Genomic Exploration of the Hemiascomycetous Yeasts: 21. Comparative functional classification of genes

Claude Gaillardin[a,]*, Guillemette Duchateau-Nguyen[b], Fredj Tekaia[c], Bertrand Llorente[c], Serge Casaregola[a], Claire Toffano-Nioche[b], Michel Aigle[d], François Artiguenave[e], Gaëlle Blandin[c], Monique Bolotin-Fukuhara[b], Elisabeth Bon[a], Philippe Brottier[e], Jacky de Montigny[f], Bernard Dujon[c], Pascal Durrens[d], Andrée Lépingle[a], Alain Malpertuy[c], Cécile Neuvéglise[a], Odile Ozier-Kalogéropoulos[c], Serge Potier[f], William Saurin[e], Michel Termier[b], Micheline Wésolowski-Louvel[g], Patrick Wincker[e], Jean-Luc Souciet[f], Jean Weissenbach[e]

[a] *Collection de Levures d'Intérêt Biotechnologique, Laboratoire de Génétique Moléculaire et Cellulaire, INRA UMR 216, CNRS URA 1925, INA-PG, BP01, F-78850 Thiverval-Grignon, France*

[b] *Informatique et Génomes, Institut de Génétique et Microbiologie, Université de Paris Sud, F-91405 Orsay, France*

[c] *Unité de Génétique Moléculaire des Levures, Institut Pasteur/URA2171 CNRS and UFR925 Université Pierre et Marie Curie, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France*

[d] *Laboratoire de Biologie Cellulaire de la Levure, IBGC, 1 rue Camille Saint-Säens, F-33077 Bordeaux Cedex, France*

[e] *Genoscope, Centre National de Séquençage, 2 rue Gaston Crémieux, F-91006 Evry Cedex, France*

[f] *Laboratoire de Génétique et Microbiologie, UPRES-A 7010 ULP/CNRS, Institut de Botanique, 28 rue Goethe, F-67000 Strasbourg, France*

[g] *Microbiologie et Génétique ERS 2009, CNRS/UCB/INSA, bat. 405 R2, Université Lyon I, 43 boulevard du 11 novembre 1918, F-69622 Villeurbanne Cedex, France*

Received 3 November 2000; accepted 9 November 2000

First published online 29 November 2000

Edited by Horst Feldmann

Abstract We explored the biological diversity of hemiascomycetous yeasts using a set of 22 000 newly identified genes in 13 species through BLASTX searches. Genes without clear homologue in *Saccharomyces cerevisiae* appeared to be conserved in several species, suggesting that they were recently lost by *S. cerevisiae*. They often identified well-known species-specific traits. Cases of gene acquisition through horizontal transfer appeared to occur very rarely if at all. All identified genes were ascribed to functional classes. Functional classes were differently represented among species. Species classification by functional clustering roughly paralleled rDNA phylogeny. Unequal distribution of rapidly evolving, ascomycete-specific, genes among species and functions was shown to contribute strongly to this clustering. A few cases of gene family amplification were documented, but no general correlation could be observed between functional differentiation of yeast species and variations of gene family sizes. Yeast biological diversity seems thus to result from limited species-specific gene losses or duplications, and for a large part from rapid evolution of genes and regulatory factors dedicated to specific functions. © 2000 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

Key words: Biodiversity; Speciation; Pathway; Gene family

1. Introduction

Yeasts have evolved to thrive in very different environments: trees and fruits in the case of the sugar fermenting *Saccharomyces*, soil and environments rich in decaying organic compounds in the case of saprophytic organisms like *Yarrowia lipolytica* or occasional pathogens like *Candida tropicalis*. Some grow under quite inhospitable conditions like the marine, osmotolerant yeast *Debaryomyces hansenii*. Adaptation to these conditions is correlated with quite different metabolic orientations, ranging for example from mostly fermentative in the case of *Saccharomyces sensu stricto* to strictly respiratory in the case of *Y. lipolytica*. Physiological diversity among yeast species is also exemplified by the occurrence of specific metabolic pathways, like nitrate assimilation in the case of *Pichia angusta*, that have been extensively used as taxonomic markers.

Two mutually non-exclusive possibilities can be envisioned to account for this diversity: (i) all species have basically the same set of genes, but they favor specific adaptations and colonize specific habitats through selection of regulatory mechanisms at large: sensory cascades, level of expression, specific regulatory networks, etc., (ii) individual species retain specific sets of genes from their ancestor which are amplified and specialized as paralogous copies, or lose entire sets of genes while acquiring completely new pathways through horizontal transfer. Analysis of whole genomes from very different, mostly prokaryotic organisms gave clear evidences for the last type of events [1], and limited gene transfer was shown to occur between closely related yeast species [2]. On the other hand, gene duplication and emergence of gene families have been proposed to play a major role in the evolution of species, both to amplify gene expression and to permit divergence and appearance of novel functions [3]. Massive gene duplications followed by massive losses have been proposed to account for the observed level of gene duplication in *Saccharomyces cere-*

*Corresponding author.
E-mail: claude@grignon.inra.fr

PII: S0014-5793(00)02292-4

visiae and these events may have been selected in a function-oriented way [4].

We investigated on a large set of genes newly identified among 13 species of hemiascomycetous yeasts [5–17] how partial genomic data can be used to depict this biological diversity. This was facilitated by ascribing each gene to a common category of the functional catalogue. Detailed statistics on these genes showed that indeed prominent differences in function representation could be detected when the 13 sets of genes were compared. We further attempted to correlate the size of gene families [18] and the over-representation of genes belonging to functional categories, in order to approach the nature of events that led to the physiological diversity of these species.

2. Materials and methods

2.1. Nucleic acid sequences

Sequences from the 13 species analyzed here have been obtained within the Génolevures project [19] funded by Genoscope (Evry, France).

2.2. Classifying genes into functional categories

We made the hypothesis that genes having a match in *S. cerevisiae* shared the same function as their *S. cerevisiae* homologue. This permitted using MIPS's functional catalogue of 6415 *S. cerevisiae* open reading frames (ORFs) (http://www.mips.biochem.mpg.de/proj/yeast/catalogues/index.html) (April 6, 2000 release), see also Mewes et al. [20]. In this catalogue, 3725 ORFs are assigned to at least one of 204 functional sub-classes grouped into 13 main classes, 148 ORFs are assigned to class 98 (classification not yet clear cut) and 2534 ORFs remain unclassified (class 99). A given ORF may be present in different classes. We used a simplified version of the MIPS catalogue where several small-sized sub-classes were fused: 09.07, 09.08 and 09.09 into 09.09 (biogenesis of endoplasmic reticulum (ER), Golgi and intracellular transport vesicles), 09.22 and 09.25 into 09.25 (biogenesis of endosomes, vacuoles and lysosomes), 11.13 and 11.99 into 11.99 (other cell rescue activities). Matches against GPROTEOME [21], or against *S. cerevisiae* or *Schizosaccharomyces pombe* ORFs absent from the above catalogue, were assigned to one or several sub-classes, using the classification proposed by MIPS for related genes whenever possible. Most analyses were done on main classes or first-level sub-classes for statistical significance. Since a given gene may be present in different sub-classes of the same functional class, duplicates were systematically removed before combining sub-classes.

2.3. Using correspondence analysis as a descriptive method for global analysis of functions

Our aim was to compare the *n* considered species according to the distribution of genes into *m* functional classes and to extract all synthetic relationships between species and functional categories. The gene distribution observed in a given species *i* among the *j*th functional category was computed as described above (see Table 1). The most appropriate method to handle such data tables is correspondence analysis [22,23]. This procedure represents functional classes and yeast species in a multidimensional space. It then extracts an orthogonal system of axes (or factors), in decreasing order of information amount (or inertia) each factor represents. Thus, the first axis represents the maximal species and function dispersion, the second axis, orthogonal to the previous one, represents the maximal portion of the remaining inertia, and so on, for the axes to follow. As a result, points that are close to the barycenter of the cloud appear as non-significant contributors to the global inertia, whereas points far away from it identify major contributors.

Correspondence analysis allows displaying species and functional categories on a given factorial plane, generally defined by the first and second factors, which corresponds to the largest proportions of the total information included in the data table. Careful inspection is needed in order to avoid misinterpretation of observed neighborhood. On the figures shown here, functional categories or yeast species are indicated by boxes when they were well represented on one axis at least and contributed strongly to the inertia of the whole data set.

Each species (and each functional category) is represented by its coordinates in this system, allowing calculation of distances between species and/or functions. Classification of these data according to their neighborhood in the factorial space leads to a hierarchical tree, which gives a graphical representation of subsets of species having similar functional profiles, or of functions similarly distributed between the different species.

Correspondence analysis was performed using Xlstat (Data Analysis and Statistical Solution for Microsoft EXCEL, http://www.xlstat.com).

2.4. Estimation of total gene numbers in the functional classes

From the known number G_i of genes assigned in *S. cerevisiae* to a functional class *i*, the estimated total number g_i of genes in class *i* of a given species was estimated as $g_i = G_i/G_T \times g_T$, where G_T is the total number of genes identified in *S. cerevisiae* and g_T the total number of genes identified through homology searches in that species. Since g_T derives from minimal gene number estimates [18,21] and not from relative physical genomic coverage, g_i values are underestimates, especially for species that are phylogenetically distant from *S. cerevisiae*.

3. Results and discussion

3.1. Nature of genes having no clear homologues in *S. cerevisiae*

Within the Génolevures project [19], random sequence tags (RSTs) of 13 hemiascomycetous yeasts were screened for the presence of homologous genes identified through BLASTX searches against ORFs either present in *S. cerevisiae* or in GPROTEOME, a non-redundant database derived from SwissProt and completely sequenced genomes [21]. Matches in GPROTEOME identified in the 13 species 590 genes without close orthologue in the sequenced *S. cerevisiae* strain. In the different species including *Saccharomyces bayanus* var. *uvarum*, which is *S. cerevisiae*'s closest relative in the set of species considered, these genes often matched the same GPROTEOME ORF (see Table 1). This is the case for the most frequently identified gene, a homologue of *S. pombe MLO2* involved in mitosis, which was identified in six species, and e.g., for the uric acid-xanthine permease gene homologous to *uapA* in *Emericella nidulans* found in three species. Sixty seven of the 590 hits (11%) are shared by two species or

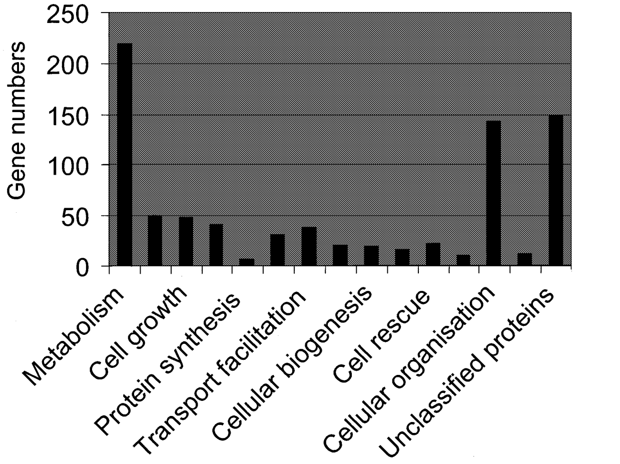

Fig. 1. Proteins identified through similarity to GPROTEOME correspond mainly to genes involved in metabolism and cellular organization. Minimal gene number values of ORFs identified in the 13 species through GPROTEOME searching (ordinates) are plotted as a function of functional class assignation (abscissa). Classes 98 and 99 were combined.

more, showing that these GPROTEOME homologues are well conserved in other species. This suggests that these genes either rapidly diverged or were even lost in the branch leading to *S. cerevisiae*.

