Genome evolution in yeasts

Bernard Dujon[1], David Sherman[5,6], Gilles Fischer[1], Pascal Durrens[6,7], Serge Casaregola[8], Ingrid Lafontaine[1], Jacky de Montigny[9], Christian Marck[10], Cécile Neuvéglise[8], Emmanuel Talla[1], Nicolas Goffard[6], Lionel Frangeul[2], Michel Aigle[7], Véronique Anthouard[11], Anna Babour[8], Valérie Barbe[11], Stéphanie Barnay[8], Sylvie Blanchin[8], Jean-Marie Beckerich[8], Emmanuelle Beyne[5,6], Claudine Bleykasten[9], Anita Boisramé[8], Jeanne Boyer[1], Laurence Cattolico[11], Fabrice Confanioleri[12], Antoine de Daruvar[6], Laurence Despons[9], Emmanuelle Fabre[1], Cécile Fairhead[1], Hélène Ferry-Dumazet[6], Alexis Groppi[6], Florence Hantraye[3], Christophe Hennequin[1], Nicolas Jauniaux[9], Philippe Joyet[8], Rym Kachouri[13], Alix Kerrest[1], Romain Koszul[1], Marc Lemaire[14], Isabelle Lesur[5], Laurence Ma[2], Héloïse Muller[1], Jean-Marc Nicaud[8], Macha Nikolski[5], Sophie Oztas[11], Odile Ozier-Kalogeropoulos[1], Stefan Pellenz[1], Serge Potier[9], Guy-Franck Richard[1], Marie-Laure Straub[9], Audrey Suleau[8], Dominique Swennen[8], Fredj Tekaia[1], Micheline Wésolowski-Louvel[14], Eric Westhof[13], Bénédicte Wirth[9], Maria Zeniou-Meyer[9], Ivan Zivanovic[12], Monique Bolotin-Fukuhara[12], Agnès Thierry[1], Christiane Bouchier[2], Bernard Caudron[4], Claude Scarpelli[11], Claude Gaillardin[8], Jean Weissenbach[11], Patrick Wincker[11] & Jean-Luc Souciet[9]

[1]*Unité de Génétique Moléculaire des Levures (URA 2171 CNRS and UFR 927 Université Pierre et Marie Curie),* [2]*Plate-forme génomique, Pasteur Génopole Ile-de-France,* [3]*Unité de Génétique des interactions macromoléculaires (URA 2171 CNRS), and* [4]*Groupe Logiciels et Banques de données, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France*
[5]*Laboratoire Bordelais de Recherche en Informatique (LaBRI, UMR 5800 CNRS), 351 cours de la Libération, 33405 Talence Cedex, France*
[6]*Centre de Bioinformatique de Bordeaux, Université Victor Ségalen (Bordeaux 2), 146 rue Léo Saignat, 33076 Bordeaux Cedex, France*
[7]*Institut de Biochimie et Génétique Cellulaires (UMR 5095 CNRS), Université Victor Ségalen (Bordeaux 2), 1 rue Camille Saint-Säens, 33077 Bordeaux Cedex, France*
[8]*Collection de Levures d'Intérêt Biotechnologique et Laboratoire de Génétique Moléculaire et Cellulaire (UMR 216 INRA and URA 1925 CNRS), INA-PG, PO Box 01, 78850 Thiverval-Grignon, France*
[9]*Laboratoire de Dynamique, Evolution et Expression des Génomes de Microorganismes (FRE 2326 CNRS), Université Louis Pasteur, 28 rue Goethe, 67000 Strasbourg, France*
[10]*Service de Biochimie et de Génétique Moléculaire, CEA/Saclay, 91191 Gif-sur Yvette, France*
[11]*Génoscope (UMR 8030 CNRS), 2 rue Gaston Crémieux, 91057 Evry Cedex, France*
[12]*Institut de Génétique Moléculaire (UMR 8621 CNRS), Université de Paris Sud, Bâtiment 400, 91405 Orsay Cedex, France*
[13]*Modélisations et Simulations des Acides Nucléiques, IBMC (UPR 9002 CNRS), 15 rue René Descartes, 67000 Strasbourg, France*
[14]*Laboratoire de Génétique des Levures (UMR 5122 CNRS), Université Claude Bernard, Bâtiment Lwoff, 43 Boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France*

Identifying the mechanisms of eukaryotic genome evolution by comparative genomics is often complicated by the multiplicity of events that have taken place throughout the history of individual lineages, leaving only distorted and superimposed traces in the genome of each living organism. The hemiascomycete yeasts, with their compact genomes, similar lifestyle and distinct sexual and physiological properties, provide a unique opportunity to explore such mechanisms. We present here the complete, assembled genome sequences of four yeast species, selected to represent a broad evolutionary range within a single eukaryotic phylum, that after analysis proved to be molecularly as diverse as the entire phylum of chordates. A total of approximately 24,200 novel genes were identified, the translation products of which were classified together with *Saccharomyces cerevisiae* proteins into about 4,700 families, forming the basis for interspecific comparisons. Analysis of chromosome maps and genome redundancies reveal that the different yeast lineages have evolved through a marked interplay between several distinct molecular mechanisms, including tandem gene repeat formation, segmental duplication, a massive genome duplication and extensive gene loss.

Comparative genomics in eukaryotes has been limited by the considerable phylogenetic distances between the first sequenced organisms: a yeast (*S. cerevisiae*)[1], a nematode (*Caenorhabditis elegans*)[2], an insect (*Drosophila melanogaster*)[3], a dicotyledonous plant (*Arabidopsis thaliana*)[4] and man (*Homo sapiens*)[5]. New sequencing programmes have recently extended these comparisons to organisms of the same phylogenetic group, but the number of completely sequenced genomes remains low, and intriguing differences appear between the groups. For example, two sequenced Diptera (*Anopheles gambiae*[6] and *D. melanogaster*) believed to have diverged from their common ancestor about 250 million years (Myr) ago show a larger sequence divergence than two vertebrate species (*H. sapiens* and the fish *Takifugu rubripes*[7]), which diverged 450 Myr ago. The recently published draft sequence of the *Caenorhabditis briggsae* genome reveals extensive map collinearity with *C. elegans*, from which it diverged 100 Myr[8] ago. By contrast, two cultivars of rice, *Oryza sativa*[9,10], show a very limited synteny with *A. thaliana*, from which they diverged 200 Myr ago.

Yeasts and fungi are ideal organisms for comparative genomic studies in eukaryotes because of their small and compact genomes and because they include a number of species, such as *Neurospora crassa*[11], *S. cerevisiae* and *Schizosaccharomyces pombe*[12], that have been, and continue to be, used extensively in genetic studies. However, the divergence between these three species is ancient (estimated to be at least 300 Myr old) and the organization of their genomes is quite different. The diversity of the hemiascomycetes—a group of ascomycetes that contains most of the known yeast species—was first explored four years ago using low-coverage sequencing of 13 distinct species[13]. More recently, deeper sequencing coverages were applied to a few *Saccharomyces* species very closely related to *S. cerevisiae* (the *sensu stricto* group), plus one more distant species, *Saccharomyces kluyveri*, in order to identify conserved regulatory elements[14,15]. While this article was submitted, the complete genome sequences of *Ashbya gossypii*[16], a filamentous yeast, and *Kluyveromyces waltii*[17] were used to map and analyse the ancient genome duplication in the ancestry of *S. cerevisiae*.

Instead of focusing on closely related species or on the origin of *S. cerevisiae*, we decided to explore the evolution of the hemiascomycete phylum as broadly as possible. On the basis of our previous estimates[13] we selected four yeast species representing various and distant branches among the hemiascomycetes for complete sequencing. *Candida glabrata* was chosen because it has become the second causative agent of human candidiasis, and because, despite its name, it is phylogenetically more closely related to *S. cerevisiae*

than to *C. albicans*, the major human fungal pathogen with which it shares only a few properties. *Kluyveromyces lactis* is a yeast species commonly used for genetic studies, and it occupies an interesting position within the phylogeny of hemiascomycetes. *Debaryomyces hansenii* was selected because it is a halotolerant yeast, related to *C. albicans* and other pathogenic yeasts, that is often found on fish and salted dairy products. *Yarrowia lipolytica*, an alkane-using yeast commonly used in genetic studies, is very distantly related to the rest of the yeasts; instead it shares a number of common properties with filamentous fungi. For each species, the haploid type strain was sequenced. Of importance for evolutionary studies, the four yeast species display different mechanisms of sexuality (see ref. 13). *Yarrowia lipolytica* has a haplo-diplontic cycle (that is, it alternates between haploid and diploid phases of similar importance), whereas *D. hansenii* is a homothallic yeast with an essentially haplontic life cycle. Both species have only one mating-type locus (*MAT*), whereas the other two have two silent mating-type cassette homologues, similar to *S. cerevisiae*. As is often the case with pathogens, *C. glabrata* displays no known sexual cycle, despite the fact that haploid strains of the two distinct mating types are regularly isolated from patients. Finally, *K. lactis* is a heterothallic species with a predominantly haplontic cycle, in contrast to *S. cerevisiae* in which the predominantly diplontic cycle is pseudo-heterothallic owing to mating-type switching.

