). Note that the *MATalpha2* gene of *K. capsulata* contains three introns, contrary to a maximum of two for the *Saccharomycetaceae* genes known so far and is cooriented with the *MATalpha1* gene, contrary to the situation observed in *Saccharomycetaceae* with two diverging genes but in agreement with its orientation in *O. polymorpha* and *Ko. pastoris*. No additional *MAT* gene could be detected in the entire *K. capsulata* genome, suggesting that haploid strain CBS 1993^T is of mating-type *alpha* and cannot undergo mating-type switching. Additional data are needed to determine whether *MATa* strains exist among the different *K. capsulata* isolates.

The two *MAT* genes are flanked by homologs of *SLA2* and *DIC1*, as previously observed in other yeasts (*Torulaspora delbrueckii* and *Lachancea kluyveri*), a gene arrangement considered as ancestral and prone to evolutionary erosion in yeast species undergoing mating-type switching (Gordon et al. 2011). The same arrangement is also observed for *Ko. pastoris* (with an inverted *DIC1* orientation) and *O. polymorpha* (fig. 6). In the latter case, however, the *MAT* locus also contains a *MAT a1* gene. The simultaneous presence of *MATa* and *MATalpha* genes at the same locus has also been observed

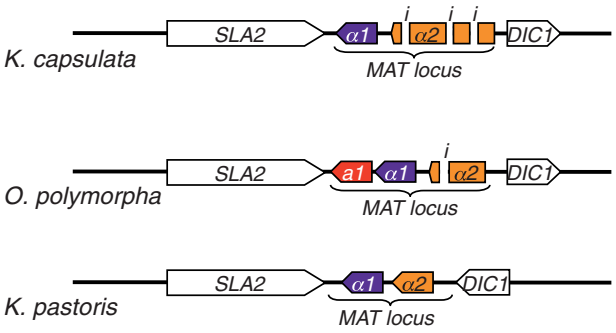

Fig. 6.—Schematic representation of *MAT* locus and flanking genes in *Kuraishia capsulata* and related species. Note the introns in the *MAT alpha2* gene and the presence of a *MAT a1* gene in *Ogataea polymorpha*. In *Saccharomyces cerevisiae*, *SLA2* encodes a transmembrane actin binding protein involved in membrane cytoskeleton assembly and cell polarization and *DIC1* encodes a mitochondrial dicarboxylate carrier.

for yeasts of the CTG clade such as *De. hansenii*, *Sc. stipitis*, and *M. sorbitophila* (Leh-Louis et al. 2012).

Most genes involved in the primary steps of the sexual cycle in *S. cerevisiae* (pheromone production and modification, pheromone sensing, and signal transduction) have homologs in *K. capsulata* (supplementary table S7, Supplementary Material online), suggesting similar functional pathways. A gene (KUCAT00004947001) encoding the precursor of the alpha factor pheromone was also detected in *K. capsulata*, with its characteristic internal duplications. The gene for the precursor of the a-factor pheromone escaped our detection, probably due to its smaller size and weaker conservation, but we suspect its existence because genes for the MATa pheromone receptor and for enzymes involved in its farnesylation, methylation, and maturation were found. No gene was detected corresponding to an endonuclease of the HO type or similar LAGLIDADG intein.

Only 86 of the 135 genes encoding proteins involved in meiosis in *S. cerevisiae* could be identified in *K. capsulata* (supplementary table S8, Supplementary Material online). Note that the important meiosis-specific gene *SPO11* is present, although it was overlooked by the automated annotation. The most frequent lack of ortholog identification concerns genes encoding proteins of the synaptonemal complex and, to a lesser proportion, those involved in spore assembly and chromosome cohesion/recombination. Proteins active in heteropolymeric complexes may show differential conservation, as previously noted by Richard et al. (2005). In the case of the MRX complex, for example, *K. capsulata* has homologs to *MRE11* and *RAD50* but no homolog to *XRS2*. The same was observed for *De. hansenii* and *Y. lipolytica* (Richard et al. 2005). Similarly, no homolog to *DMC1* or *MEI5*, two elements of the meiosis-specific complex Mei5p/Sae3p protein complex mediating Dmc1p activity (Ferrari et al. 2009), could be found in the *K. capsulata* genome, while a poorly conserved homolog of *SAE3* is present.