A total of 447 different GPROTEOME ORFs were thus identified. They were distributed into 18, 71, 384 and 117 orthologues of genes identified in Archaea, bacteria, fungi and other eukarya, respectively. Yeast species distantly related to *S. cerevisiae* tend to have more homologues: 14 and 161 matches were identified for *S. bayanus* var. *uvarum* and *Y. lipolytica* (respectively) for which ca. 5000 RSTs were analyzed, vs. three and 38 matches for *S. servazzii* and *C. tropicalis* (respectively), two species with 2500 RSTs. Genes of known function (327 or 75% of the total) are mostly involved in metabolism or cellular organization (see Fig. 1). A conservative estimate identifies at least 275 different activities.

For most of them, related genes do exist in *S. cerevisiae*, but their similarity scores were judged non-significant, whereas much better scores were observed with GPROTEOME. For example, the *CLN2* and *STE7* genes of *Candida albicans* permitted identification of homologues in *P. sorbitophila* and *Y. lipolytica*, and in *C. tropicalis* and *P. sorbitophila*, respectively. However, *Ca*Cln2p and *Ca*Ste7p share 32% and 36% identity with their *S. cerevisiae* homologues. Similarly, a direct search of *P. angusta* genes deposited in GenBank permitted identification of *PaURA3*, a gene that remained undetected when *P. angusta* RSTs were searched using *S. cerevisiae* ORFs [13]. Other genes that exist in *S. cerevisiae* were identified through an even more distant species search: whereas genes encoding dihydrorotate dehydrogenase (an uracil biosynthetic enzyme) could be identified in *Zygosaccharomyces rouxii* and *K. thermotolerans* through their *S. pombe* homologue, they were detected in *S. kluyveri* and *P. angusta* when compared to *Arabidopsis thaliana* and *Rattus norvegicus*. These genes seem to diverge rapidly in distant species and are therefore elusive *S. cerevisiae* homologues.

A few genes are absent from the sequenced *S. cerevisiae* strain, but are still present and even amplified in several other *S. cerevisiae* strains like the *MEL* genes (α-galactosidase) [24] found in *S. kluyveri* and *S. bayanus* var. *uvarum*.

About 100 genes with a known function in another organism seem to have no homologue at all in *S. cerevisiae*. A non-exhaustive list of the corresponding enzymatic functions encompasses β-galactosidase, sorbitol utilization enzymes, arylsulfatases, renal dipeptase (a microsomal zinc-dependent dipeptidase), uricase (a peroxisomal enzyme oxidizing uric acid to allantoin), iron/ascorbate reductase, complex I genes (see below) or queuine tRNA-ribosyltransferase. All these genes are present and conserved in several ascomycetes and higher eukaryotes (like *S. pombe* or *Aspergillus nidulans*), and even in bacteria.

Finally, a few matches identify genes that have not been found at all in ascomycetes until now, like genes for β-lactamase or pristinamycine synthase subunit A. Dibenzothiophene (DBT) desulfurization enzymes, converting DBT to sulfite and 2-hydroxybiphenyl, have so far been identified in bacteria only [25], and are involved in the metabolism of sulfur-containing fossil fuel compounds: homologues of the *soxA* gene of *Rhodococcus* sp. encoding FAD-linked monooxygenase were found in *Y. lipolytica, P. angusta* and *S. kluyveri*. Very significant similarities were found between two *C. tropicalis* ORFs and bacterial genes: a formamidase gene of *Methylophilus methylotrophus* also found in *P. angusta* and a putative monooxygenase of *Rhizobium* sp. [16]. Whether these genes may represent cases of recent horizontal gene transfer awaits more extensive characterization.

An interesting situation was found for hydantoin utilization genes. Several bacteria are able to convert 5-substituted hydantoins into the corresponding L-amino acids in three steps, a process that might be relevant for industrial synthesis of amino acids [26–29], through a three-step pathway which is the reverse of the one used for pyrimidine biosynthesis. Genes resembling *hyuA* and *hyuC* were found in *P. angusta* and *S. kluyveri*, whereas homologues of *hyuC* were identified in *K. thermotolerans* and *Y. lipolytica*. These genes were not identified through similarity to YKL215c, a *S. cerevisiae* gene of unknown function similar to *Pseudomonas aeruginosa hyuA*, suggesting that they are ancient in yeasts and have strongly diverged. Since no *hyuB* or *hyuE* homologues were found and since yeasts to our knowledge have not been reported to use 5-substituted hydantoins, these genes may alternatively define a pyrimidine degradation pathway, similar to the one recently described in *S. kluyveri* [30]. A similar case may possibly be made for the four dihydroorotate dehydrogenase genes found in *S. kluyveri, P. angusta, K. thermotolerans* and *Z. rouxii*: some of them may actually encode the first step of the uracil degradation pathway instead of the third biosynthetic step of orotic acid.

3.2. Species-specific metabolic functions identified through comparison with GPROTEOME

Although GPROTEOME permitted identification of only a limited set of proteins absent from *S. cerevisiae*, we wondered whether these could be associated with functions characteristic of the different yeast species.

A clear case can be made with the mitochondrial complex I which catalyzes oxidation of NADH by ubiquinone, linked to proton transfer across the mitochondrial membrane. It is present in all higher eukarya, and is encoded partly by the nucleus, partly by organellar DNA (mitochondrial or chloroplast DNA). Both sets of genes have apparently been lost in the branch leading to the *Saccharomyces* genus [31]. Since there are more than 20 nuclear encoded subunits, finding or not complex I genes in our sequence samples would be highly indicative of presence or absence of complex I. Nuclear genes for complex I were found in *Y. lipolytica, C. tropicalis, D. hansenii, P. angusta, P. sorbitophila*, but conspicuously not in any of the *Saccharomyces* species, neither in *Z. rouxii* nor in any of the *Kluyveromyces* species tested. These results correlate well with the data on mitochondrial DNA (mtDNA) when available (see [6–17]). Our data do not exclude that partial loss occurred in some species, like in *S. pombe* where mtDNA encoded subunits were all lost, whereas one nuclear gene (24 kDa subunit) could still be identified through systematic sequencing (V. Wood, personal communication).

Several of the species studied here are known to metabolize substrates that *S. cerevisiae* cannot. *C. tropicalis, D. hansenii, Kluyveromyces lactis, Kluyveromyces marxianus* and the two *Pichia* species are able to use β-D-glucosides (cellobiose, arbutin, salicin), whereas the remaining species cannot. Genes encoding β-glucosidase were found in five out of the six species growing on β-glucosides. All these six species plus *K. thermotolerans, Z. rouxii* and *S. bayanus* var. *uvarum* are able to use sorbitol (D-glucitol) as a carbon source: 1–3 copies of homo-

Table 1
Matches in GPROTEOME shared by two species at least

ORF in GPROTEOME	Number of hits	*C. tropicalis*	*D. hansenii*	*K. lactis*	*K. marxianus*	*K. thermotolerans*	*P. angusta*	*P. sorbitophila*	*S. exiguus*	*S. kluyveri*	*S. servazzii*	*S. uvarum*	*Y. lipolytica*	*Z. rouxii*	Species	Function, gene name
SPBC4.05	6			x	x	x				x	x			x	*S. pombe*	involved in mitosis MLO2
P78771	5				x	x	x			x				x	*S. pombe*	hypothetical protein
P87219	5	x			x	x				x				x	*C. albicans*	sorbitol dehydrogenase SOU1
P07337	4		x	x	x		x								*K. marxianus*	β-glucosidase precursor BGLS
P48777	4					x	x	x						x	*E. nidulans*	wide specificity purine permease UAPC
P49374	4			x	x								x	x	*K. lactis*	glucose transporter high affinity HGT1
P51691	4			x		x	x						x		*P. aeruginosa*	arylsulfatase
P87218	4			x	x	x	x								*C. albicans*	sorbitol utilization protein SOU2
Q12556	4			x			x						x	x	*Aspergillus niger*	copper amine oxidase 1 AMO1
SPAC11D3.09	4	x					x	x						x	*S. pombe*	agmatinase precursor (putative)
SPAC12B10.16C	4					x		x		x			x		*S. pombe*	hypothetical protein
SPAC19G10.13	4		x	x	x		x								*S. pombe*	hypothetical protein
SPAC8F11.02C	4				x		x		x		x				*S. pombe*	hypothetical protein
SPBC354.15	4		x	x			x			x					*S. pombe*	fructosyl amino acid oxidase (putative)
SPCC1450.07C	4			x	x		x						x		*S. pombe*	D-amino acid oxidase (putative)
SPCC965.12	4		x	x			x	x							*S. pombe*	dipeptidase cytoplasmic (putative)
HI0588	3			x		x				x					*Haemophilus influenzae*	*N*-carbamyl-L-amino acid amidohydrolase
P24918	3						x	x					x		*Neurospora crassa*	NADH-ubiquinone oxidoreductase (CI-78KD)
P54995	3						x			x			x		*Rhodococcus* sp.	DBT desulfurization enzyme A *soxA*
Q07307	3						x			x				x	*E. nidulans*	uric acid-xanthine permease UAPA
Q53389	3					x				x			x		*Bacillus stearothermophilus*	*N*-carbamyl-L-amino acid amidohydrolase
Q99042	3			x		x	x								*Trigonopsis variabilis*	D-amino acid oxidase OXDA
SPAC1F8.03C	3		x	x			x								*S. pombe*	major facilitator family
SPAC22H10.08	3			x			x			x					*S. pombe*	hypothetical protein
SPAC8E4.03	3	x		x			x								*S. pombe*	agmatinase precursor (putative)
SPBC3B8.07C	3	x		x			x								*S. pombe*	fatty acid desaturase (putative)
SPBC660.12C	3			x		x	x								*S. pombe*	aminotransferase pyridoxal pH-dependent (class V)
SPCC1494.01	3		x	x			x								*S. pombe*	hypothetical protein
SPCC550.07	3				x	x								x	*S. pombe*	acetamidase
Q10193	3			x			x			x					*S. pombe*	putative splicing protein SRP1, RNA binding
P07921	3		x	x			x								*K. lactis*	lactose permease (LAC12)

Table 1 (*continued*)