This work, which represents the first multispecies exploration of genome evolution across an entire eukaryotic phylum, reveals the variety of events and mechanisms that have taken place, and should allow useful comparisons with other phyla of multicellular organisms when more genome sequences are determined.

Overview of the four yeast genomes

Sequencing status of the four yeast species is summarized in Table 1. Genome sizes and chromosome numbers vary within an approximately twofold range between the four species, but without direct correlation. In contrast, the total number of protein-coding genes varies only by 1.3 times between the four species (Table 2), whereas the total number of transfer RNA genes varies by more than three times between *Y. lipolytica* (510 genes) and *K. lactis* (162 genes). The overall gene density is significantly lower in *Y. lipolytica* (one gene per 3 kilobases (kb)) than in the other yeasts (one gene per 2 kb, as in *S. cerevisiae*). We have identified all of the centromeres from *C. glabrata*, *K. lactis* and *Y. lipolytica* (see Supplementary Table S1), but were unable to detect the centromeres of *D. hansenii*. We also identified telomeric repeats and proteins of the telomerase complex (see Supplementary Table S2). Transposable elements will be reported elsewhere (S.C., C.N. and P.W., unpublished data).

Non-coding RNA genes

Ribosomal RNA genes

There is a large diversity in the organization of rDNA repeats among yeasts. Compared with the single intrachromosomal rDNA repeat locus of *S. cerevisiae* and *K. lactis*, three distinct intrachromosomal loci are found in *D. hansenii*, whereas seven and two loci are found in subtelomeric regions of *Y. lipolytica* and *C. glabrata*, respectively. Variability also prevails for the 5S rRNA gene copies. In contrast with *S. cerevisiae* and *K. lactis*, where a single copy of this gene is located in opposing orientation between each repeat unit of the 35S primary transcript (the precursor of the 18S, 5.8S and 26S rRNA molecules), two copies occur in tandem in each repeat in *C. glabrata* and *D. hansenii*, and the same gene is dispersed in 105 copies (plus 11 pseudogenes) throughout the genome of *Y. lipolytica*.

Transfer RNA genes

The type and number of tRNA genes (tDNA) are described in Supplementary Table S3. *Candida glabrata* and *K. lactis* display exactly the same 42-tDNA set as *S. cerevisiae*, whereas *D. hansenii* uses a slightly different 43-tDNA set, and *Y. lipolytica* uses the same 44-tDNA set as higher eukaryotes[18]. The CUG codon (leucine) is used as a serine codon in *D. hansenii*[13], and is read by the special single-copy tRNA-Ser (CAG), as in *C. albicans*. Note that this modification of the genetic code does not exist in *C. glabrata*, another indication of the artificial nature of this genus. Introns are found in about one-quarter of the tDNAs in *C. glabrata*, *K. lactis* and *D. hansenii* (see Supplementary Table S4), as in *S. cerevisiae*; however, in *Y. lipolytica* 26 of the 44 tDNAs contain introns, a proportion that has not been observed so far in other eukaryotes[18]. As in *S. cerevisiae*, tRNA genes are scattered throughout the genomes of the four yeast species. We found no gene clusters, except in *D. hansenii* where eight identical copies of a tDNA-Lys (CTT) are repeated in tandem separated by intergenic distances sufficient for independent transcription (188–1,855 bp). Notably, a number of tDNA pairs are observed in which the distance separating the two

Table 1 **Genome assemblies of the four yeast species**

Species	Strain	Number of chromosomes	Total reads	Coverage (sequence)	Coverage (clones)	N50 contigs (kb)	N50 scaffolds (kb)	Total gaps	Assembly size (without rDNA) (kb)
C. glabrata	CBS138	13	188,853	× 8	× 30	1,000	1,025	6	12,280
K. lactis	CLIB210	6	152,071	× 11.4	× 56	1,670	1,670	0	10,631
D. hansenii	CBS767	7	150,570	× 9.7	× 36	102	2,038	207	12,221
Y. lipolytica	CLIB99	6	247,279	× 10	× 59	704	3,453	11	20,503

Sequencing and assembly were performed as described in Methods. Sequences of *C. glabrata*, *K. lactis* and *Y. lipolytica* are finished (no gap) or contain very few gaps. The sequence of *D. hansenii* is in the form of a high-quality draft. In all cases, each chromosome of each yeast is either complete (single contig) or represented by a single super-contig (scaffold). Most remaining gaps are small or contain repeated sequences. Some subtelomeric regions are missing from the assembly because they are too similar to one another to be assigned to a specific chromosome. rDNA repeats are assembled separately. N50, median values.

Table 2 **General characteristics of the yeast genomes and predicted proteomes**

Species	Genome size (Mb)	Average G+C content (%)	Total CDS	Total tRNA genes	Average gene density (%)	Average G+C in CDS (%)	Average CDS size (codons)	Median CDS size (codons)	Maximum CDS size (codons)
S. cerevisiae	12.1	38.3	5,807	274	70.3	39.6	485	398	4,911
C. glabrata	12.3	38.8	5,283	207	65.0	41.0	493	409	4,881
K. lactis	10.6	38.7	5,329	162	71.6	40.1	461	381	4,916
D. hansenii	12.2	36.3	6,906	205	79.2	37.5	389	307	4,190
Y. lipolytica	20.5	49.0	6,703	510	46.3	52.9	476	399	6,539

Figures are calculated from final chromosome sequences or scaffolds, after annotation. Genome sizes do not include rDNA. Average gene density represents the fraction of each genome occupied by the protein-coding genes (other genetic elements are not considered). Figures for *D. hansenii* are only tentative; figures for *S. cerevisiae* were recently recomputed from http://mips.gsf.de/genre/proj/yeast.

Table 3 **Classification of yeast proteins in families**

Family class	Number of families	Number of CDS					
		SACE	CAGL	KLLA	DEHA	YALI	Total
Robust (identical + reconciled)	3,410	4,094	3,651	3,504	3,832	3,296	18,377
Consensus	1,311	1,287	1,201	1,176	1,831	1,894	7,389
Non-assigned	–	426	431	649	1,243	1,513	4,262
Total	4,721	5,807	5,283	5,329	6,906	6,703	30,028

The table shows the total number of protein families in each class and the corresponding numbers of CDS in each yeast species. Families were classified as explained in Methods. See Fig. 1 for species abbreviations.

genes is shorter than the minimal 5′ sequence required for transcription (see Supplementary Table S5). Consistent with the idea of co-transcription, the two members of each pair are always co-oriented. Such structures must result from independent formation followed by subsequent duplications in each phylogenetic branch, as judged from the fact that they are distinct for each species and are often present in multiple copies dispersed throughout the genome.