Horizontal Gene Transfers from Prokaryotes

Application of published methods to identify putative horizontal transfers (see Materials and Methods) revealed 13 *K. capsulata* genes of probable bacterial origin distributed on six of the seven scaffolds (supplementary table S9, Supplementary Material online). This list is not necessarily exhaustive. These genes were identified from their high number of homologs among bacterial genomes and their absence or only occasional presence among other yeast or fungal genomes. A large proportion of them (8 of 13) are in subtelomeric locations (scaffolds 2 and 6), in tandems (scaffold 3), or form a pair of dispersed paralogs (scaffolds 1 and 6). These features are consistent with recent acquisitions observed in other yeasts (Rolland et al. 2009). The two genes expected to encode a 5-formyltetrahydrofolate cyclo-ligase (scaffold 1) and a chitinase-like protein (scaffold 4) may have been acquired in an ancestor of *K. capsulata* and related species because they are also found in *O. parapolymorpha*. The lack of synteny between the two species, however, does not allow a definitive conclusion because horizontal transfers from prokaryotes may be recurrent in yeast evolution, especially for genes encoding metabolic functions of possible immediate usefulness (Marcet-Houben and Gabaldón 2010). This was already documented for genes coding proline racemase, for example (Fitzpatrick et al. 2008). The commonality of six *K. capsulata* gene candidates with other *Saccharomycotina* yeasts (*Sc. stipitis* for a phospho-lipid binding protein, *Blastobotrys adeninivorans* for beta-lactamase, *Y. lipolytica* for arsenate reductase, *Ko. pastoris* for NADP oxidoreductase, *K. lactis* for a protein of unknown function, and *S. cerevisiae* for a protein with glutathione *S*-transferase domain) suggests such recurrent transfers (supplementary table S9, Supplementary Material online). The same is true for the commonality with other *Ascomycota* (*Botrytis cinerea*, *Grosmannia clavigera*, and *Fusarium verticilloides*) as well as with the basidiomycetous yeast *Cryptococcus neoformans*. The three gene copies for NADP oxidoreductase forming a common tandem array on scaffold 3 are reminiscent of the frequent amplification of horizontally acquired genes in other yeast genomes (Rolland et al. 2009).

Discussion

The genome of *K. capsulata* documents a novel evolutionary lineage of *Saccharomycotina* yeasts, unexplored before at the genomic level. It shares common signatures with genomes of the *Dekkera–Ogataea* group, two other nitrate-assimilating yeasts, although no trace of synteny conservation could be identified with them, suggesting a very ancient common ancestry. Conserved signatures include a striking reduction of tRNA gene redundancy, a novel tRNA-sparing strategy, and a cluster of nitrate assimilation genes, as observed in filamentous fungi but not in all other yeast genomes known so far. Consistent with phylogenetic reconstructions based on individual gene trees, the tRNA sparing strategy places the group of nitrate-assimilating yeasts, including *K. capsulata* at an intermediate position between *Ko. pastoris*, a member of the *Phaffomycetaceae* family, and the large group of CTG yeasts. The first one uses the eukaryotic sparing strategy for both the Arg CGN and Leu CUN codon families, like *Y. lipolytica*, while the second uses the bacterial sparing strategy for the Arg CGN codon family and read CUG as serine (with some degrees of ambiguity). For memory, almost all members of the *Saccharomycetaceae* family use the bacterial sparing strategy for both families of codons. *K. capsulata* and the other nitrate-assimilating yeasts use the bacterial sparing strategy for the Arg CGN codon family but keep the eukaryotic one for the Leu CUN codon family, supporting the idea of a monophyletic group of yeasts despite an important evolutionary divergence. The homogeneity of this group is also supported by the presence in their genomes of a nitrate gene cluster, composed of homologs to the *crnA*, *niaD*, and *niiA* genes of filamentous fungi encoding enzymes for nitrate import and nitrate and nitrite reduction. The presence of this cluster, specific to *K. capsulata*, *O. polymorpha*, and *D. bruxellensis*, but not found in other yeasts including *Ko. pastoris*, explains their ability to assimilate nitrate and suggests an ancestral acquisition of this cluster, possibly from a filamentous fungus. Additional copies of the *niaD* gene (but not the other two) are found dispersed in genomes of other yeasts and filamentous fungi. They are generally in duplicate, suggesting a very ancient duplication of this gene in fungi by a clearly distinct process from the acquisition of the cluster. The mechanism of this acquisition remains, of course, hypothetical but transfer of a specific chromosomal segment of dozens of kb from a fungus to a yeast is not without analogy to the recently documented introgressions of large chromosomal segments originating from distantly related species in several yeast genomes.