ORF in GPROTEOME	Number of hits	*C. tropicalis*	*D. hansenii*	*K. lactis*	*K. marxianus*	*K. thermo-tolerans*	*P. angusta*	*P. sorbito-phila*	*S. exiguus*	*S. kluyveri*	*S. servazzii*	*S. uvarum*	*Y. lipolytica*	*Z. rouxii*	Species	Function, gene name
CEZK455.4	2								x	x					*Caenorhabditis elegans*	sphingomyelin phosphodiesterase (similar to)
Cj1199	2				x								x		*Campylobacter jejuni*	iron/ascorbate-dependent oxidoreductase (putative)
MTRv0773c	2			x			x								*Mycobacterium tuberculosis*	γ-glutamyl transpeptidase (putative)
O05691	2		x				x								*Rhodococcus erythropolis*	non-heme haloperoxidase THCF
P05982	2					x			x						*R. norvegicus*	NAD(P)H dehydrogenase (quinone)
P00723	2			x	x										*K. lactis*	β-galactosidase. LAC4 protein
P15559	2									x	x				*Homo sapiens*	NAD(P)H dehydrogenase [quinone]
P16932	2					x								x	*Burkholderia cepacia*	2,2-dialkylglycine decarboxylase
P22506	2	x	x												*S. fibuligera*	β-glucosidase 1 or 2 precursor
P30887	2	x					x								*Y. lipolytica*	acid phosphatase precursor PHO2
P32747	2					x								x	*S. pombe*	dihydroorotate dehydrogenase ura3
P38680	2			x									x		*N. crassa*	methyltryptophane transport system
P41946	2									x		x			*S. cerevisiae*	α-galactosidase precursor MEL1, MEL2...
P43062	2							x					x		*C. albicans*	G1-specific cyclin (Cln2)
P46463	2						x						x		*P. pastoris*	peroxisome biosynthesis protein PAS1
P46599	2	x						x							*C. albicans*	serine/threonine protein kinase STE7
P52958	2	x					x								*Haematonec. haematococca*	cutinase transcription factor 1 α CT1A
P55441	2	x												x	*Rhizobium* sp.	monooxygenase Y4FC
P78609	2	x								x					*Pichia jadinii*	uricase (urate oxidase)
Q00673	2		x				x								C. maltosa	NADH-ubiquinone oxidoreductase (CI-30.4 kd)
Q01264	2						x			x					*Pseudomonas* sp. NS672	hydantoin utilization protein C HYUC
Q16739	2			x			x								*H. sapiens*	ceramide glucosyltransferase CEGT
Q50228	2	x					x								*M. methylotrophus*	formamidase
Q64536	2	x					x								*R. norvegicus*	pyruvate dehydrogenase (lipoamide) kinase PDK2
SPAC11D3.03C	2					x	x								*S. pombe*	hypothetical protein
SPAC12G12.12	2				x		x								*S. pombe*	hypothetical protein
SPAC1327.01C	2		x										x		*S. pombe*	transcriptional regulator (Zn2-Cys6 binuclear cluster)

Table 1 (*continued*)

ORF in GPROTEOME	Number of hits	*C. tropicalis*	*D. hansenii*	*K. lactis*	*K. marxianus*	*K. thermotolerans*	*P. angusta*	*P. sorbitophila*	*S. exiguus*	*S. kluyveri*	*S. servazzii*	*S. uvarum*	*Y. lipolytica*	*Z. rouxii*	Species	Function, gene name
SPAC24C9.05C	2				x								x		*S. pombe*	hypothetical protein
SPAC27D7.03C	2						x						x		*S. pombe*	mei2, required for meiosis
SPAC2F3.16	2				x					x					*S. pombe*	hypothetical zinc finger protein
SPAC6B12.07C	2	x											x		*S. pombe*	hypothetical protein zinc finger
SPBC16H5.12C	2		x				x								*S. pombe*	hypothetical protein
SPBC4F6.09	2				x		x								*S. pombe*	major facilitator family
SPCC622.19	2						x			x					*S. pombe*	hypothetical protein
SPCC757.05C	2						x						x		*S. pombe*	putative acetylornithine deacetylase
SPSIN1	2	x											x		*S. pombe*	stress activated MAP kinase

logues of the sorbitol utilization genes *SOU1* and *SOU2* of *C. albicans* were found in seven of these nine species. Interestingly *Y. lipolytica* which grows after a 7 day lag on salicin contains a gene similar to *Saccharomycopsis fibuligera* β-D-glucosidase. Similarly, *K. lactis*, *K. marxianus* and dairy isolates of *D. hansenii* (but not the type strain nor several non-dairy isolates) are able to use lactose as a carbon source. RSTs carrying homologues of the *K. lactis LAC4* or *LAC12* gene (β-galactosidase and lactose permease, respectively) were observed in these three species. *C. tropicalis* and *Y. lipolytica* are able to utilize alkanes as carbon sources, and duplicated copies of the Cyt P450 genes needed in the first oxidation step were identified in *Y. lipolytica* (*ALK2, 3, 5* and *7* homologues). *P. angusta* is the only yeast in our set able to use nitrate and nitrite as nitrogen source, and methanol as a carbon source: homologues of genes encoding nitrate and nitrite reductase, as well as the methanol oxidase gene were identified in this species.

Taken together, these results show that several metabolic peculiarities of the yeast set considered here can be traced down to the presence of species-specific genes that were apparently absent from other species.

3.3. *Impact of speciation on gene assignment to specific functional categories*

All newly identified genes in the different species were distributed into 15 main functional classes (see Table 2), using the classification proposed by MIPS for its functional catalogue of *S. cerevisiae* ORF (see Section 2). This assumes that genes have conserved in the different species a function close to that defined for their homologues in *S. cerevisiae*. This assumption is obviously not completely true, but is likely to be generally valid as interspecific complementation of e.g. *S. cerevisiae* mutants often resulted in the isolation of functionally equivalent homologues from gene libraries of various origins [32–34].

This analysis relies on the assumption that genes are identified (and thus assigned to a functional class) independently of the general sequence divergence between the species, i.e. that among all newly identified genes, the fraction attributed in a given species to a specific functional class is independent from the taxonomic position. On the contrary, if genes of the different classes would evolve at different rates, this would lead to apparent over-representation of classes assembling conserved genes and under-representation of functional classes, where most genes tend to evolve rapidly and would therefore be missed in our search.

In order to test this possibility, we first computed the quality of the matches observed for genes predicted in each functional class. Only ORFs having a homologue in *S. cerevisiae* were considered. This analysis was based on the percentage of identity of aligned amino acid segments identified by BLASTX search on each validated RST match. When several segments of a given RST were aligned on the same ORF with different scores (as in the case of frameshift or of poor conservation of local parts of the protein), only the best score was kept. Representative results for *S. bayanus* var. *uvarum*, *K. marxianus* and *Y. lipolytica* are shown on Fig. 2. For these species, which range from very close to very distant from *S. cerevisiae*, this analysis shows that genes from all 14 main functional classes exhibit a similar distribution of identity percentages, with peaks around 90%, 60% and 40% in the

Table 2
Distribution into functional categories of the genes identified in the 14 species and percentage of ascomycete-specific genes

Functional class	MIPS class no.	*Y. lipolytica*		*C. tropicalis*		*D. hansenii*		*P. sorbitophila*		*P. angusta*		*K. marxianus*		*K. lactis*		*K. thermotolerans*		*S. kluyveri*		*Z. rouxii*		*S. servazzii*		*S. exiguus*		*S. uvarum*		*S. cerevisiae*	
		Mini	% Asco	Mini	% Asco	Mini	% Asco	Mini	% Asco	Mini	% Asco	Mini	% Asco	Mini	% Asco	Mini	% Asco	Mini	% Asco	Mini	% Asco	Mini	% Asco	Mini	% Asco	Mini	% Asco	Mini	% Asco
Metabolism	01	319	15	265	12	265	11	314	7	583	15	298	12	535	18	287	15	326	15	490	14	288	14	306	14	550	15	1044	16
Energy	02	59	7	73	10	58	12	68	6	141	14	65	18	110	18	59	7	78	13	110	12	67	10	70	4	118	15	240	18
Cell growth, cell division and DNA synthesis	03	179	21	145	12	180	11	229	9	345	14	186	16	363	22	203	23	210	19	329	20	254	21	264	21	442	25	788	28
Transcription	04	153	24	115	22	153	19	176	15	315	17	190	28	335	30	211	25	182	30	283	25	219	25	209	28	420	30	728	30
Protein synthesis	05	67	3	59	7	56	7	105	9	153	8	82	16	122	7	80	11	64	14	133	11	82	15	88	7	165	13	345	13
Protein destination	06	141	10	113	6	120	13	157	10	284	13	123	13	235	15	157	17	171	21	241	12	141	13	172	10	284	17	539	18
Transport facilitation	07	115	5	87	7	109	3	94	6	187	10	77	5	147	10	80	4	95	4	131	5	71	6	84	7	161	7	301	9
Intracellular transport	08	122	5	97	7	115	7	135	7	226	7	130	10	208	15	119	9	138	7	189	12	117	10	130	7	247	13	448	14
Cellular biogenesis	09	57	18	38	9	44	11	60	17	93	12	48	11	84	19	63	17	51	29	83	14	60	17	69	19	104	24	187	27
Cellular communication/ signal transduction	10	30	17	19	16	29	14	44	7	62	12	29	0	57	11	29	21	37	14	50	12	49	22	43	14	71	21	126	20
Cell rescue, defense, cell death and ageing	11	95	8	86	9	85	4	100	4	158	6	78	12	133	12	94	11	93	8	133	8	90	9	90	11	186	18	354	21
Ionic homeostasis	13	39	5	33	8	31	0	38	8	59	8	26	12	52	15	37	9	24	0	46	9	30	10	30	7	72	13	118	14
Cellular organization	30	536	9	449	11	517	11	641	7	1093	13	568	17	978	18	593	17	640	16	937	15	639	18	635	16	1160	19	2181	21
Classification unclear	98	30	23	25	22	45	20	32	9	74	18	33	10	61	22	45	11	26	12	54	24	33	15	36	17	80	21	148	28
Unknown function	99	299	36	225	32	313	40	269	29	657	35	384	43	671	43	480	49	419	46	642	46	411	46	420	43	978	48	2568	52
Total number of genes identified		1235		965		1165		1321		2476		1343		2283		1468		1432		2126		1415		1456		2898		6320	

Data concern minimal gene number, and have been adjusted to the nearest integer. *S. cerevisiae* data have been added for comparison. The sum of each column is larger than the total number of genes identified in each species (last line) since the same gene may be present in different functional classes.