Other non-coding RNA genes

Most RNA-polymerase-III-transcribed genes and other non-coding RNA genes are not well conserved in yeasts. Nevertheless, we have identified the U1–U6 small nuclear RNAs, as well as the RNA components of the RNase P, the signal recognition particle (SRP) and, in two species, the telomerase complex (see Supplementary Table S6). Some RNAs show extensive size variation between species. Although most RNA genes are unique in each yeast, the U3 gene is duplicated in *D. hansenii* and is triplicated in *Y. lipolytica*; the U1 and SRP genes are duplicated in *Y. lipolytica*; and the U4 gene is duplicated in tandem in *D. hansenii.*

Protein families and genome redundancy

Classification of yeast proteins and sequence conservation

Together with previous data from *S. cerevisiae*[1], our four new yeast sequences offer a unique collection of 30,028 proteins from five phylogenetically related species. As a first step towards identifying homologies between these genomes, the proteins were classified into families of homologues (see Methods and Table 3). When sorted according to their phyletic patterns (presence or absence in each species, Table 4), over 40% of the families (2,014 families and 17,153 genes) are common to the five yeasts (pattern 'sckdy', where 's' indicates *S. cerevisiae*, 'c' indicates *C. glabrata*, and so on). Most of these 'universal' families (1,208) contain a single protein from each species (1:1:1:1:1 relationship), considered here as 'direct orthologues'. The remaining 806 universal families contain at least one paralogous pair in at least one species, creating various situations of paralogy subtypes as defined in ref. 19. In agreement with the phylogeny of the five yeast species (see Supplementary Fig. S7), the next most abundant patterns are sck– and —dy, followed by sckd-. All families common to our five yeasts (pattern sckdy) have homologues in other hemiascomycete species, 98% of them have homologues in other ascomycetes, and 92% have homologues in basidiomycetes (see Table 4). Thus, most of these proteins seem to be universally or largely conserved in evolution. Evolutionary conservation is not as widespread for non-sckdy protein families, but remains consistent with the phylogeny (for example, compare protein families in the sck– pattern, which are poorly conserved beyond yeasts, with those in the —dy pattern, which are more often conserved beyond yeasts). In total, approximately 800 homologous protein families seem to be ascomycete-specific, of which about 660 are specific to hemiascomycetes only (not shown). The functions of these families seem to be highly diversified.

Overall genome redundancy as deduced from protein families

The global degree of genome redundancy, estimated from the number of paralogous gene copies per protein family, illustrates the importance of ancestral gene duplications in all yeasts (Fig. 1). Quantitative variations are, however, observed between species. *Kluyveromyces lactis* has the least duplicated genome (501 sets of

Table 4 **Phyletic patterns of yeast protein families and conservation in other fungi**

Pattern	Families	Proteins	Conservation (%)			
			Hemiascomycetes	Ascomycetes	Basidiomycetes	All
Families universal to all						
sckdy	2,014	17,153	100	98	92	100
Families restricted to *S. cerevisiae*, *C. glabrata* and/or *K. lactis*						
sck--	572	1,827	100	36	19	100
sc---	245	547	100	19	6	100
s-k--	102	212	100	25	11	99
-ck--	34	77	97	31	19	94
Families restricted to *D. hansenii* and *Y. lipolytica*						
---dy	488	1,121	62	77	52	89
Families missing in one species						
-ckdy	17	100	100	94	82	94
s-kdy	49	312	100	98	90	98
sc-dy	44	215	100	93	80	98
sck-y	81	382	100	92	76	99
sckd-	316	1,445	100	79	57	100
Species-specific families						
s----	31	95	90	13	0	87
-c---	8	51	63	50	50	50
--k--	16	38	50	13	6	44
---d-	135	501	42	40	29	55
----y	177	492	31	44	32	45
All other combinations						
sc-d-	16	53	100	68	55	95
sc--y	7	24	100	88	75	88
s-kd-	32	125	100	79	62	97
s--dy	14	46	100	93	87	93
s-k-y	8	27	100	71	57	86
-c-dy	19	63	100	75	50	95
-ckd-	13	43	100	50	33	92
-ck-y	3	9	100	100	100	67
--kdy	103	380	95	86	59	96
s--d-	51	110	98	59	44	96
s---y	7	15	83	83	50	83
-c-d-	18	53	100	77	64	95
-c--y	17	45	82	71	47	82
--kd-	58	143	85	51	33	92
--k-y	26	62	92	88	65	92
Total	4,721	25,766				

The phyletic pattern of each protein family indicates the presence (s, c, k, d, y) or absence (-) of its members in each yeast species (see text for definition of single-letter abbreviations). The total number of families and proteins are indicated. Note that the number of species-specific families may be overestimated because, in the present stage of annotation, some predicted ORFs might not correspond to actual genes. For each protein family a position-specific scoring matrix was computed for two rounds using BLASTpgp, with the longest member of the family serving as representative. Each position-specific scoring matrix and protein family representative was used to search a combined data bank of completely or partially sequenced fungal species, using psitBLASTn, with an expected cutoff value of 1×10^{-10} and no filtering of low-complexity regions. A total of 239,464 matches were recorded. The sequenced organisms included in the comparisons are: (1) hemiascomycetes: *Saccharomyces bayanus*, *S. castellii*, *S. kluyveri*, *S. kudriavzevii*, *S. mikatae*, *S. paradoxus* and *C. albicans*; (2) other ascomycetes: *Aspergillus fumigatus*, *A. nidulans*, *Coccidioides immitis*, *Fusarium graminearum*, *Magnaporthe grisea*, *N. crassa*, *Pneumocystis carinii*, *P. carinii carinii* and *S. pombe*; and (3) basidiomycetes: *Phanerochaete chrysosporium* and *Ustilago maydis*. The table shows the percentage of families of each pattern conserved in at least one other species of hemiascomycetes, other ascomycetes, basidiomycetes, or all simultaneously (All).

paralogues), whereas *D. hansenii* has the most duplicated one (901 sets). *Saccharomyces cerevisiae* and *Y. lipolytica* have an intermediate and similar level of duplication, and *C. glabrata* stands between *S. cerevisiae* and *K. lactis*. The maximum range of variation in global genome redundancy is only 1.6-fold (31.8% for *K. lactis* versus 51.5% for *D. hansenii*), a figure that, compared to duplication events discussed below, illustrates the importance of ancestral duplications that occurred before divergence of hemiascomycetes and the extent of gene loss in each branch.

Divergence between orthologous and paralogous proteins

The distributions of amino acid sequence divergence between orthologous proteins were calculated from all pairwise comparisons between the five yeast species. All distributions are unimodal, and their mean and median values are consistent with phylogeny (Fig. 2a–c). Notably, despite their similar morphologies and lifestyles, the five yeast species appear to be more diverse at the molecular level than, for example, the entire phylum of chordates. The average sequence identity between orthologous proteins of mammals (man or mouse) and fishes (*Takifugu* or *Tetraodon)* is about 70% (refs 7, 20, and O. Jaillon and H. Roest-Crollius, personal communication) compared with only 65% between *S. cerevisiae* and *C. glabrata*, or 60–61% between *K. lactis* and either *S. cerevisiae* or *C. glabrata* (Fig. 2a). Similarly, the average level of amino acid identity between *Y. lipolytica* and any of the other yeasts is 48–49%, a figure similar to that found between the urochordate *Ciona intestinalis* and any of the vertebrates (ref. 21, and O. Jaillon and H. Roest-Crollius, personal communication). Such a broad evolutionary range within the hemisacomycete yeasts was anticipated from our first low-coverage sequence exploration[13], and justifies *a posteriori* the present selection of species. In contrast to orthologous proteins, sequence divergence between paralogues in each species shows bimodal distributions with an abundance of low sequence identities (25–50%)—probably representing ancient duplications—and a moderate excess of highly similar proteins (90–100%) relative to the intermediate range (Fig. 2d). This last fraction, which is nearly absent in *C. glabrata*, must reflect recent duplications and/or sequence homogenization by gene conversion. Interestingly, the major peaks have similar modes for all five species (roughly 35% identity), indicating that these paralogues largely correspond to duplications having occurred before species divergence. By contrast, the paralogous pairs showing over ~60% identity must largely reflect duplication events that occurred after the speciation of these five yeasts, as this fraction quantitatively varies between species.

Expansion and contraction within universal protein families

Specific gene duplications in some phylogenetic branches and/or gene losses in others are represented by families containing unequal numbers of proteins between species. Such a situation is found in over 700 universal protein families (sckdy). In some cases, adaptative evolution, as reported in *S. cerevisiae*[22], or functionally significant gene dosage effects may exist. But as most of these variations concern small numbers of genes (1–3), they cannot be statistically distinguished from random accidents with no specific phenotypic effect. By contrast, 14 protein families were identified showing statistically significant (with a probability of 10^{-3}) expansion in *Y. lipolytica*, *D. hansenii* or both. Judging from their *S. cerevisiae* members, these families encode acylglycerol lipases, proteins similar to sphingomyelinases, α-1,4-glucan glucosidases, alkaline extracellular proteases, GPI-anchored aspartyl proteases, choline or allantoate transporters, peroxisomal 2,4-dienoyl-CoA reductase, C-22 sterol desaturase and other cytochrome P450 enzymes, or NADPH dehydrogenase. One family remains of unknown function. Multigenic families encoding multidrug resistance proteins and hexose transporters are specifically more expanded in *D. hansenii* than in the other four yeasts. In nearly all cases, the family expansions correspond to additional gene copies dispersed in each genome, suggesting ancient and multiple gene duplication events. Tandem gene repeat formation has a negligible role in family size expansions, except in *D. hansenii* (see below).