The number of protein-coding genes in *K. capulata* (6,029 predicted in total) is similar to what exists in many CTG yeasts and higher than for most *Saccharomycetaceae* yeasts, including those that emerged after the ancestral genome duplication. Part of this situation may be explained by a high global genome redundancy observed in *K. capsulata* (1.37×), due numerous dispersed paralogs and to some large gene families. The low gene duplication rate in this lineage, however, indicates that most of the redundancy observed in the *K. capsulata* genome is inherited from ancient duplications rather than from species-specific gene family expansions. Surprisingly, the genome of *K. capsulata* shows a relatively high frequency of protein-coding genes interrupted by introns (13.5%), a figure close to *Y. lipolytica* but twice as that of other *Saccharomycotina*. The similarity of *K. capsulata* with *Y. lipolytica* is not only illustrated by its high GC content (45.7%, a likely convergence) but also by the presence of dispersed copies of 5S RNA gene, the presence of a common Ile-tRNA$^{[AAU]}$–Gln-tRNA$^{[UUG]}$ transcriptional fusion (a rare case), and, even more strikingly, by the close proximity of their "protein conservation indexes" with other

yeasts and fungi, clearly differentiating them from the *Saccharomycetaceae* and the CTG yeasts (supplementary fig. S7, Supplementary Material online). More yeast sequences are clearly needed to clarify whether these similarities represent evolutionary convergences or conservation of ancestral features between the two phylogenetically distinct lineages. Again, the centromeres of *K. capsulata* correspond to GC-poor troughs of similar size as those originally identified in *Y. lipolytica* and also present in several CTG yeasts.

Supplementary Material

Supplementary figures S1–S8 and tables S1–9 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Antonia Doyen and Bart Theelen for technical advices and assistance, and Laurent Chatre for scientific discussions and expert microphotographies of yeasts. Sequencing was supported by CEA/Genoscope and performed using HPC resources from GENCI-TGCC (Grants 2012076389 and 2013036389). Data analysis and experimental aspects of the work were supported in part 1) by contract DYGEVO from ANR (2011SVSE6) to B.D., 2) by the Spanish ministry of Economy and Competitiveness (BIO2012-37161) and the European Research Council (Grant Agreement n° ERC-2012-StG-310325) to T.G., and 3) by the Qatar National Research Fund grant (NPRP 5-298-3-086) to T.B. and T.G.