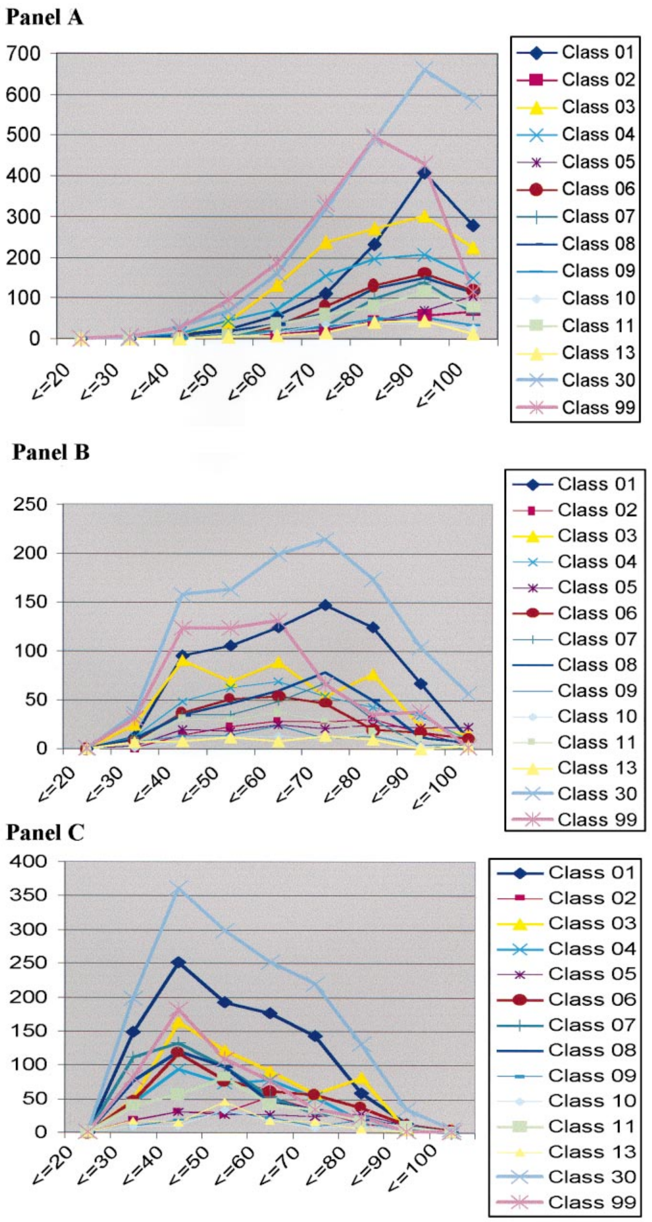

Fig. 2. Gene conservation through functional classes. The distribution of the percentage of amino acid identity observed between translated RST and proteins from *S. cerevisiae* or GPROTEOME was computed for genes assigned to the following functional classes: 01 = metabolism; 02 = energy; 03 = cell growth; 04 = transcription; 05 = protein synthesis; 06 = protein destination; 07 = transport facilitation; 08 = intracellular transport; 09 = cellular biogenesis; 10 = cellular communication; 11 = cell rescue; 13 = cell organization; 99 = unclassified proteins. Data are shown for the following species: A = *S. bayanus* var. *uvarum*; B = *K. marxianus*; C = *Y. lipolytica*.

case of *S. bayanus* var. *uvarum, K. marxianus*, *Y. lipolytica* (respectively), as expected from their phylogenetic distance from *S. cerevisiae* [19]. We noticed however that class 99 (unclassified proteins) and 03 (cell growth) tended to be less conserved than the average and to have a more scattered distribution, whereas, e.g., class 01 (metabolism) tended to be more conserved than the average.

3.4. *Ascomycete-specific genes are not equally distributed among functional classes*

Malpertuy et al. [35] proposed to divide *S. cerevisiae* genes into 'conserved' and 'ascomycete-specific', the latter evolving more rapidly (questionable ORFs have been ignored here). The fact that the mean percentage of protein identity appeared somewhat dependent on the type of associated function suggested that some functional classes might be enriched in ascomycete-specific genes (Asco-genes). In order to test this hypothesis, we classified all genes known in *S. cerevisiae* or identified in the 13 species as follows: (i) the Asco-genes without homologue outside *S. cerevisiae* but identified in one of the 13, (ii) the conserved genes identified through GPROTEOME in non-ascomycetous organisms, (iii) the Asco-genes not found in the 13 studied hemiascomycetes but found exclusively in other ascomycetous fungi. Thus, about 20% of the total number of genes identified were Asco-genes (see also [35]). They contributed differentially to the functional classes (see Table 2), over a range from 0% to 49%. Functional classes that contain more than 20% Asco-genes in *S. cerevisiae* (like transcription) were called 'Asco-class'. As shown below, these classes tend to be under-represented in species distant from *S. cerevisiae*, likely reflecting rapid evolution blurring detectable similarity with *S. cerevisiae* homologues. 'Conserved classes' like transport facilitation (7% Asco-genes) may reciprocally appear over-represented in those species.

No Asco-genes were identified for class 10 (cellular communication) in *K. marxianus*, nor for class 13 (ion homeostasis) in *D. hansenii* and *S. kluyveri* (see Table 2). Although sampling bias may explain these variations, the fact that class 13 is over-represented in both *D. hansenii* and *S. kluyveri* (see below) may also suggest that these Asco-genes have evolved beyond recognition, possibly to meet specific needs. A special situation seems to occur in *P. sorbitophila* where all classes show a deficit in Asco-genes: since a lower than expected number of genes was identified in this species, this may indicate that many Asco-genes have evolved rapidly in this species and were missed.

Functional sub-classes identified in the yeast species appear much more heterogeneous by this criterion. Some of them contain large sets of genes of the conserved type, typically more than 90% of the total, and thus appear well represented in all species. For example, 117 out of the 123 genes involved in amino acid biosynthesis were of the conserved type. The same holds true for several basic metabolic and cellular functions like ABC transporters, tRNA synthetases, lipid and fatty acid transporters, biosynthesis of vitamins/cofactors/prosthetic groups, amino acid transporters, glycolysis and gluconeogenesis, allantoin and allantoate transporters, pyrimidine-ribonucleotide metabolism, protein folding and stabilization, tricarboxylic acid pathway, drug transporters, deoxyribonucleotide metabolism, cytokinesis, purine-ribonucleotide metabolism, anion transporters, purine and pyrimidine transporters, translation, vacuolar transport, vacuolar and lysosomal biogenesis. On the contrary, a few small-sized functional classes (10–30 genes) appear to be mainly composed of Asco-genes (more than 60%), like organization of cell wall, extracellular/secreted proteins, ageing, amine metabolism, retrotransposon and plasmid proteins, and their size may thus have been underestimated in species distant from *S. cerevisiae*. The first two classes of those just mentioned have been extensively used as taxonomic markers [36], suggesting that Asco-classes strongly contributed to adaptation and diversification of yeast species. Interestingly, several sub-classes involved in transcriptional control are also enriched in ascomycete-specific genes. For example, Zn_2-Cys_6 binuclear transcriptional activators have been hitherto found exclusively in fungi [37,38]. A total of 48 putative orthologues of the 56 Zn_2-Cys_6 genes identified in *S. cerevisiae* were identified in the species analyzed (A. Goffeau, personal communication); 31 were found in *S. bayanus* var. *uvarum* (ca. 5000 RSTs) and 18 in *S. servazzii* (ca. 2500 RSTs), between 10 and 16 in all other species except *C. tropicalis, D. hansenii* (ca. 2500 RSTs each), *Y. lipolytica, P. sorbitophila* (ca. 5000 RSTs each) where only six, four, four and three putative orthologues were identified, respectively. This suggests rapid evolution of these factors in the hemiascomycetous branch. Similarly, Asco-genes were predominant among genes encoding regulators of specific pathways, such as lipid/fatty acids and sterol metabolism, amino acid metabolism, nitrogen or sulfur or carbohydrate utilization. It is

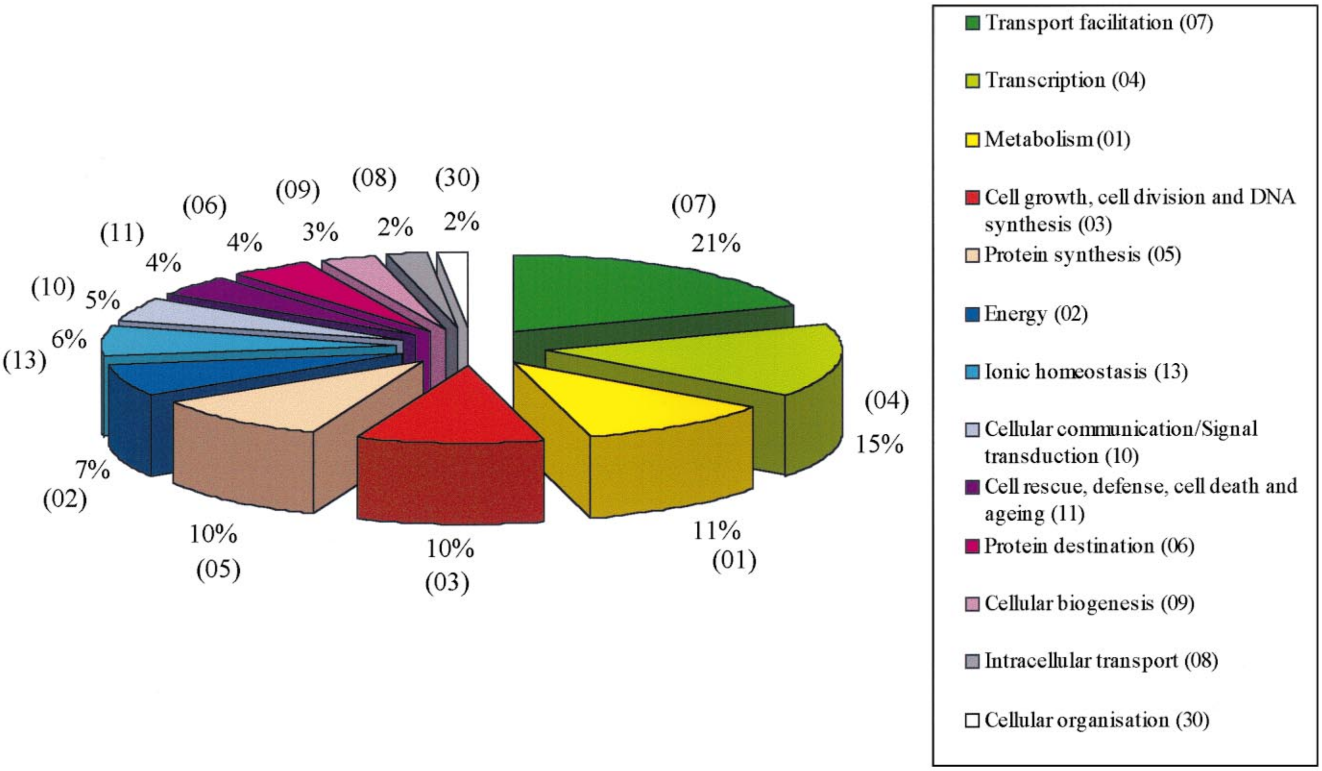

Fig. 3. Percentage of inertia distribution. The inertia of each functional class was computed. Six classes among the 13 existing ones: transport facilitation, transcription, metabolism, cell growth/cell division/DNA synthesis, protein synthesis and energy contribute mostly (74%) to the total inertia of the whole data set. The less variable classes are: cellular biogenesis, cellular organization, intracellular transport, which represent together only 7% of the inertia of the whole data set.