Genetic map organization and genome redundancy

Analysis of the genetic maps of the four yeasts and comparisons between them and *S. cerevisiae* reveals an interplay of several distinct mechanisms that have contributed unequally to the evolution of these genomes in the different lineages.

Tandem gene duplications

As in *S. cerevisiae*, a few dozen tandem gene arrays, mostly composed of two or three gene copies, are dispersed throughout the genomes of *C. glabrata*, *K. lactis* and *Y. lipolytica*. However, the genome of *D. hansenii* was found to be much richer in such structures than the other yeast species (Fig. 3). A total of 247 arrays (329 gene pairs) are distributed all over this genome, significantly contributing to its global redundancy. Furthermore, arrays of 8 to 9 repeats are encountered in this species (Fig. 3). The quantitative difference between *D. hansenii* and the other yeasts was unexpected. It may result from a more efficient creation of tandem arrays (or a recent massive episode) or a less efficient destruction of newly formed tandem repeats (for example by 'pop-out'). Although their sequence divergence is generally lower than that of dispersed paralogues, tandem paralogues are generally not strictly identical in sequence, indicating possible functional specialization (and limiting their destruction by pop-out). Frameshifted pseudogenes and gene fragments are often found in large arrays. Finally, in 27 cases the repeats are made of two genes forming alternating arrays. The functions of genes included in tandem arrays appear to be extremely diverse (see Supplementary Table S8). As only two arrays are common to all yeasts, it is probable that new arrays are constantly formed (and possibly rapidly resolved) in all lineages.

Blocks of ancestral duplications

A significant part of the genome redundancy in *S. cerevisiae* is due to

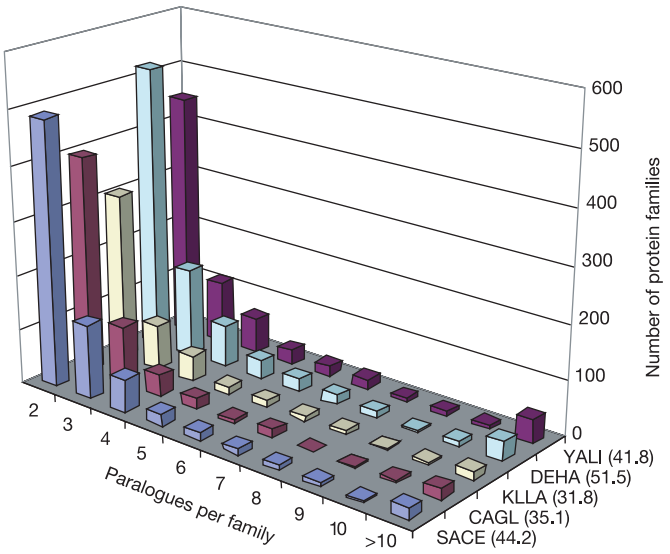


Figure 1 Overall genome redundancy as deduced from protein families. Shown for each yeast species is the total number of protein families distributed according to their size. A similar pattern (not shown) is obtained when considering only the universal protein families (sckdy pattern). The overall genome redundancy for each species, defined as the ratio (in per cent) of the number of CDS belonging to multigene families over the total number of CDS, is indicated in brackets next to the species abbreviation. SACE, *S. cerevisiae*; CAGL, *C. glabrata*; KLLA, *K. lactis*; DEHA, *D. hansenii*; YALI, *Y. lipolytica*.

the presence of sister chromosomal regions or blocks[23] thought to be the remnants of an ancestral whole-genome duplication that occurred after its divergence from *K. lactis* and before its divergence from *C. glabrata*[24,25]. According to this theory, similar numbers of blocks are expected in *C. glabrata* and *S. cerevisiae*, and no block is expected in other yeasts. Rather than this simple all-to-none transition, our results instead show gradual quantitative differences between the species (Fig. 4). Compared with the 56 blocks scattered throughout the genome map in *S. cerevisiae* (plus 21 blocks in subtelomeric regions), only 20 blocks are found in *C. glabrata* (approximately three times less), 8 blocks (1 tandem) in *K. lactis*, 5 blocks (3 in tandem) in *D. hansenii* and 2 blocks in *Y. lipolytica*. A closer examination of the results reveals that 18 of the 20 blocks identified in *C. glabrata* correspond to blocks also present in *S. cerevisiae* (see Supplementary Table S9). This high coincidence strongly suggests that the duplicated blocks of *C. glabrata* have the same ancestral origin as those of *S. cerevisiae*, indicating that a major duplication occurred before the divergence of the two species and after their divergence from *K. lactis*. Consistently, none of the eight duplicated blocks in *K. lactis* coincide with the above, indicating a distinct formation; the same applies for the few duplicated blocks of *D. hansenii*.

The marked difference between *C. glabrata* and *S. cerevisiae* needs, however, to be explained in order to be reconciled with a whole-genome duplication event in their common ancestry (this whole-genome duplication is also documented by recent comparisons to *A. gossypii*[16] and *K. waltii*[17]). Because the *C. glabrata* blocks tend to be smaller, on average, than those of *S. cerevisiae* (Fig. 4), a plausible explanation for their reduced number is that, after the duplication, a higher rate of gene loss occurred in *C. glabrata* compared with *S. cerevisiae*, thus effacing many blocks inherited from their common ancestor and reducing the size of others. This possibility is consistent with the reductive genome evolution mentioned below for this pathogenic yeast, and with its reduced conservation of synteny with other yeasts (not shown). It is also consistent with the lower global genome redundancy of *C. glabrata* compared with *S. cerevisiae* (see Fig. 1) and with the reduced number of highly similar paralogues (see Fig. 2).

The duplicated blocks observed in *K. lactis*, *D. hansenii* and *Y. lipolytica* are too few to indicate other possible massive genome duplications in their own lineages. Instead, like some of the blocks that do not coincide between *C. glabrata* and *S. cerevisiae*, they may have originated from independent segmental duplications, a mechanism now directly demonstrated experimentally in *S. cerevisiae*[26]. In general, the absence of a direct correlation between the total number of duplicated blocks in the genome maps and the global levels of genome redundancy judged from individual genes (see above), illustrates the multiplicity of events that have occurred in the various lineages. If an ancestral genome duplication event accounts for a large part of the redundancy of the