Literature Cited

Akaike H. 1973. Presented at the 2nd International Symposium on Information Theory, Budapest, Hungary.
Aury J-M, et al. 2008. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. BMC Genomics 9:603.
Borneman AR, et al. 2011. Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of *Saccharomyces cerevisiae*. PLoS Genet. 7:e1001287.
Brettauer RK, Kozak LP, Irwin WE. 1969. Phosphate and mannose transfer from guanosine diphosphate mannose to yeast mannan acceptors. Biochem Biophys Res Commun. 37:820–827.
Breus NA, Ryazanova LP, Dmitriev VV, Kulakovskaya TV, Kulaev IS. 2012. Accumulation of phosphate and polyphosphate by *Cryptococcus humicola* and *Saccharomyces cerevisiae* in the absence of nitrogen. FEMS Yeast Res. 12:617–624.
Brito N, Avila J, Perez MD, Gonzalez C, Siverio JM. 1996. The genes *YNI1* and *YNR1*, encoding nitrite reductase and nitrate reductase respectively in the yeast *Hansenula polymorpha*, are clustered and co-ordinately regulated. Biochem J. 317:89–95.
Butler G, et al. 2009. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. Nature 459:657–662.
Capella-Gutiérrez S, Silla-Martinez JM, Gabaldon T. 2009. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.
Crawford NM, Aarst HN. 1993. The molecular genetics of nitrate assimilation in fungi and plants. Annu Rev Genet. 27:115–146.
Cruz J, Westhof E. 2011. Identification and annotation of noncoding RNAs in *Saccharomycotina*. C R Biol. 334:671–678.
Curtin CD, Borneman AR, Chambers PJ, Pretorius IS. 2012. De-novo assembly and analysis of the heterozygous triploid genome of the wine spoilage yeast *Dekkera bruxellensis* AWRI1499. PLoS One 7(3): e33840.
D'Souza CA, et al. 2011. Genome variation in *Cryptococcus gattii*, an emerging pathogen of immunocompetent host. MBio 11:e00342–10.
Dávila López M, Rosenblad MA, Samuelsson T. 2009. Conserved and variable domains of RNase MRP RNA. RNA Biol. 6:208–220.
De Schutter K, et al. 2009. Genome sequence of the recombinant protein production host *Pichia pastoris*. Nat Biotechnol. 27:561–566.
Dujon B. 2010. Yeast evolutionary genomics. Nat Rev Genet. 11:512–524.
Dujon B, et al. 2004. Genome evolution in yeasts. Nature 430:35–44.
Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.
Fairhead C, Dujon B. 2006. Structure of *Kluyveromyces lactis* subtelomeres: duplications and gene content. FEMS Yeast Res. 6:428–441.
Ferrari SR, Grubb J, Bishop DK. 2009. The Mei5-Sae3 protein complex mediates Dmc1 activity in *Saccharomyces cerevisiae*. J Biol Chem. 284:11766–11770.
Fitzpatrick DA, Logue ME, Butler G. 2008. Evidence of recent interkingdom horizontal gene transfer between bacteria and *Candida parapsilosis*. BMC Evol Biol. 8:181.
Fitzpatrick DA, Logue ME, Stajich JE, Butler G. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol Biol. 6:99.
Gabaldón T. 2008. Large-scale assignment of orthology: back to phylogenetics? Genome Biol. 9:235.
Galeote V, et al. 2011. Amplification of a *Zygosaccharomyces bailii* DNA segment in wine yeast genomes by extrachromosomal circular DNA formation. PLoS One 6:e17872.
Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 14:685–695.
Gillece JD, et al. 2011. Whole genome sequence analysis of *Cryptococcus gattii* from the Pacific Northwest reveals unexpected diversity. PLoS One 6:e28550.
Gordon JL, et al. 2011. Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents. Proc Natl Acad Sci U S A. 108: 20024–20029.
Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.
Howe KL, Chothia T, Durbin R. 2002. GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. Genome Res. 12:1418–1427.
Huerta-Cepas J, Gabaldón T. 2011. Assigning duplication events to relative temporal scales in genome-wide studies. Bioinformatics 27:38–45.
Huerta-Cepas J, et al. 2011. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Nucleic Acids Res. 39: D556–D560.
Iwasaki O, Noma K-I. 2012. Global genome organization mediated by RNA polymerase III-transcribed genes in fission yeast. Gene 493: 195–200.
Johnstone IL, et al. 1990. Isolation and characterisation of the crnA-niiA-niaD gene cluster for nitrate assimilation in *Aspergillus nidulans*. Gene 90:181–192.
Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38:D355–D360.
Katoh K, Kuma K-i, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33: 511–518.