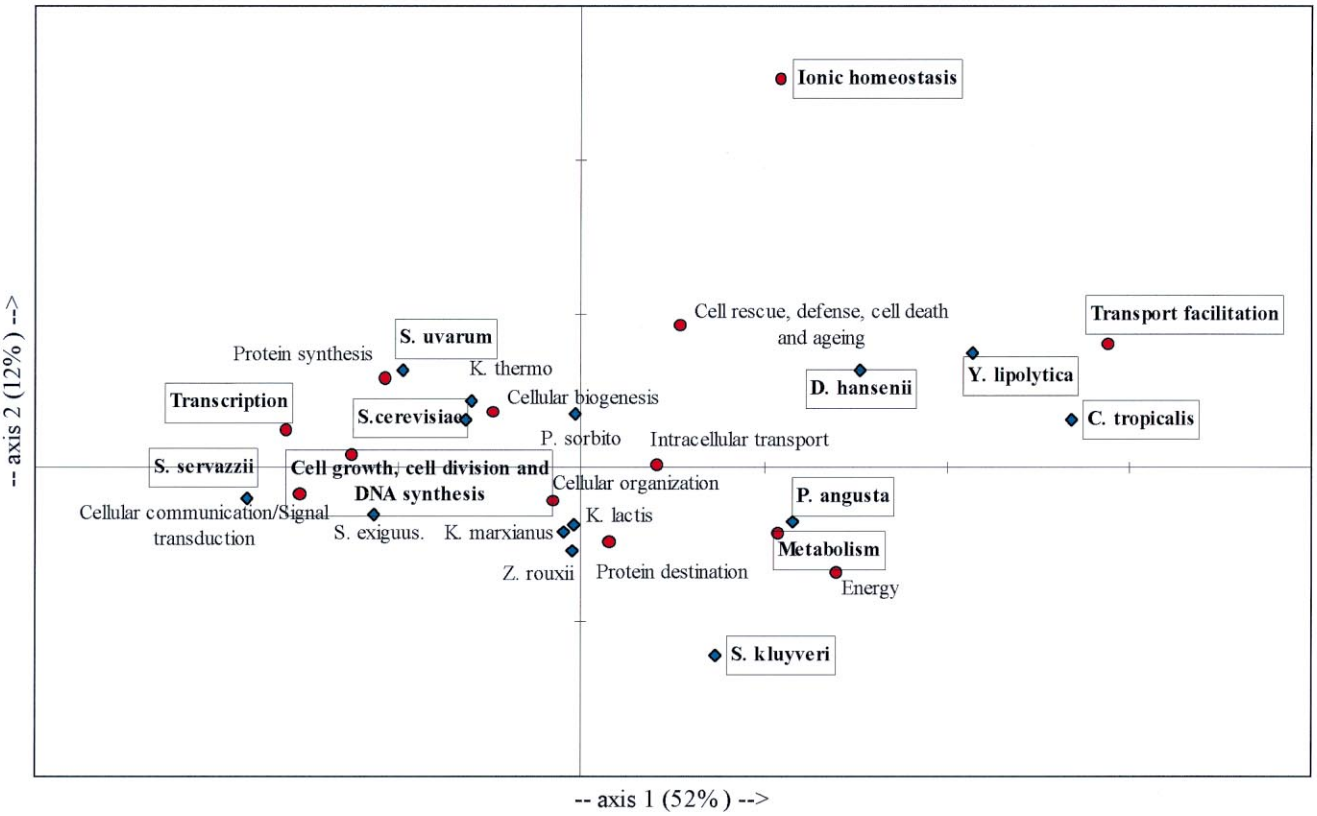

Fig. 4. Correspondence analysis of the distribution of main functional categories in the 14 species including *S. cerevisiae*. Projection on the two first axes of the data sets. The 14 yeast species appear as navy blue diamond. The 13 functional categories appear as red dots. The two first axes account for 64% of the global inertia of the data sets. The categories and yeast species which contribute significantly to the inertia of the factorial plane appear boxed and in bold characters.

tempting to speculate that rapid evolution of transcriptional activators was somehow selected during speciation. The same probably holds true for unclassified proteins (29–49% Asco-genes) which appear strongly biased against in species distant from *S. cerevisiae*.

This also indicates that gene number estimates in the main functional classes will be moderately biased by the presence of Asco-genes, but that representation of specific sub-classes might be strongly affected.

3.5. *Yeast species can be differentiated on the basis of their global functional profiles*

With the above caveats in mind, we compared the representation of the different functional classes in the 14 species including *S. cerevisiae*. We excluded classes 98 and 99 corresponding to genes of undecided and unknown function (respectively). We also checked that no significant differences were observed when analyses were based on minimal (like shown below) or on maximal gene number estimates.

A first χ^2 test was done on the entire set of data: it showed a significant difference ($P < 10^{-4}$) in the distribution of the 14 yeast species between the 13 different functional classes. This clearly showed that there was no 'average yeast functional profile'.

3.6. *Distinct major functional classes contribute to profile differences*

The former analysis does not tell which of the functional classes contribute to the observed differences. Correspondence analysis (see Section 2) was performed on Table 2 data using minimal values of gene occurrence [18,21], and excluding classes 98 and 99.

Done on the 14 different species and 13 functional classes, this analysis showed that six functional classes: transport facilitation (07), transcription (04), cell growth/cell division and DNA synthesis (03), metabolism (01), protein synthesis (05), and energy (02) account for 74% of the total inertia of the data set (see Fig. 3). On the other hand, intracellular transport, cellular organization, and cellular biogenesis contributed poorly to the global inertia of the data set. Rather than indicating that no significant differences exist between the 14 studied species for these classes, this suggests that these classes are either too large or too loosely delimited in the MIPS catalogue to be discriminant. Five species contributed predominantly to the inertia of the data set: *C. tropicalis, Y. lipolytica, S. servazzii, D. hansenii*, and *P. angusta*. The less contributing yeast species were *K. lactis* and *K. marxianus*.

Fig. 4 shows the distribution of the species and functional categories on the first factorial space which represents 64% of the total inertia (or information). Note that this is a significant proportion, since the mean contribution of a given axis to the total inertia is 100/13 = 7.7%.

Metabolism and transport facilitation on one hand, transcription and cell growth on the other, strongly differentiated the eight species *C. tropicalis, S. servazzii, Y. lipolytica, P. angusta, S. bayanus* var. *uvarum, D. hansenii, S. cerevisiae*, and *S. kluyveri* from the remaining species. Metabolism and transport facilitation were over-represented in *Y. lipolytica,*

Table 3
Over-representation of functional sub-classes

Class		G_i	g_i	Mini	Mini/g_i	% Asco	Species
02.25	β-Oxidation of fatty acids	7	4	**8**	1.90	0	*Z. rouxii*
01.02	Nitrogen and sulfur metabolism	74	14	26.6	1.93	19	*C. tropicalis*
02.01	Glycolysis and gluconeogenesis	35	7	12.6	1.93	3	*C. tropicalis*
01.02	Nitrogen and sulfur metabolism	74	17	32	1.94	19	*Y. lipolytica*
30.09	Organization of transport vesicles	42	9	17	1.95	14	*D. hansenii*
30.09	Organization of transport vesicles	42	9	13.5	1.95	14	*K. marxianus*
09.25	Vacuolar and lysosomal biogenesis	17	4.58	9	1.96	6	*S. exiguus*
30.19	Peroxisomal organization	39	17	33.9	1.96	13	*P. angusta*
30.19	Peroxisomal organization	39	8	16	1.98	13	*D. hansenii*
30.19	Peroxisomal organization	39	8	16	1.98	13	*K. marxianus*
05.07	Translational control	30	18	**36**	1.99	40	*Z. rouxii*
30.02	Organization of plasma membrane	142	29	59	2.00	13	*D. hansenii*
05.10	tRNA synthetases	37	9	19	2.06	0	*P. sorbitophila*
07.99	Other transport facilitators	56	13	26	2.08	23	*Y. lipolytica*
07.13	Lipid transporters	7	4	**9**	2.13	0	*Z. rouxii*
07.25	ABC transporters	28	7	15	2.15	0	*P. sorbitophila*
07.25	ABC transporters	28	5	11.3	2.18	0	*C. tropicalis*
07.28	Drug transporters	35	7	16	2.20	0	*D. hansenii*
07.28	Drug transporters	35	7	16	2.20	3	*K. marxianus*
09.16	Mitochondrial biogenesis	15	3	7	2.25	21	*D. hansenii*
09.16	Mitochondrial biogenesis	15	3	7	2.25	21	*K. marxianus*
09.13	Biogenesis of chromosome structure	18	4.85	11	2.27	28	*S. exiguus*
07.10	Amino acid transporters	25	5	12	2.31	0	*D. hansenii*
07.10	Amino acid transporters	25	5	12	2.31	0	*K. marxianus*
02.10	Tricarboxylic acid pathway	24	4	11.2	2.50	8	*C. tropicalis*
07.25	ABC transporters	28	6	16	2.56	0	*Y. lipolytica*
30.19	Peroxisomal organization	39	7	18.8	2.60	13	*C. tropicalis*
08.10	Peroxisomal transport	13	2	7.0	2.90	23	*C. tropicalis*
07.19	Allantoin and allantoate transporters	9	4	**14.4**	3.61	0	*P. angusta*
09.99	Other cellular biogenesis activities	3	2	**8**	4.42	0	*Z. rouxii*
02.25	β-Oxidation of fatty acids	7	1	**7**	5.38	0	*C. tropicalis*
07.19	Allantoin and allantoate transporters	9	2	**11**	5.89	0	*D. hansenii*
07.19	Allantoin and allantoate transporters	9	2	**11**	5.89	0	*K. marxianus*
09.19	Peroxisomal biogenesis	2	0	**5**	11.19	0	*Y. lipolytica*

Species are shown in the left column. G_i = size of the class in *S. cerevisiae*. g_i = expected size in the species (see Section 2), Mini = minimal of genes assigned to a functional sub-class in the species, in the different species, % Asco = percentage of ascomycete-specific genes in the functional sub-class in *S. cerevisiae*. Only cases where mini/g_i was higher than 1.9 are shown. Bold figures represent sub-classes where the absolute number of identified members is higher in a given species than in *S. cerevisiae* (mini > G_i).

C. tropicalis, P. angusta, D. hansenii, and *S. kluyveri* (see Fig. 4, right side of first axis). On the contrary, transcription and cell growth were over-represented in *S. servazzii, S. bayanus* var. *uvarum, S. cerevisiae* and *S. exiguus* (Fig. 4, left side of first axis).

Y. lipolytica and *C. tropicalis* were found to exhibit similar functional profiles, as did *S. servazzii* and *S. exiguus*, but both profiles were clearly different. Interestingly, *S. kluyveri* appeared rather different from the other *Saccharomyces* species: metabolism and transport facilitation classes were over-represented in *S. kluyveri* while they were under-represented in the other *Saccharomyces* species. Conversely, genes assigned to transcription and cell growth appeared under-represented in *S. kluyveri*.

At a first glance, this analysis indicates that representation of genes dedicated to specific major functions permits clear differentiation of most species. It should be stressed however that this representation is conspicuously biased by phylogenetic distances and gene conservation in the different functions. Species close to *S. cerevisiae* appear on the left and are associated with functional classes containing more than 20% Asco-genes ('Asco-classes'), which are likely best identified in these species. More distant species appear on the right and are associated with 'conserved classes' (less than 20% Asco-genes in *S. cerevisiae*). The only clear exception concerns the class protein synthesis.