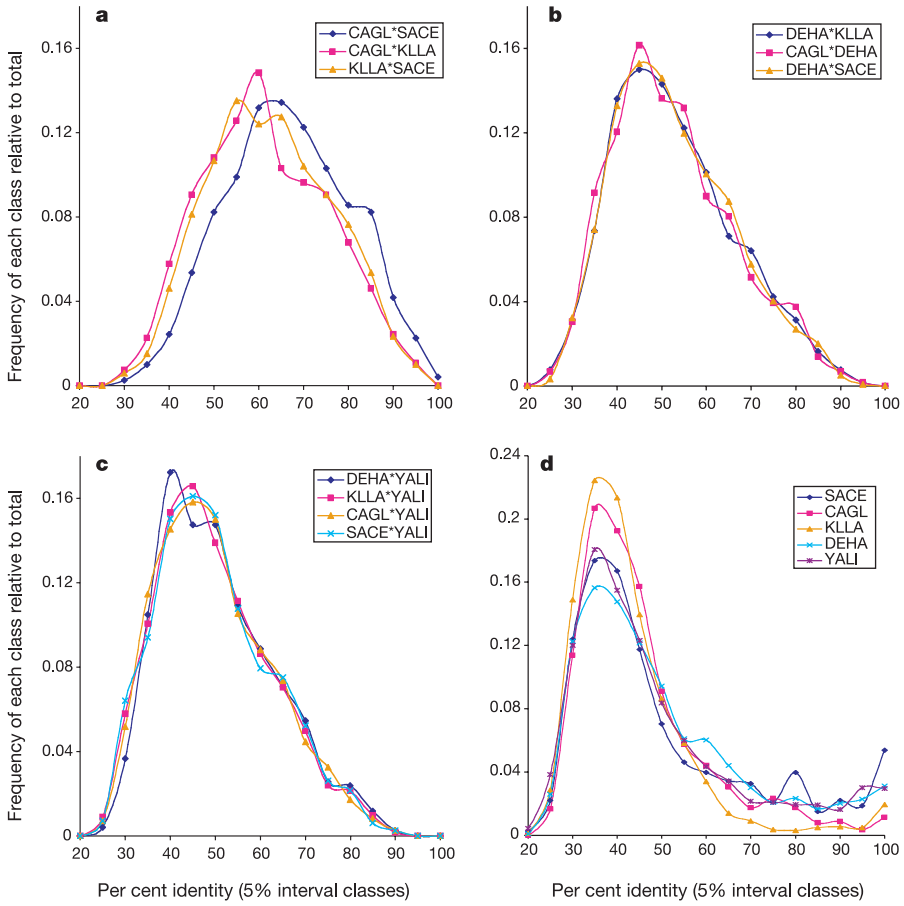


Figure 2 Distribution of the percentage identity between pairs of homologous proteins. Pairs of direct orthologues (**a**–**c**) and paralogues (**d**) defined after classification of the proteins (see text) were used to compute the distributions. Amino acid identities were calculated from Smith–Waterman alignments (see Methods). Each distribution was computed from over 1,100 pairwise alignments (orthologues) and from 2,200 to 5,700 pairwise alignments (paralogues). Mean identity values between all orthologous proteins allow meaningful comparisons with other classes of organisms as indicated in the text (48–49% for all comparisons including *Y. lipolytica* (**c**), about 51% for the comparisons between *D. hansenii* with the three other yeasts (**b**), and 60–65% for *S. cerevisiae*, *C. glabrata* and *K. lactis* (**a**)). Species abbreviations as in Fig. 1.

genomes of *S. cerevisiae* and *C. glabrata*, other yeasts have acquired equivalent or even larger redundancies by other mechanisms (see Discussion).

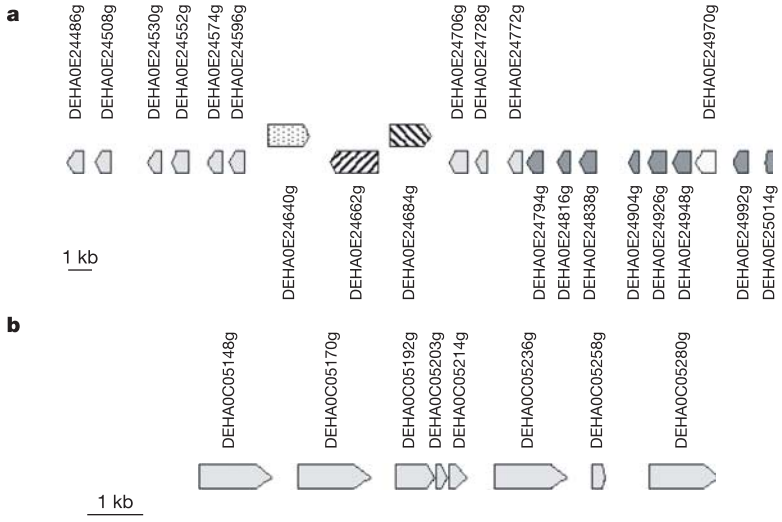
Conservation of synteny

Studying the conservation of synteny between species is another route to trace the evolutionary events that affect genomes. When applied to yeasts, this exercise reveals the considerable extent of genome reorganization that has occurred in this phylum (Fig. 5). The largest syntenic clusters between any two of our five yeasts contain only about 40 gene pairs. *Saccharomyces cerevisiae*, *C. glabrata* and *K. lactis* share about 500 syntenic clusters between them (considered in pairwise combinations), but the numbers and sizes of clusters rapidly decrease over increasing phylogenetic distances. We found only 74–149 syntenic clusters between *D. hansenii* and any other yeast species, and only 34–90 clusters between *Y. lipolytica* and any of the others. This extensive level of genomic map reshuffling contrasts with the near complete collinearity between genomes of the *Saccharomyces sensu stricto* group[14,15,27], but is coherent with the large evolutionary distance between our yeasts deduced from the degree of amino acid replacement between orthologous proteins (see above).

Despite the extent of map rearrangements, comparisons between *K. lactis* and *S. cerevisiae* reveal that 78% of the genes belonging to syntenic clusters correspond to intermingled series in which one region of the *K. lactis* genome corresponds to two (and sometimes more) distinct chromosomal regions in *S. cerevisiae*. Similar results have been reported separately by comparing the genomes of *K. waltii*[17] and *A. gossypii*[16] with *S. cerevisiae*. When *C. glabrata* is compared to *K. lactis*, only 64% of the genes belonging to syntenic clusters correspond to intermingled series in which one region of the *K. lactis* genome corresponds to two distinct chromosomal regions of *C. glabrata*. Again, this quantitative difference (64% compared with 78%) can only be reconciled with a general duplication event in the common ancestor of *S. cerevisiae* and *C. glabrata* if one assumes a more extensive rate of gene loss in the *C. glabrata* lineage than in *S. cerevisiae*.

Gene dynamics and sequence conservation

In addition to duplication and divergence, the gain and loss of specific genes may be critical for the functional differentiation between species and for evolution. Considering only the families of well conserved proteins between yeasts (to eliminate artificial threshold effects through a gradual increase in sequence divergence when phylogenetic distances increase), we have identified species-specific gene losses using, for criterion, the absence of a protein in one species and its presence in the four others (see Supplementary Table S10). The most striking example of species-specific gene losses is offered by *C. glabrata* where 29 genes are lost compared to all other yeasts. Losses seem to affect genes in a functionally coordinated manner, suggesting a reductive evolutionary scheme, possibly associated with the emergence of this yeast as a human pathogen. Specific losses in *C. glabrata* include genes involved in (1) galactose metabolism (five genes); (2) phosphate metabolism (four genes); (3) cell rescue, defence and virulence (three genes); and (4) nitrogen and sulphur metabolism (three genes). Specific gene losses were also found in *D. hansenii* (eight genes missing), *K. lactis* (five genes



c

Species	Tandem gene pairs: Total	No interruption	Interruption (1 gene)	Direct orientation (%)	Arrays: Total	Two copies	Average *E*-values: Tandem	All paralogues
S. cerevisiae	61	31	15	78.7	50	42	1.7×10^{-26}	4.9×10^{-22}
C. glabrata	47	34	5	83.0	32	24	6.4×10^{-30}	8.4×10^{-22}
K. lactis	36	27	4	72.2	33	30	1.4×10^{-24}	1.3×10^{-21}
D. hansenii	329	167	90	92.4	247	196	3.3×10^{-22}	1.4×10^{-21}
Y. lipolytica	54	39	5	72.2	48	45	3.6×10^{-23}	1.2×10^{-21}

Figure 3 Tandem gene repeats. **a**, Example of two large tandem arrays (nine gene copies plus one fragment (light grey), and eight gene copies (dark grey)) forming a partially intermingled structure on one *D. hansenii* chromosome. Genes are of unknown function. The first array is interspersed by three other genes (dotted and striped). **b**, Example of a tandem array of four gene copies encoding NADH dehydrogenase, plus a pseudogene and a gene fragment. **c**, Computation of all tandem gene arrays in the five yeast genomes (see Methods). The table indicates the total number of pairs of genes (column 2), either contiguous (column 3) or separated by one intervening gene (column 4). Column 5 indicates the proportion (%) of gene pairs in direct orientation.

missing) and *Y. lipolytica* (39 genes missing), but their functional coordination is less obvious.

Contrary to gene loss, the acquisition of new genes in a species raises the question of their origin. Horizontal gene transfer, which is quantitatively important in prokaryotes, is very rare in hemiascomycetes. A few examples of genes occurring in only one yeast species but having close homologues in Bacteria could, however, be detected (see Supplementary Table S11). This is the case for eight genes (including two pairs of paralogues) in *Y. lipolytica*, five genes (including one pair of paralogues) in *K. lactis* and one gene in *D. hansenii*. Many of them encode metabolic enzymes. No convincing case could be detected in *C. glabrata*. If these actually represent examples of horizontal gene transfer, this phenomenon accounts for less than 1% of the total gene number in yeasts. In addition, a small number of other genes seem to be species specific in the sense that they are limited to one of our five yeast species (as deduced from protein families; see Table 4) and have no significant homologue in general databases. The origins of these genes remain unclear and their function is generally unknown although, in many cases, they form paralogous families in the only species where they are found, indicating rapid expansion after their acquisition or *de novo* creation.