Ketel C, et al. 2009. Neocentromeres form efficiently at multiple possible loci in *Candida albicans*. PLoS Genet. 5:e1000400.
Konishi M, Hatada Y, Horuishi J-I. 2013. Draft genome sequence of the basidiomycetous yeast-like fungus *Pseudozyma hubeiensis* SY62, which produces an abundant amount of the biosurfactant mannosylerythritol lipids. Genome Announc. 1:e00409–13.
Küberl A, et al. 2011. High-quality genome sequence of *Pichia pastoris* CBS7435. J Biotechnol. 154:312–320.
Kurtzman CP. 1984. Synonomy of the yeast genera *Hansenula* and *Pichia* demonstrated through comparisons of deoxyribonucleic acid relatedness. Antonie Van Leeuwenhoek 50:209–217.
Kurtzman CP. 2011. A new methanol assimilating yeast, *Ogataea parapolymorpha*, the ascosporic state of *Candida parapolymorpha*. Antonie van Leeuwenhoek 100:455–462.
Kurtzman CP, Fell JW, Boekhout T. 2011. The yeasts, a taxonomic study. 5th ed. Elsevier Science.
Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.
Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol. 25:1307–1320.
Leh-Louis V, et al. 2012. *Pichia sorbitophila*, an interspecies yeast hybrid, reveals early steps of genome resolution after polyploidization. G3 2: 299–311.
Li R, Li Y, Kristiansen K, Wang J. 2008. SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714.
Liti G, et al. 2009. Population genomics of domestic and wild yeasts. Nature 458:337–341.
Loftus BJ, et al. 2005. The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. Science 307: 1321–1324.
Louis E. 1995. The chromosome ends of *Saccharomyces cerevisiae*. Yeast 11:1553–1573.
Lynch DB, Logue ME, Butler G, Wolfe KH. 2010. Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. Genome Biol Evol. 2:572–583.
Man O, Pilpel Y. 2007. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nat Genet. 39:415–421.
Marcet-Houben M, Gabaldón T. 2009. The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome. PLoS One 4:e4357.
Marcet-Houben M, Gabaldón T. 2010. Acquisition of prokaryotic genes by fungal genomes. Trends Genet. 26:5–8.
Marck C, et al. 2006. The RNA polymerase III-dependent family of genes in hemiascomycetes: comparative RNomics, decoding strategies, transcription and evolutionary implications. Nucleic Acids Res. 34: 1816–1835.
Mattanovich D, et al. 2009. Genome, secretome and glucose transport highlight unique features of the protein production host *Pichia pastoris*. Microb Cell Fact. 8:29.
Meraldi P, McAinsh A, Rheinbay E, Sorger PK. 2006. Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins. Genome Biol. 7:R23.
Morales L, Dujon B. 2012. Evolutionary role of interspecies hybridization and genetic exchanges in yeasts. Microbiol Mol Biol Rev. 76: 721–739.
Morita T, et al. 2013. Genome sequence of the basidiomycetous yeast *Pseudozyma antarctica* T-34, a producer of the glycolipid biosurfactants mannosylerythritol lipids. Genome Announc. 1(2): e0006413.
Nishida H, et al. 2011. Draft genome sequencing of the enigmatic yeast *Saitoella complicate*. J Gen Appl Microbiol. 57:243–246.
Patel RN, et al. 2004. Enatioselective microbial reduction of substituted acetophenones. Tetrahedron Asymm. 15:1247–1258.
Petersen GR, Schubert WW, Richards GF, Nelson GA. 1990. Yeasts producing exopolysaccharides with drag-reducing activity. Enzyme Microb Technol. 12:255–259.
Piccinelli P, Rosenblad MA, Samuelsson T. 2005. Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes. Nucleic Acids Res. 33:4485–4495.
Piškur J, et al. 2012. The genome of wine yeast *Dekkera bruxellensis* provides a tool to explore its food-related properties. Int J Food Microbiol. 157:202–209.