3.7. Species classification according to their neighborhood in the factorial space

Species and functions were classified according to their neighborhood in the factorial space obtained by correspond-

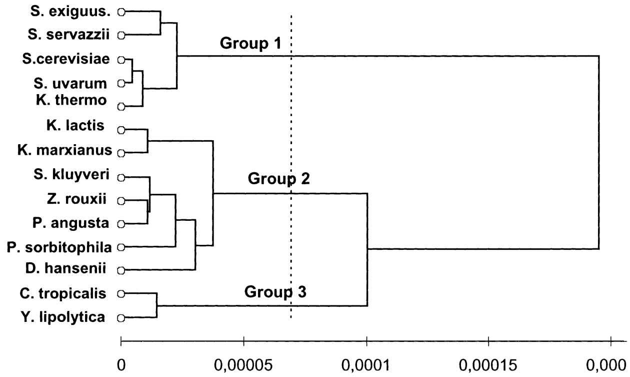

Fig. 5. Cluster analysis of the 14 studied species according to the distribution of their genes among the 13 functional classes. This tree shows the species classification according to their neighborhood in the factorial space obtained by correspondence analysis of data from Table 2 excluding classes 98 and 99. This classification takes into account the whole information of Table 2. The software Xlstat was used to perform an ascendant hierarchical cluster analysis.

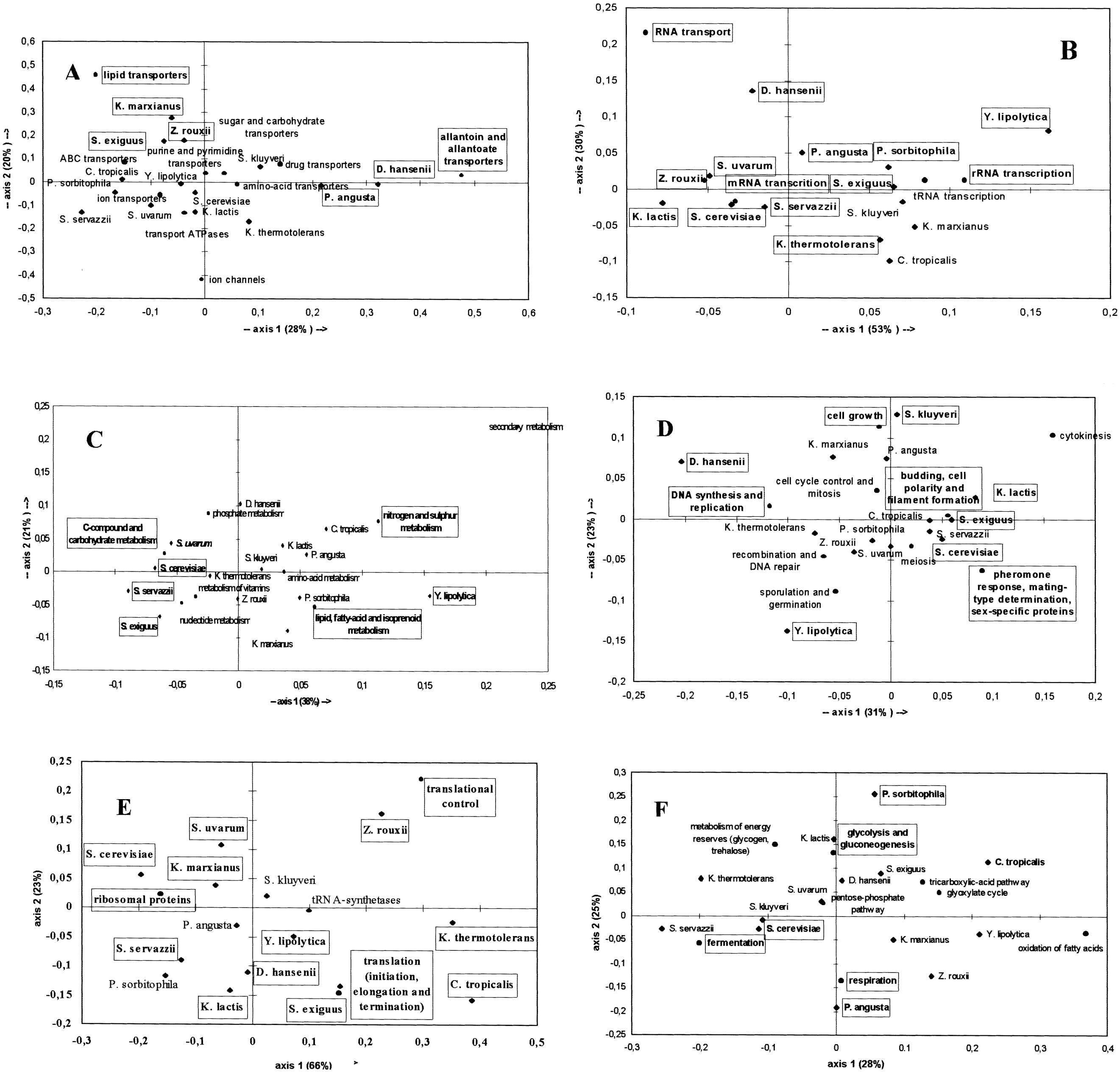

Fig. 6. Correspondence analysis of the distribution of genes from the main functional classes into the sub-classes in the 14 species. Projection on the two first axes of the data sets. The 14 yeast species appear as diamond. The sub-classes appear as dots. The sub-classes and yeast species contributing significantly to the two first axes appear in a box and in bold font. A: class 07 (transport facilitation); B: class 04 (transcription); C: class 01 (metabolism); D: class 03 (cell growth, cell division and DNA synthesis); E: class 05 (protein synthesis); F: class 02 (energy).

ence analysis taking into account the whole inertia of the data set. The dendrogram obtained for species classification is shown on Fig. 5. Species group 1 assembles all *Saccharomyces* species except *S. kluyveri* together with *K. thermotolerans*; species group 3 is made of *C. tropicalis* and *Y. lipolytica*. The rest of the species constitutes species group 2. This tree differs only slightly from the phylogenetic tree based on rRNA sequences [19], see *K. thermotolerans* and *Z. rouxii* for instance, further confirming that functional representation is strongly affected by phylogenetic relatedness. A slightly different tree was obtained when *S. cerevisiae* data were excluded from the analysis, with *P. angusta* moved to species group 3 (not shown, but see http://www-alt.pasteur.fr/~tekaia/HYG/mfctfcah.html).

Since by construction, members of the same species group have similar functional profiles, intersection between species and function groups permits characterization of each species group by its most relevant functional categories. This analysis led to the following relationships: species group I: cell organization, cell communication, cell rescue; species group II: energy, metabolism, cell rescue; species group III: metabolism, cell growth, transcription, cell organization.

As stated before, this classification may be blurred by gene identification biases. It should be noticed however that it basically relies on principles (functional clustering) similar to those used in numerical taxonomy [39] for microbial classification before the advent of molecular methods, and which resulted in a classification partially compatible with rDNA phylogeny.

3.8. *Functional sub-classes contribute to the variation of the main classes representation*

We then evaluated gene distributions into functional sub-classes which are smaller than main classes and thus may represent more sensitive indicators of quantitative variations in function representation. All identified genes were assigned to one of 204 functional sub-classes (see Section 2, not shown). Clear cases of expansion of functional sub-classes can be identified when absolute numbers of genes assigned to a sub-class are larger in one species than in *S. cerevisiae* (see classes 07.19, 07.13, 02.25, 05.07, 09.19 and 09.99 in Table 3), and others can be suspected when the expected numbers of genes in the functional sub-classes are compared to the observed ones (see Section 2 and Table 3 for a partial list). Amino acid, allantoin and allantoate, ABC and drug transporters often appeared likely to be over-represented, as well as genes of peroxisomal organization in four species (*D. hansenii, K. marxianus, P. angusta* and *C. tropicalis*). Most of the over-represented sub-classes contain around or less than 20% Asco-genes in *S. cerevisiae*, the only exceptions being found in species close to *S. cerevisiae* (*Z. rouxii, S. exiguus*), and amplification affected almost exclusively conserved genes: this strongly suggests that limited gene amplification occurred in specific sub-classes.

For a more comprehensive analysis, we resorted to correspondence analysis. A few species could be clearly singled out from all others through a specific trait (see Fig. 6): over-representation of allantoin and allantoate transporters appeared associated with *D. hansenii* and *P. angusta* (Fig. 6A), lipid and fatty acid metabolism (Fig. 6C), translational control with *Z. rouxii* (Fig. 6E), respiration with *P. angusta* which may parallel the development of a strong oxidative metabolism in this species for growth on C1 compounds (Fig. 6F).

Other associations were more difficult to interpret since biases in function representation could not be excluded. We considered that the likelihood of variation of sub-class representation was highest in cases of: (i) over-representation of Asco sub-classes (> 20% Asco-genes) or under-representation of conserved sub-classes (< 20% Asco-genes) in species distant from *S. cerevisiae*, (ii) over-representation of conserved sub-classes or under-representation of Asco sub-classes in species close to *S. cerevisiae*.

For instance, no clear conclusion could be derived in the case of transcription sub-classes since these contain 20–30% Asco-genes in *S. cerevisiae* and the observed ranking of species from left to right may reflect phylogenetic distance from *S. cerevisiae* (Fig. 6B). This was also the case for sub-classes of cell growth/cell division/DNA synthesis (Fig. 6D) which contain 21–35% Asco-genes, except for the sub-class DNA synthesis which appears over-represented in *D. hansenii*

The sub-classes of metabolism (Fig. 6C) contain 11–21% Asco-genes in *S. cerevisiae*. The over-representation of sub-class C compound and carbohydrate metabolism in *S. servazzii*, *S. cerevisiae*, *S. exiguus* and *S. bayanus* var. *uvarum* may reflect their feeding preferences. Those species are opposed to *Y. lipolytica*, where this sub-class is under-represented, possi-

Table 4
Number of genes which are present in the main functional classes and are part of gene families over-amplified when compared to *S. cerevisiae*

Class no.	*K. thermotolerans*	*S. kluyveri*	*K. marxianus*	*P. sorbitophila*	*S. servazzii*	*C. tropicalis*	*S. uvarum*	*K. lactis*	*D. hansenii*	*Y. lipolytica*	*S. exiguus*	*P. angusta*	*Z. rouxii*	Sum of genes over-amplified in the different species	Class
98	0	0	0	0	0	0	0	0	0	0	2	4	0	6	Not clear cut
10	0	0	0	0	0	0	0	0	6	0	0	0	4	10	Signal transduction
05	0	0	0	0	0	0	2	0	0	0	2	0	**12**	16	Protein synthesis
09	0	0	0	0	0	0	0	0	4	3	4	2	8	21	Cellular biogenesis
13	0	0	2	0	0	3	0	2	2	7	0	7	4	27	Ionic homeostasis
08	0	0	2	4	0	0	0	2	4	5	4	7	8	36	Intracellular transport
02	0	0	0	2	3	2	2	**12**	4	4	7	6	7	49	Energy
06	0	0	2	0	0	0	4	7	0	2	9	5	**20**	49	Protein destination
11	0	0	0	4	2	6	4	2	4	5	2	**17**	8	54	Cell rescue
04	0	0	0	0	4	0	**12**	4	4	0	8	**11**	**16**	59	Transcription
03	0	0	3	2	5	2	**12**	0	8	7	**21**	**11**	**16**	87	Cell growth
07	0	0	6	2	0	9	0	7	**16**	**15**	8	**37**	4	103	Transport facilitation
99	0	6	4	**10**	**12**	2	**11**	8	**14**	8	**34**	**43**	**55**	207	Unclassified proteins
01	0	2	6	4	4	**9**	**6**	**33**	**10**	**29**	**14**	**58**	**38**	214	Metabolism
30	0	2	2	**10**	9	**11**	**16**	**19**	**22**	**26**	**24**	**60**	**66**	267	Cellular organization
Total	0	10	26	38	39	42	69	75	94	101	139	229	266		

Data are based on minimal gene numbers. Species are ranked by increasing number of total number of over-amplified genes (last row). Functional classes are ranked by increasing number of observed amplified genes in all species (penultimate column).