Finally, the broad evolutionary range covered by our sequenced yeast genomes allowed us to investigate the sequence conservation of proteins known to be involved in protein–protein interactions in *S. cerevisiae* compared to the average level of sequence conservation for all proteins. Previous comparisons with *S. pombe* and *C. elegans* already indicated the significant conservation of such proteins[28]. Systematic examination across the evolutionary span of hemiascomycetes (Table 5) shows that homologues of the set of 785 'interacting proteins' of *S. cerevisiae* are more conserved than average (homologues to the entire *S. cerevisiae* protein set). This trend is even more pronounced for the homologues of the 1,535 proteins of the 'complex set' and becomes more apparent when the phylogenetic distances increase (compare ratios in *C. glabrata* and *Y. lipolytica*).

Discussion

The central strategy of this work was to examine eukaryotic genome evolution by confining the comparisons within a single phylum, while exploring its evolutionary range as widely as possible. Each species revealed unique signatures in its genome, reflecting the fact that distinct mechanisms have predominated in each phylogenetic branch. With its larger genome size, not linked to a significantly larger gene repertoire, the highly redundant genome of *Y. lipolytica* shows a strong tendency for map dispersion. This dispersion is visible at various levels: a near complete absence of duplicated blocks despite a high number of paralogous genes; a higher number of tRNA genes; a higher number of rDNA loci; a dispersion of the 5S RNA genes; and the specific duplication of other non-coding RNA genes. By contrast, the other yeast species show significant constraints on genome size, possibly associated with their ability to duplicate genes in an ordered manner as revealed by the presence of duplicated blocks and tandem gene repeats in their genomes. This

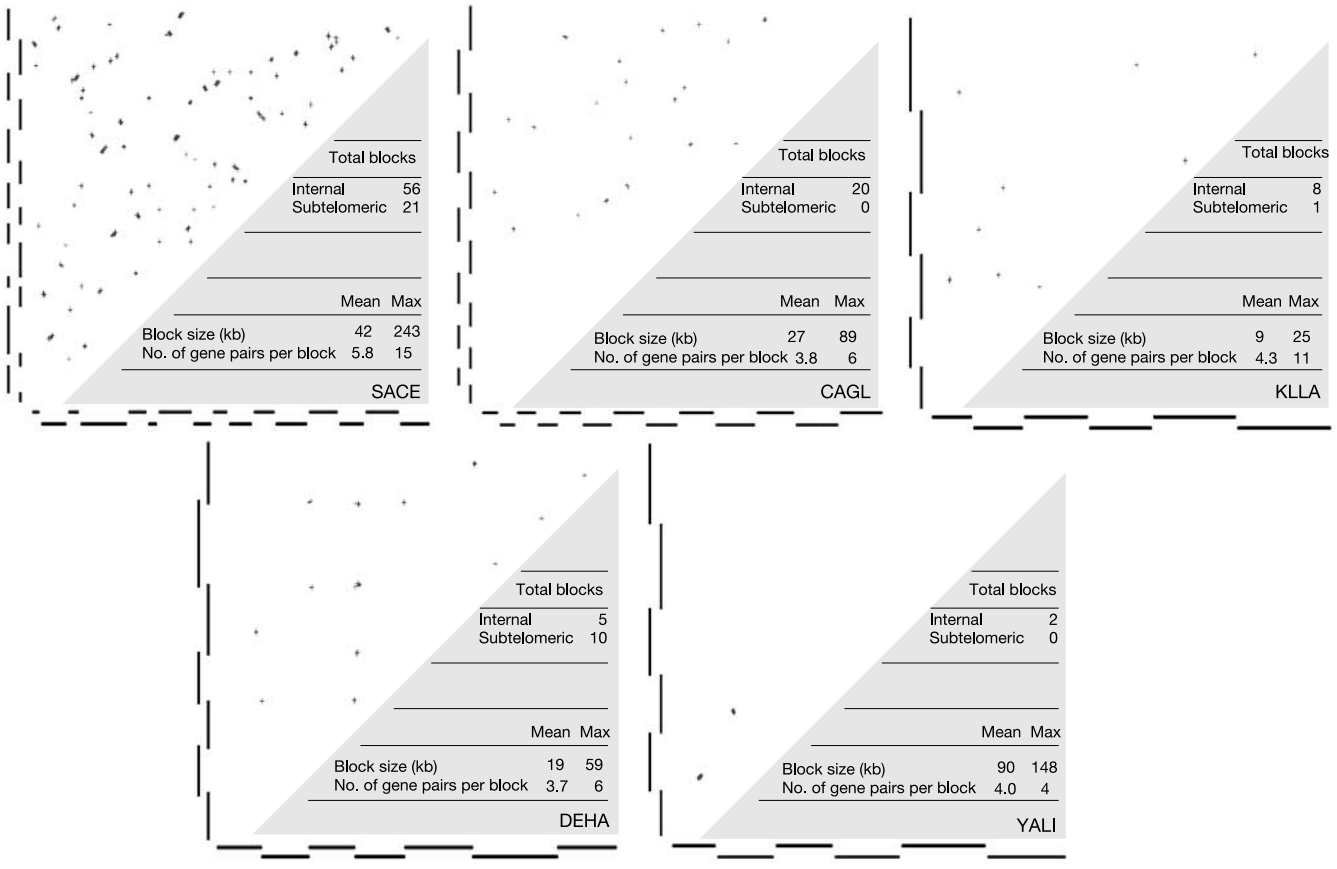
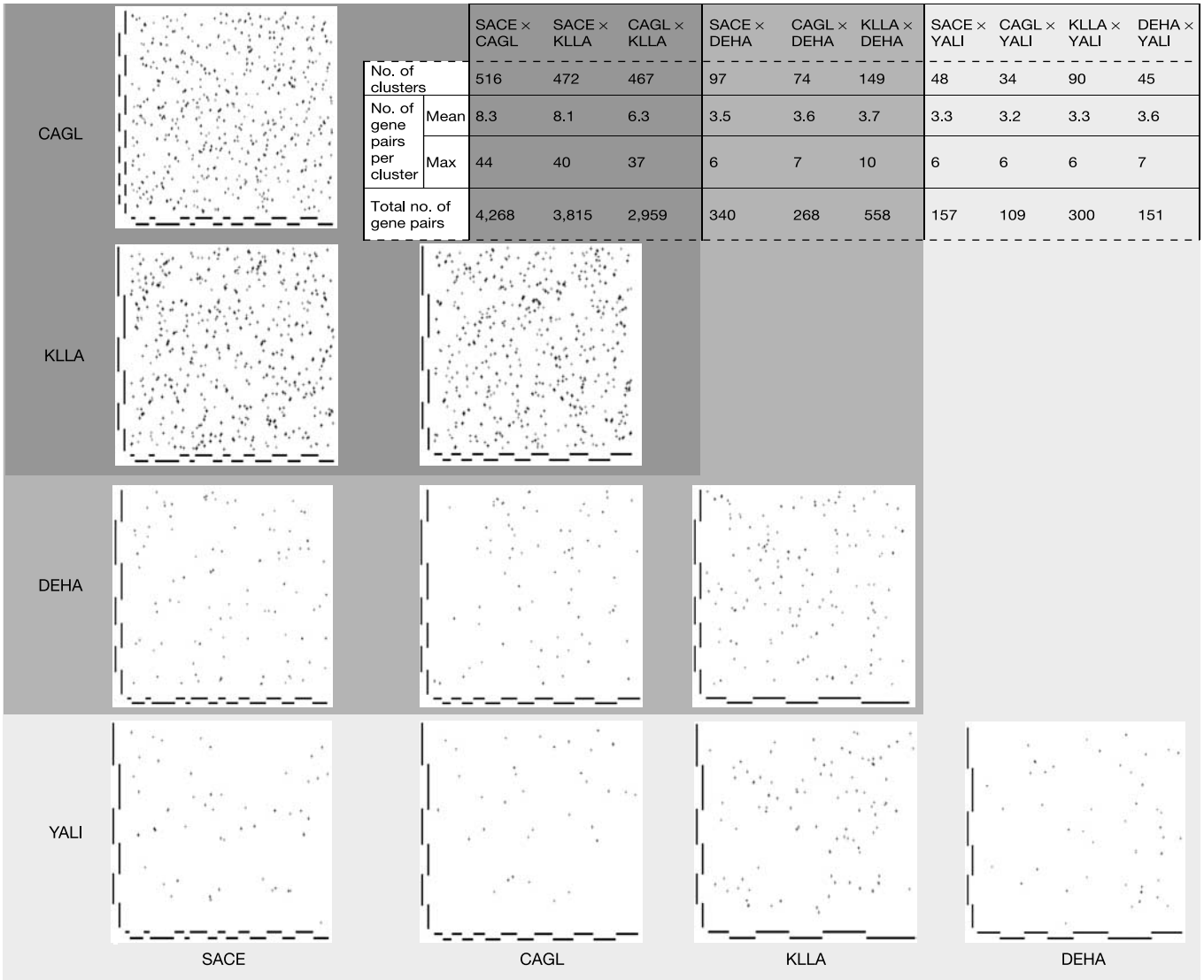