Ramezani-Rad M, et al. 2003. The *Hansenula polymorpha* (strain CBS4732) genome sequencing and analysis. FEMS Yeast Res. 4: 207–215.
Retief JD. 2000. Phylogenetic analysis using PHYLIP. Methods Mol Biol. 132:243–258.
Rhind N, et al. 2011. Comparative functional genomics of the fission yeasts. Science 332:930–936.
Richard G-F, Kerrest A, Lafontaine I, Dujon B. 2005. Comparative genomics of hemiascomycete yeasts: genes involved in DNA replication, repair, and recombination. Mol Biol Evol. 22:1011–1023.
Rolland T, Neuveglise C, Sacerdot C, Dujon B. 2009. Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes. PLoS One 4(8):e6515.
Rosenblad MA, Larsen N, Samuelsson T, Zwieb C. 2009. Kinship in the SRP RNA family. RNA Biol. 6:508–516.
Sanyal K, Baum M, Carbon J. 2004. Centromeric DNA sequences in the pathogenic yeast *Candida albicans* are all different and unique. Proc Natl Acad Sci U S A. 101:11374–11379.
Shimotoyodome A, et al. 2006. Reduction of saliva-promoted adhesion of *Streptococcus mutans* MT8147 and dental biofilm development by tragacanth gum and yeast-derived phosphomannan. Biofouling 22: 261–268.
Siverio JM. 2002. Assimilation of nitrate by yeasts. FEMS Microbiol Rev. 26: 277–284.
Souciet J-L, et al. 2009. Comparative genomics of protoploid *Saccharomycetaceae*. Genome Res. 19:1696–1709.
Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456–463.
Subramanian AR, Kaufmann M, Morgenstern B. 2008. DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms Mol Biol. 3:6.
Suh S-O, Zhou J. 2010. Methylotrophic yeasts near *Ogataea (Hansenula) polymorpha*: a proposal of *Ogataea angusta* comb. nov. and *Candida parapolymorpha* sp. nov. FEMS Yeast Res. 10:631–638.
Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278:631–637.
Tekaia F, Yeramian E. 2012. SuperPartitions: detection and classification of orthologs. Gene 492:199–211.
Wallace IM, O'Sullivan O, Higgins DG, Notredame C. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. 34:1692–1699.
Wehe A, Bansal MS, Burleigh JG, Eulenstein O. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics 24:1540–1541.
Wickerham LJ. 1952. Recent advances in the taxonomy of yeasts. Annu Rev Microbiol. 6:317–332.
Wickerham LJ. 1970. *Hansenula* H. et P. Sydow. In: Lodder J, editor. The yeasts, a taxonomic study, 2nd ed.. Amsterdam: North-Holland. p. 226–315.
Wood V, et al. 2002. The genome sequence of *Schizosaccharomyces pombe*. Nature 415:871–880.
Xu J, et al. 2007. Dandruff-assocated *Malassezia* genomes reveal convergent and divergent virulence traits shared with pant and human fungal pathogens. Proc Natl Acad Sci U S A. 104:18730–18735.

Yamada Y, Maeda K, Mikata K. 1994. The phylogenetic relationships of the hat-shaped ascospore-forming, nitrate-assimilating *Pichia* species, formerly classified in the genus *Hansenula* Sydow et Sydow, based on the partial sequences of 18S and 26S ribosomal RNAs (*Saccharomycetaceae*): the proposals of three new genera, *Ogataea*, *Kuraishia*, and *Nakazawaea*. Biosci Biotechnol Biochem. 58: 1245–1257.

Yamane T, Ogawa T, Matsuoka M. 2008. Derivation of consensus sequence for protein binding site in *Yarrowia lipolytica* centromere. J Biosci Bioeng. 105:671–674.

Yanai T, Sato M. 2000. Purification and characterization of a novel alpha-L-arabinofuranosidase from *Pichia capsulata* X91. Biosci Biotechnol Biochem. 64:1181–1188.

Zimmer C, Fabre E. 2011. Principles of chromosomal organization: lessons from yeast. J Cell Biol. 192:723–733.

Zwieb C, van Nues RW, Rosenblad MA, Brown JD, Samuelsson T. 2005. A nomenclature for all signal recognition particle RNAs. RNA 11: 7–13.

Associate editor: Emmanuelle Lerat