Table 5
Number of genes which are present in functional sub-classes and are part of amplified gene families in the 13 species

Sub-classes	*S. kluyveri*	*K. marxianus*	*P. sorbitophila*	*S. servazzii*	*C. tropicalis*	*S. uvarum*	*D. hansenii*	*Y. lipolytica*	*K. lactis*	*S. exiguus*	*Z. rouxii*	*P. angusta*	Sum in the 13 species	*S. cerevisiae*	
01.04	0	0	0	0	0	0	0	0	0	0	0	2	2	31	phosphate metabolism
07.13	0	0	0	0	0	0	0	0	0	0	0	2	2	7	lipid transporters
02.13	0	0	0	0	0	0	0	0	0	0	0	2	2	79	respiration
03.25	0	0	0	0	0	0	2	0	0	0	0	0	2	30	cytokinesis
04.99	0	0	0	0	0	0	0	0	0	0	2	0	2	57	other transcription activities
05.99	0	0	0	0	0	2	0	0	0	0	0	0	2	14	other protein synthesis activities
07.22	0	2	0	0	0	0	0	0	0	0	0	0	2	41	transport ATPases
07.25	0	0	2	0	0	0	0	0	0	0	0	0	2	28	ABC transporters
09.10	0	0	0	0	0	0	0	0	0	2	0	0	2	9	nuclear biogenesis
09.13	0	0	0	0	0	0	0	0	2	0	0	2	18		biogenesis of chromosome structure
09.16	0	0	0	0	0	0	2	0	0	0	0	0	2	14	mitochondrial biogenesis
10.03	0	0	0	0	0	0	2	0	0	0	0	0	2	17	osmosensing
10.04	0	0	0	0	0	0	2	0	0	0	0	0	2	24	nutritional response pathway
10.05	0	0	0	0	0	0	2	0	0	0	0	0	2	35	pheromone response generation
11.99	0	0	2	0	0	0	0	0	0	0	0	0	2	10	other cell rescue activities
13.01	0	0	0	0	0	0	0	2	0	0	0	0	2	56	homeostasis of metal ions
30.05	0	0	0	0	0	2	0	0	0	0	0	0	2	29	organization of centrosome
30.25	0	2	0	0	0	0	0	0	2	0	0	0	4	54	vacuolar and lysosomal organization
01.20	0	0	0	0	0	0	2	0	0	0	0	2	4	4	secondary metabolism
02.16	0	0	0	0	0	0	0	2	0	2	0	0	4	34	fermentation
03.13	0	0	0	0	0	0	0	0	0	2	2	0	4	99	meiosis
04.07	0	0	0	2	0	0	0	0	0	0	2	0	4	27	RNA transport
05.07	0	0	0	0	0	0	0	0	0	0	4	0	4	30	translational control
08.01	0	0	2	0	0	0	0	0	2	0	0	0	4	48	nuclear transport
08.13	0	2	0	0	0	0	0	0	0	0	2	0	4	51	vacuolar transport
30.08	0	0	0	0	0	0	0	0	4	0	0	4	66		organization of Golgi
30.13	0	0	0	0	0	2	0	0	0	2	0	0	4	38	organization of chromosome structure
03.19	0	0	0	0	0	2	0	3	0	0	0	0	5	89	recombination and DNA repair
04.03	0	0	0	0	0	0	0	0	0	2	2	1	5	81	tRNA transcription
03.99	0	0	0	0	0	0	0	0	0	4	0	2	6	13	cell division and DNA synthesis
03.10	0	0	0	0	0	2	0	0	0	0	4	0	6	106	sporulation and germination
10.99	0	0	0	0	0	0	0	2	0	0	4	0	6	36	other signal transduction activities
30.09	0	0	0	0	0	0	0	0	0	0	4	2	6	42	organization of transport vesicles
98	0	0	0	0	0	0	0	0	0	2	0	4	6	148	classification not clear cut
04.01	0	0	0	0	0	2	0	0	0	0	2	3	7	100	rRNA transcription
02.19	0	0	0	0	0	0	2	0	5	0	0	0	7	37	metabolism of energy reserves
08.99	0	0	0	0	0	0	2	3	0	2	0	0	7	31	other intracellular transport activities
02.10	0	0	0	0	0	0	0	2	1	2	2	0	7	24	tricarboxylic acid pathway
02.25	0	0	0	0	2	0	2	0	0	0	0	4	8	7	oxidation of fatty acids
01.07	0	0	0	2	0	0	2	0	3	2	0	0	9	79	metabolism of vitamins and cofactors
06.07	0	0	0	0	0	0	0	2	0	5	2	0	9	167	protein modification
30.04	0	0	0	3	0	2	0	0	0	4	0	0	9	97	organization of cytoskeleton
02.99	0	0	0	3	0	0	0	0	1	3	2	0	9	15	other energy generation activities
03.01	0	0	0	0	0	0	2	0	0	0	2	6	10	75	cell growth
02.01	0	0	2	0	0	2	0	0	3	0	3	0	10	35	glycolysis and gluconeogenesis
05.01	0	0	0	0	0	0	0	0	0	2	8	0	10	205	ribosomal proteins
06.04	0	0	0	0	0	0	0	0	2	2	6	0	10	128	protein targeting, sorting, translocation
08.07	0	0	2	0	0	0	0	0	0	2	4	2	10	106	vesicular transport (Golgi network, etc.)
11.04	0	0	2	2	0	0	0	3	0	0	2	1	10	83	DNA repair
07.28	0	0	0	0	2	0	2	0	0	0	0	7	11	35	drug transporters
08.19	0	0	0	0	0	0	2	2	0	0	2	5	11	102	cellular import
30.01	0	0	0	0	3	0	0	0	5	0	6	0	14	30	organization of cell wall
30.90	2	0	0	0	3	0	2	5	0	2	0	0	14	20	extracellular/secretion proteins
03.04	0	0	0	0	0	0	0	4	0	7	2	2	15	170	budding, polarity, filament formation
09.01	0	0	0	0	0	0	2	3	0	0	8	2	15	102	biogenesis of plasma membrane, cell wall
11.01	0	0	0	0	0	2	2	2	0	2	2	5	15	165	stress response
30.07	0	0	2	2	0	0	0	2	0	2	6	2	16	154	organization of ER
03.07	0	2, 5	2	2	2	2	2	0	0	2	2	0	17	159	pheromone, mating type and sex-specific
03.16	0	0	0	2	0	6	2	0	0	0	4	3	17	86	DNA synthesis and replication
30.19	0	0	0	0	2	0	4	0	1	0	2	8	17	39	peroxisomal organization
07.07	0	2	0	0	2	0	0	0	2	4	0	8	18	44	sugar and carbohydrate transporters
01.03	0	0	0	0	0	0	0	3	0	2	8	5	18	139	nucleotide metabolism
06.10	0	2	0	0	0	4	0	0	2	2	8	0	18	92	assembly of protein complexes

Table 5 (*continued*)

Sub-classes	*S. kluyveri*	*K. marxianus*	*P. sorbitophila*	*S. servazzii*	*C. tropicalis*	*S. uvarum*	*D. hansenii*	*Y. lipolytica*	*K. lactis*	*S. exiguus*	*Z. rouxii*	*P. angusta*	Sum in the 13 species	*S. cerevisiae*	
07.19	0	0	0	0	0	0	7	0	0	0	0	**13**	20	9	allantoin and allantoate transporters
13.04	0	2	0	0	3	0	2	5	2	0	4	7	25	63	homeostasis of other ions
06.13	0	0	0	0	0	0	0	0	**10**	0	**10**	5	25	145	proteolysis
07.04	0	2	0	0	3	0	4	7	2	0	4	4	26	74	ion transporters
11.07	0	0	0	0	6	4	2	0	0	0	4	**11**	27	101	detoxification
07.99	0	1, 5	0	0	2	0	5	8	3	4	0	4	28	56	other transport facilitators
03.22	0	0	0	3	0	2	0	0	0	**12**	8	5	30	331	cell cycle control and mitosis
01.02	0	0	0	0	5	0	0	3	5	0	6	**16**	35	75	nitrogen and sulfur metabolism
04.05	0	0	0	2	0	**10**	4	2	4	6	8	8	44	522	mRNA transcription
30.03	0	0	4	0	2	2	0	2	4	4	**18**	**12**	48	557	organization of cytoplasm
01.06	0	0	0	2	2	0	0	**11**	8	2	**13**	**11**	48	205	lipid, fatty acid metabolism
30.16	0	0	0	0	0	4	4	8	3	4	**18**	7	49	344	mitochondrial organization
30.02	0	0	0	0	5	0	**10**	**10**	2	0	2	**21**	50	143	organization of plasma membrane
01.01	0	6	2	0	2	2	0	2	**10**	2	6	**21**	53	205	amino acid metabolism
30.10	0	0	4	4	0	8	2	5	4	8	**12**	**11**	58	756	nuclear organization
01.05	2	2	2	0	2	4	6	**11**	**19**	4	**12**	**15**	79	411	carbohydrate metabolism
99	6	4	**10**	**12**	2	**11**	**14**	8	8	**34**	**55**	**45**	209	2182	unclassified proteins
Total	10	30	38	41	50	77	100	122	124	149	289	295			

Numbers of genes in families over-amplified in each species are compared to the total number of genes found in a given sub-class in *S. cerevisiae*. Species are ranked by increasing total number of over-amplified genes (last row). Functional sub-classes are ranked by increasing number of observed amplified genes in all species. No case of gene amplification was observed in *K. thermotolerans*

bly reflecting the restricted range of sugar compounds this species is able to assimilate.

In the case of sub-classes linked to protein synthesis (Fig. 6E), the ribosomal protein sub-class (14% Asco-genes) appeared over-represented in *S. cerevisiae* and *S. servazzii*, under-represented in *K. thermotolerans, Z. rouxii, C. tropicalis* and to a lesser extent in *Y. lipolytica* and *S. exiguus*. No conclusion could be reached for translational control (40% Asco-genes), nor for translation initiation/elongation/termination (2% Asco-genes) or tRNA synthetases (0% Asco-genes).

Among the sub-classes of energy (Fig. 6F), the highly conserved sub-class fermentation (6% Asco-genes) appears over-represented in *S. cerevisiae* (but surprisingly not in *S. bayanus* var. *uvarum*) and under-represented in *Y. lipolytica* which may correlate the fermentation-negative phenotype of this species.