Figure 4 Detection of ancient duplicated blocks in each yeast genome. Chromosomal blocks containing at least three pairs of paralogous genes with a probability lower than 0.01 to occur by chance (see Methods) are plotted on two-dimensional diagrams each representing an entire yeast genome (chromosomes, represented by shifted bars, are ordered by increasing numbers from the lower-left corner). Each cross on the plots corresponds to a pair of paralogous genes belonging to a defined block. The shaded triangle in each plot gives numerical descriptions of the duplicated blocks found for each species. Species abbreviations as in Fig. 1.

last trend is particularly acute in *D. hansenii* but exists in all species. The remaining three species have acquired novel features such as the triplication of the *MAT* cassettes and short centromeres. Compared with *K. lactis*, which has the shortest and least redundant genome of the five yeast species studied, *S. cerevisiae* and *C. glabrata* partly share the traces of an extensive duplication event in their common ancestry (shared duplicated blocks). However, they present very different levels of duplication, which can only be reconciled with a single whole-genome duplication event, as proposed previously[16,17,24], if one assumes a significantly higher rate of loss of



		SACE × CAGL	SACE × KLLA	CAGL × KLLA	SACE × DEHA	CAGL × DEHA	KLLA × DEHA	SACE × YALI	CAGL × YALI	KLLA × YALI	DEHA × YALI
No. of clusters		516	472	467	97	74	149	48	34	90	45
No. of gene pairs per cluster	Mean	8.3	8.1	6.3	3.5	3.6	3.7	3.3	3.2	3.3	3.6
	Max	44	40	37	6	7	10	6	6	6	7
Total no. of gene pairs		4,268	3,815	2,959	340	268	558	157	109	300	151

Figure 5 Conservation of synteny between yeast genomes. Syntenic clusters (defined in Methods) are indicated by crosses on all possible pairwise comparisons of the five yeast genomes (same representation of chromosomes as in Fig. 4). The dark grey background indicates comparisons between the three closest relatives (*S. cerevisiae*, *C. glabrata* and *K. lactis*), medium grey indicates comparisons between these species and *D. hansenii*, and light grey indicates all comparisons including *Y. lipolytica*. The numerical characterization of syntenic clusters between any two yeast species is given in the inset (top). Species abbreviations as in Fig. 1.

Table 5 **Sequence conservation in interacting proteins and multiprotein complexes**

Protein sets	Total number of proteins in set	BLAST					ClustalW
		BDBH	*S. cerevisiae* proteins with a BDBH (%)				Alignment score
			C. glabrata	*K. lactis*	*D. hansenii*	*Y. lipolytica*	
S. cerevisiae set	5,807	2.62	74.1	73.0	63.8	51.5	13,587.4
Interacting set	785	3.26	89.8	87.9	80.3	68.0	14,351.9
Complex set	1,535	3.47	93.3	92.6	86.0	74.8	16,362.4

Experimentally defined protein–protein interactions in *S. cerevisiae* were taken from ref. 34. For each protein set of *S. cerevisiae*, sequence conservation in other genomes was estimated by bi-directional best hits (BDBH) in BLASTP alignments with *E*-value $\leq 10^{-8}$ (BLAST heading), and the average ClustalW alignment scores for *S. cerevisiae* proteins giving significant BDBHs with all four other genomes (ClustalW heading). The table indicates the average number (over each protein set) of the four yeast genomes giving a significant BDBH with *S. cerevisiae* proteins (column 3), and the percentage of proteins of each set giving a significant BDBH with each other yeast species (columns 4–7). The last column indicates the average ClustalW alignment scores for each set.

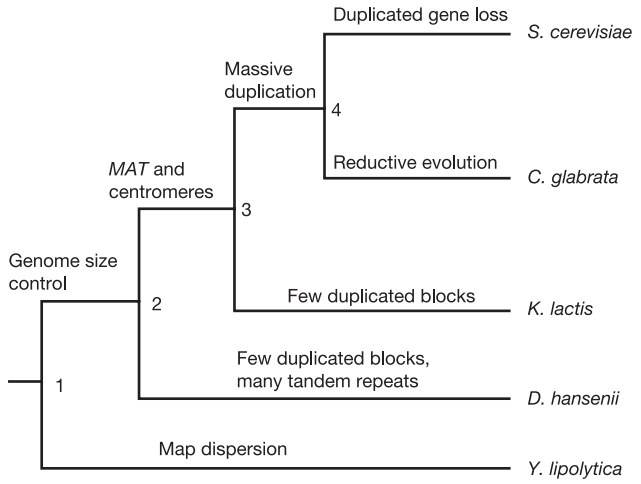


Figure 6 Major evolutionary events in the genomes of hemiascomycetes. Shown is a cartoon of the evolutionary history of the four sequenced yeasts, plus *S. cerevisiae*. The tree topology is based on 25S rDNA sequences (see Supplementary Fig. S17). The most conspicuous evolutionary signatures are summarized on each branch. The tendency for map dispersion of *Y. lipolytica* is visible at a variety of levels (see text). The other yeasts share an ability to duplicate genes in an ordered manner. An accidental whole-genome duplication event occurred in the common ancestor of *S. cerevisiae* and *C. glabrata*. It has been followed by extensive gene loss of paralogous copies.

the duplicated gene copies in the *C. glabrata* lineage than in *S. cerevisiae*. The reductive evolution of the *C. glabrata* genome was not anticipated and may be related to the adaptation of this yeast to life as a human pathogen.

Despite the conspicuous organizational differences between the five yeast genomes, all results are consistent with the topology of their phylogenetic tree (see Fig. 6). Accepting the risks inherent in the delicate exercise of reconstructing an evolutionary past in which accidental events may be misleading, we propose the following steps in the evolution of hemiascomycetes. At the separation between *Y. lipolytica* and the four other yeasts (node 1, Fig. 6), one branch lost the DNA transposons, kept the retrotransposons and responded to stronger genome size constraints (the nature of which remains to be elucidated) by simultaneously reducing its number of introns. In this context, the coordinated duplication of genes, either as large chromosome segments or as tandem repeats, became the major route to create the paralogous copies on which subsequent diversification is based. At the next separation (node 2), one branch, keeping a low level of segmental duplication, underwent extensive formation of tandem gene repeats, whereas the other, perhaps associated with the appearance of new centromeres providing for better chromosomal segregation, underwent segmental duplication to create the three *MAT* cassettes that changed the sexual capabilities and, consequently, the whole evolutionary future of species on this branch. At node 3, while some of these yeasts retained these features, leading to *K. lactis* (which can be regarded as representative of the ancestor of this branch), others underwent a whole-genome duplication event subsequently compensated by extensive gene loss among the newly formed paralogous pairs. The loss of paralogues sometimes leaves behind visible relics, as with *S. cerevisiae*[27]. In *C. glabrata*, this loss of paralogues has been so extensive as to result in a reductive evolution with loss of function and a general degree of genome redundancy nearly equivalent to that of *K. lactis* (and significantly lower than that of *Y. lipolytica* and, even more so, *D. hansenii*).

The diversity of evolutionary routes taken by each branch was unexpected at the start of our work. However, it becomes less surprising if one considers that, despite their morphological and biological similarities, the five hemiascomycete yeast species encompass an evolutionary span as large as the entire phylum of chordates. Relative to this broad evolutionary spectrum, the number of yeast species now entirely or partially sequenced remains limited. With over 700 species described[29], other interesting novelties may lie hidden in the genomes of other hemiascomycetes. Together with the fact that yeasts are such powerful experimental systems, this should stimulate future studies with impacts that reach far beyond yeasts. □

Methods

Sequencing and sequence assembly

Plasmid libraries (insert sizes 3–5 kb) were constructed using randomly sheared total DNA purified from the haploid type strain of each species and used for shotgun sequencing (Table 1). Bacterial artificial chromosome (BAC) libraries were used for sequence finishing, as required, and for assembly verification. Sequence electrophoresis and base calling were performed on either a 3700 Genetic Analyser (ABI PRISM BigDye Terminator) or a Licor 4200L DNA sequencer (dye primers). The traces were assembled using Phrap and the resulting contigs were checked and ordered using in-house software (Cover and Coverparse; M. Levy, unpublished). Gaps were closed by the primer-walking method. Regions of insufficient quality were resequenced using either new primers or a different chemistry. Correct assembly and collinearity of contigs and scaffolds were systematically (*C. glabrata*) or partially (*K. lactis*, *Y. lipolytica*) verified using BAC fingerprints.

Annotation

Genome annotation was conducted using the CAAT-Box system[30]. Automatic predictions were manually curated and each open reading frame (ORF) was classified as coding (CDS), non-coding (false ORF) or pseudogene according to Génolevures annotation standards (http://cbi.labri.fr/Genolevures). The codon usage matrix used to screen for ORFs was trained on a sample of coding and non-coding sequences, initially produced by automatic identification, and subsequently using curated CDS. Annotation was performed iteratively across successive sequence assemblies (each file is conserved if no sequence edition occurs, and it is replaced otherwise). In a final stage, 291 small ORFs missed by the automatic procedures were identified with the help of the protein family classification (each protein of incomplete families was used as a tBLASTn query against all remaining intergenic sequences). Introns were identified from consensus splice sites and branch points, and from comparisons to *S. cerevisiae*[31]. Genes encoding tRNA molecules were identified according to ref. 18 (see Supplementary Table S3). All other non-coding RNA molecules (Supplementary Table S6), rRNAs, transposons, centromeres (Supplementary Table S1) and telomeric repeats (Supplementary Table S2) were identified by sequence comparisons and manual curation.

All annotated genetic elements were designated using a new nomenclature system (http://cbi.labri.fr/Genolevures). Briefly, elements are numbered serially along each sequence contig or scaffold from the left to right of each chromosome using 11 incremental steps (to limit errors and offer the possibility for subsequent insertion of newly recognized elements). The element nomenclature indicates the species (four letters), the project or strain number (one numeral), the chromosome (one letter) followed by the serial number (for example, CAGL0G08492g). The suffix identifies the type of element ('g' stands for any element whose RNA product may be translated by the genetic code; 'r' for elements whose RNA product is not translated; 's' for a *cis*-acting element; and 'v' for intergenes (intervening)).

Classification of proteins into families

A data set of 34,824 amino acid sequences corresponding to all annotated and predicted proteins (not necessarily subsequently confirmed) from our four yeast species plus the annotated proteins of *S. cerevisiae* was used to construct protein families. Two sets of sequence alignments were produced separately using Smith–Waterman and BLAST algorithms. After filtration for statistical significance, the pairs considered as homologous were clustered using the MCL method[32] with a variety of inflation parameters (see Supplementary S12). Results of clusterings applied separately on the Smith–Waterman and BLAST pairs were compared using a graph-based technique producing a best coincidence of 3,016 strictly identical protein families (exactly the same set of proteins using the two alignment methods). Another set of 394 families, corresponding to exactly the same set of proteins split into two families (or merged into a single one) when comparing the two methods, was easily reconciled using motif search programs. The two sets form the 'robust' families. Reconciliation was not attempted for the more complex cases in which the same set of proteins was classified into partially overlapping families by each of the two methods. Instead, the union of all such proteins was retained as 'consensus' families (a total of 1,311 families).

Identification of duplicated blocks and syntenic clusters

Using homologous gene pairs as defined from protein families, maps were compared with the ADHoRe program[33] (r^2 cutoff = 0.8, maximum gap size = 35 genes, minimum number of pairs = 3). Results were filtered such that the maximum probability for a segment to be generated by chance was lower than 0.01. Paralogous pairs were used to define ancient duplicated blocks within each genome (Fig. 4), whereas orthologous pairs were used to define conserved syntenic clusters between two genomes (Fig. 5).

Identification of tandem arrays

Each genome was systematically examined for the presence of repeated gene arrays. Arrays are defined as a succession of genes encoding paralogous proteins (as defined by protein families and with a BLASTP *E*-value $<10^{-20}$) allowing a maximum of ten intervening genes between two closest members of an array. Arrays were defined regardless of gene orientation.

Received 3 February; accepted 19 April 2004; doi:10.1038/nature02579.

1. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 563–567 (1996).
2. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).
3. Amanatides, P. G. *et al.* The genome sequence of *Drosophila melanogaster. Science* **287,** 2185–2195 (2000).
4. The *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
5. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
6. Holt, R. A. *et al.* The genome sequence of the malaria mosquito *Anopheles gambiae. Science* **298,** 129–149 (2002).
7. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes. Science* **297,** 1301–1310 (2002).
8. Stein, L. D. *et al.* The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol.* **1,** 166–192 (2003).
9. Stephen, A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296,** 92–100 (2002).
10. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science* **296,** 79–92 (2002).
11. Galagan, J. E. *et al.* The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422,** 857–868 (2003).
12. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe. Nature* **415,** 871–880 (2002).
13. Souciet, J. L. *et al.* Genomic exploration of the hemiascomycetous yeasts. *FEBS Lett.* **487,** 3–147 (2000).
14. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423,** 241–254 (2003).
15. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301,** 71–76 (2003).
16. Dietrich, F. S. *et al.* The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304,** 304–307 (2004).
17. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428,** 617–624 (2004).
18. Marck, C. & Grosjean, H. tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA* **8,** 1189–1232 (2002).
19. Soonhammer, E. L. & Koonin, E. V. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* **18,** 619–620 (2002).
20. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
21. Dehal, P. *et al.* The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298,** 2157–2167 (2002).
22. Brown, C. J., Todd, K. M. & Rosenzweig, R. F. Multiple duplications of yeast hexose transport gene in response to selection in a glucose-limited environment. *Mol. Biol. Evol.* **15,** 931–942 (1998).
23. Lalo, D., Stettler, S., Mariotte, S., Slonimski, P. P. & Thuriaux, P. Two yeast chromosomes are related by a fossil duplication of their centromeric regions. *C. R. Acad. Sci. III* **316,** 367–373 (1993).
24. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387,** 708–713 (1997).
25. Wong, S., Butler, G. & Wolfe, K. H. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc. Natl Acad. Sci. USA* **99,** 9272–9277 (2002).
26. Koszul, R., Caburet, S., Dujon, B. & Fischer, G. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J.* **23,** 234–243 (2004).
27. Fischer, G., Neuveglise, C., Durrens, P., Gaillardin, C. & Dujon, B. Evolution of gene order in the genomes of two related yeast species. *Genome Res.* **11,** 2009–2019 (2001).
28. Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. Evolutionary rate in the protein interaction network. *Science* **296,** 750–752 (2002).
29. Kurtzman, C. P. & Fell, J. W. *The Yeasts, a Taxonomic Study* 4th edn (Elsevier, Amsterdam, 1998).
30. Frangeul, L. *et al.* CAAT-Box. Contigs-Assembly and Annotation Tool-Box for genome sequencing projects. *Bioinformatics* **20,** 790–797 (2004).
31. Bon, E. *et al.* Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res.* **31,** 1121–1135 (2003).
32. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (2002).
33. Vandepoele, K., Saeys, Y., Simillion, C., Raes, J. & Van De Peer, Y. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res.* **12,** 1792–1801 (2002).
34. Bader, G. D., Betel, D. & Hogue, C. W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31,** 248–250 (2003).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank M. Tichit, S. Duthoy, S. Ferry and N. Zidane for technical assistance; A. Louis for help and expertise in the use of the INFOBIOGEN computational facilities; O. Jaillon, H. Roest-Crollius and their colleagues for sharing unpublished results on *Tetraodon negroviridis*; and A. Goffeau, H. Feldmann, A. Nicolas, N. Huu-Vang, J.-P. Latgé, I. Moszer and M. Vergassola for discussions and advice. This work was supported by the Consortium National de Recherche en Génomique (to Génoscope and to Institut Pasteur Génopole), the CNRS (GDR 2354, Génolevures sequencing consortium) and the 'Conseil Régional d'Aquitaine'. B.D. is a member of Institut Universitaire de France.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.-L.S. (souciet@gem.u-strasbg.fr) or B.D. (bdujon@pasteur.fr). Sequences have been deposited in EMBL under accession numbers CR380947–CR380959 for *C. glabrata*, CR382121–CR382126 for *K. lactis*, CR382127–CR382132 for *Y. lipolytica* and CR382133–CR382139 for *D. hansenii*.