For intracellular transport, peroxisomal transport (23% Asco-genes) is over-represented in *C. tropicalis* (see also Table 3).

For cellular communication, osmosensing (12% Asco-genes) appears under-represented in *S. cerevisiae* and possibly over-represented in *K. marxianus*, *P. angusta* and *D. hansenii*.

For cell rescue, the detoxification sub-class (15% Asco-genes) may be over-represented in *P. angusta, Y. lipolytica* and *D. hansenii*, and under-represented in *K. lactis* and *P. sorbitophila*.

Concerning cellular organization, organization of plasma membrane (13% Asco-genes) and peroxisomal organization (13% Asco-genes) may respectively be over-represented in *D. hansenii* and in *C. tropicalis*, in agreement with data shown in Table 3.

3.9. *Amplified genes and functions*

Gene amplification may permit emergence of new, species-specific functions [3]. We thus tried to evaluate the impact of genetic redundancy [18] on function representation in the individual species. Genes identified in the different species through BLASTX matches to *S. cerevisiae* ORFs are distributed in *S. cerevisiae* into 722 paralogous gene families of *n* size, *n* varying between 1 (singleton) and 108 [40]. We considered that over-amplification of *S. cerevisiae* gene families occurred when we observed an absolute number of paralogues in a given family which was larger than its corresponding *n* value in *S. cerevisiae*. All paralogues identified in a given species through GPROTEOME were counted as amplifications. This approach underestimates the number of functions where genes are actually amplified, and introduces a bias between species where 2500 or 5000 RSTs were analyzed. It permits however comparing absolute gene numbers in the 13 species to gene family sizes known in *S. cerevisiae*. It thus allows analyzing small-sized sub-classes of the functional catalogue that may better discriminate species in terms of functions.

Using these criteria, several main classes of functions are affected by gene amplifications in all species but *K. thermotolerans* (see Table 4). More than 10 genes are amplified in cellular organization, metabolism, transport facilitation, transcription, and cell Growth in 3–9 species. The last four classes were previously shown to differentiate functionally the 13 species (see above). Interestingly, 10–58 genes encoding proteins without known functions were found to be over-amplified in seven species.

Analysis of functional sub-classes evidenced amplification in the same species (see Table 5). Allantoin and allantoate transporters were amplified in *P. angusta* and *D. hansenii* (13 and seven genes, respectively), in agreement with previous correspondence analysis results. Genes involved in detoxification were found to be amplified in *P. angusta* and *C. tropicalis*, which correlates with correspondence analysis in the case of *P. angusta* only.

Other examples of observed amplifications do not coincide with correspondence analysis results. Protein synthesis and protein destination are amplified in the case of *Z. rouxii*, energy in the case of *K. lactis*, cell rescue in the case of *P. angusta*, with 12–20 genes over-amplified as compared to *S. cerevisiae*. Other cases of amplification concern proteolysis in *Z. rouxii* and *K. lactis*, and *C. tropicalis*, nitrogen metabolism in *P. angusta*, lipid and fatty acid metabolism in *Y. lipolytica*, *Z. rouxii*, and *P. angusta*, but surprisingly not in *C. tropicalis*, possibly reflecting the limited sample of RSTs analyzed in this species. *S. exiguus* exhibited an expansion of 10 genes involved in cell cycle control and mitosis. Ten or more genes involved in the metabolism of carbohydrates were found to be amplified in *Z. rouxii*, *K. lactis, Y. lipolytica* and *P. angusta*, whereas no function appeared affected by family expansions larger than four in *S. kluyveri, S. servazzii, K. marxianus* or *P. sorbitophila*. On the other hand, *S. bayanus* var. *uvarum* which is close to *S. cerevisiae* showed amplification of gene families involved in mRNA transcription.

Most of the expansions appear however limited and scattered through several functional sub-classes: this suggests that yeast biodiversity has not been primarily achieved by amplification of genes involved in specific functions. Together with data reported by Llorente et al. [18], evidencing overall conservation of the internal structure of gene families across hemiascomycetous yeasts, these observations strongly suggest that species differentiation involved a detectable, but quantitatively limited reshaping of a mostly conserved genetic repertoire.

4. Conclusion

Biological diversity of the 13 yeast species examined within the Génolevures project was reflected at the genomic level by: (i) presence of a few species-specific genes that were absent from *S. cerevisiae*, (ii) non-random distribution of rapidly evolving, ascomycete-specific, genes among functional sub-classes, (iii) in a limited number of cases, amplification of genes defining specific functional sub-classes. These results suggest that one of the main determinants of species differentiation among hemiascomycetous yeasts was gene sequence drift. Changes in the structure of both structural, enzymatic and regulatory proteins may thus have been favored to permit emergence of species-specific regulatory pathways.

Acknowledgements: This work was supported in part by the BRG (Bureau des Ressources Génétiques) (Grant no. 11-0926-99) and by the EC Grant QLRI CT-1999-01333 to one of us (E.B.).

References

[1] Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L. and Dodson, R.J. et al. (1999) Nature 399, 323–329.

[2] Marinoni, G., Manuel, M., Petersen, R.F., Hvidtfeldt, J., Sulo, P. and Piskur, J. (1999) J. Bacteriol. 181, 6488–6496.

[3] Ohno, S. (1970) Allen, G. and Unwin (Eds.), London.

[4] Seoighe, C. and Wolfe, K.H. (1999) Curr. Opin. Microbiol. 2, 548–554.

[5] Bon, E., Neuvéglise, C., Casaregola, S., Artiguenave, F., Wincker, P. et al. (2000) FEBS Lett. 487, 37–41 (this issue).

[6] Bon, E., Neuvéglise, C., Lépingle, A., Wincker, P., Artiguenave, F. et al. (2000) FEBS Lett. 487, 42–46 (this issue).

[7] Casaregola, S., Lépingle, A., Neuvéglise, C., Bon, E., Nguyen, H.V. et al. (2000) FEBS Lett. 487, 47–51 (this issue).

[8] de Montigny, J., Straub, M.L., Potier, S., Tekaia, F., Dujon, B. et al. (2000) FEBS Lett. 487, 52–55 (this issue).

[9] Neuvéglise, C., Bon, E., Lépingle, A., Wincker, P., Artiguenave, F. et al. (2000) FEBS Lett. 487, 56–60 (this issue).

[10] Malpertuy, A., Llorente, B., Blandin, G., Artiguenave, F., Wincker, P. and Dujon, B. (2000) FEBS Lett. 487, 61–65 (this issue).

[11] Bolotin-Fukuhara, M., Toffano-Nioche, C., Artiguenau, F., Duchateau-Nguyen, G., Lemaire, M., et al. (2000) FEBS Lett. 487, 66–70 (this issue).

[12] Llorente, B., Malpertuy, A., Blandin, G., Wincker, P., Artiguenave, F. and Dujon, B. (2000) FEBS Lett. 487, 71–75 (this issue).

[13] Blandin, G., Llorente, B., Malpertuy, A., Wincker, P., Artiguenave, F. and Dujon, B. (2000) FEBS Lett. 487, 76–81 (this issue).

[14] Lépingle, A., Casaregola, S., Neuvéglise, C., Bon, E., Nguyen, V.H. et al. (2000) FEBS Lett. 487, 82–86 (this issue).

[15] de Montigny, J., Spehner, C., Souciet, J.L., Tekaia, F., Dujon, B. et al. (2000) FEBS Lett. 487, 87–90 (this issue).

[16] Blandin, G., Ozier-Kalogeropoulos, O., Wincker, P., Artiguenave, F. and Dujon, B. (2000) FEBS Lett. 487, 91–94 (this issue).

[17] Casaregola, S., Neuvéglise, C., Lépingle, A., Bon, E., Feynerol, C. et al. (2000) FEBS Lett. 487, 95–100 (this issue).

[18] Llorente, B., Durrens, P., Malpertuy, A., Aigle, M., Artiguenave, F. et al. (2000) FEBS Lett. 487, 101–112 (this issue).

[19] Souciet, J.L., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M. et al. (2000) FEBS Lett. 487, 3–12 (this issue).

[20] Mewes, H.W., Albermann, K., Bähr, M., Gleissner, G., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, G. and Zollner, A. (1997) Nature 387, 7–8.

[21] Tekaia, F., Blandin, G., Malpertuy, A., Llorente, B., Durrens, P. et al. (2000) FEBS Lett. 487, 17–30 (this issue).

[22] Benzécri, J.-P. (1973) L'analyse des données, Vol. 2, l'analyse des correspondances, Dunod, Paris.

[23] Greenacre, M. (1984) Theory and Application of Correspondence Analysis, Academic Press, London.

[24] Naumova, E.S., Turakainen, H., Naumov, G.I. and Korhola, M. (1996) Mol. Gen. Genet. 253, 111–117.

[25] Piddington, C.S., Kovacevich, B.R. and Rambosek, J. (1995) Appl. Environ. Microbiol. 61, 468–475.

[26] Watabe, K., Ishikawa, T., Mukohara, Y. and Nakamura, H. (1992) J. Bacteriol. 174, 7989–7995.

[27] Watabe, K., Ishikawa, T., Mukohara, Y. and Nakamura, H. (1992) J. Bacteriol. 174, 3461–3466.

[28] Ishikawa, T., Watabe, K., Mukohara, Y. and Nakamura, H. (1997) Biosci. Biotechnol. Biochem. 61, 185–187.

[29] May, O., Nguyen, P.T. and Arnold, F.H. (2000) Nat. Biotechnol. 18, 317–320.

[30] Gojkovic, Z., Jahnke, K., Schnackerz, K.D. and Piskur, J. (2000) J. Mol. Biol. 295, 1073–1087.

[31] Nosek, J. and Fukuhara, H. (1994) J. Bacteriol. 176, 5622–5630.

[32] Kataoka, T., Powers, S., Cameron, S., Fasano, O., Goldfarb, M., Broach, J. and Wigler, M. (1985) Cell 40, 19–26.

[33] D'Enfert, C., Gensse, M. and Gaillardin, C. (1992) EMBO J. 11, 4205–4211.

[34] Tang, Z., Kuo, T., Shen, J. and Lin, R.J. (2000) Mol. Cell. Biol. 20, 816–824.

[35] Malpertuy, A., Tekaia, F., Casaregola, S., Aigle, M., Artiguenave, F. et al. (2000) FEBS Lett. 487, 113–121 (this issue).

[36] Phaff, H.J. (1998) Chemotaxonomy based on the polysaccharide composition of the cell walls and capsules, in: The Yeasts, a Taxonomic Study (Kurtzman, C.P. and Fell, J.W. (Eds.), pp. 45–47, Elsevier, Amsterdam.

[37] Marmorstein, R., Carey, M., Ptashne, M. and Harrison, S.C. (1992) Nature 356, 408–414.

[38] Marmorstein, R. and Harrison, S.C. (1994) Genes Dev. 8, 2504–2512.

[39] Sneath, P.H.A. (1995) Syst. Biol. 44, 281–298.

[40] Blandin, G., Durrens, P., Tekaia, F., Aigle, M., Bolotin-Fukuhara, M. et al. (2000) FEBS Lett. 487, 31–36 (this issue).