Genomic Exploration of the Hemiascomycetous Yeasts: 19. Ascomycetes-specific genes

Alain Malpertuy[a,1], Fredj Tekaia[a,1], Serge Casarégola[b], Michel Aigle[c], Francois Artiguenave[d], Gaëlle Blandin[a], Monique Bolotin-Fukuhara[e], Elisabeth Bon[b], Philippe Brottier[d], Jacky de Montigny[f], Pascal Durrens[c], Claude Gaillardin[b], Andrée Lépingle[b], Bertrand Llorente[a], Cécile Neuvéglise[b], Odile Ozier-Kalogeropoulos[a], Serge Potier[f], William Saurin[d], Claire Toffano-Nioche[e], Micheline Wésolowski-Louvel[g], Patrick Wincker[d], Jean Weissenbach[d], Jean-Luc Souciet[f], Bernard Dujon[a,*]

[a] *Unité de Génétique Moléculaire des Levures, URA 2171 CNRS and UFR927 Université P. et M. Curie, Institut Pasteur, 25 Rue du Dr Roux, F-75724 Paris Cedex 15, France*
[b] *Collection de Levures d'Intéret Biotechnologique, Laboratoire de Génétique Moléculaire et Cellulaire, INRA UMR216, CNRS URA1925, INA-PG, BP01, F-78850 Thiverval-Grignon, France*
[c] *Laboratoire de Biologie cellulaire de la Levure, IBGC, 1 rue Camille Saint-Saens, F-33077 Bordeaux Cedex, France*
[d] *Génoscope, Centre National de Séquenage, 2 rue Gaston Crémieux, BP191, Evry Cedex, France*
[e] *Institut de Génétique Moléculaire, CNRS/UPS UMR 8621, Batiment 400, Université de Paris Sud, F-91405 Orsay, France*
[f] *Laboratoire de Génétique et Microbiologie, UPRES-A 7010 ULP/CNRS, Institut de Botanique, 28 rue Goethe, F-67000 Strasbourg Cedex, France*
[g] *Microbiologie et Génétique, CNRS/UCB/INSA ERS 2009, Bat. 405 R2, Université Lyon I, F-69622 Villeurbanne Cedex, France*

Received 9 November 2000; accepted 11 November 2000

First published online 30 November 2000

Edited by Horst Feldmann

Abstract Comparisons of the 6213 predicted *Saccharomyces cerevisiae* open reading frame (ORF) products with sequences from organisms of other biological phyla differentiate genes commonly conserved in evolution from 'maverick' genes which have no homologue in phyla other than the Ascomycetes. We show that a majority of the 'maverick' genes have homologues among other yeast species and thus define a set of 1892 genes that, from sequence comparisons, appear 'Ascomycetes-specific'. We estimate, retrospectively, that the *S. cerevisiae* genome contains 5651 actual protein-coding genes, 50 of which were identified for the first time in this work, and that the present public databases contain 612 predicted ORFs that are not real genes. Interestingly, the sequences of the 'Ascomycetes-specific' genes tend to diverge more rapidly in evolution than that of other genes. Half of the 'Ascomycetes-specific' genes are functionally characterized in *S. cerevisiae*, and a few functional categories are over-represented in them.

Key words: Orphan; Maverick; Questionable open reading frame; Novel gene; Gene number

1. Introduction

It is now a common observation, with every novel organism sequenced, that a sizeable proportion of the predicted proteins are of unknown function and do not have convincing sequence homology with other proteins already existing in the databases. The corresponding genes, often called 'orphans', are one of the signatures of today's genomics. But the existence of 'orphans' in DNA sequences can be traced back to the early days of DNA sequencing. The complete sequence of human mtDNA, for example, revealed eight predicted open reading frames (ORFs) of unknown function that were without homology in the sequence databases of the time [1]. It was only several years later that the identity and function of such genes (seven of them encode NADH dehydrogenase subunits) were discovered by direct experimental investigation [2]. Those human mitochondrial genes have remained 'orphans' for several years for the sole reason that their equivalents were absent from the mitochondrial genome of the yeast *Saccharomyces cerevisiae*, the only other mtDNA molecule extensively studied at that time [3]. This historical example illustrates the ambiguity of 'orphans'. First, genes of unknown function may become functionally characterized with time, as has been the case for many *S. cerevisiae* genes since the completion of its sequence [4]. Second, absence of homology depends upon the spectrum of organisms compared. It happens that mtDNA molecules lack the genes encoding NADH dehydrogenase subunits in *S. cerevisiae*, but not in most other organisms. When discovered later, such genes were not called 'orphans' because the function of their human homologues was already characterized.

If the notion of 'orphans' is not novel, their abundance was first recognized with the yeast genome sequencing program [5,6]. The sequence of chromosome III, back in 1992 [7], came as a surprise by showing that, even in an extensively studied organism such as *S. cerevisiae*, nearly half of the protein-coding genes discovered from the sequence, and not yet experimentally studied, had no clearcut homologue in the sequence databases of the time [8]. The reality of such genes was debated as well as their actual proportion. But despite recent publications on this issue [9,10], the question is no longer pending because direct experimental studies have shown that a majority of the ca. 3000 initial yeast 'orphan' genes are actively transcribed, and that a sizeable fraction of them

*Corresponding author. Fax: (33)-1-40 61 34 56.
E-mail: bdujon@pasteur.fr

[1] These authors contributed equally to this work.

0014-5793/00/$20.00
PII: S0014-5793(00)02290-0

(ca. 13%, not very different from the general average) are essential for life [11]. Yet, 4 years after completion of its sequence, the actual number of genes in *S. cerevisiae* is not precisely known, and a number of questionable ORFs, predicted from the sequence, are often confused with actual 'orphan' genes.

If 'orphans' are the signature of both the actual diversity of the living world and the manner molecular geneticists look at it, the present sequencing project including a number of species from a unique phylogenetic group offers a unique opportunity to distinguish between the two views. This was indeed one of our initial motivations to undertake this project (see [12]). The goal of this article is to review the results of comparisons between the predicted *S. cerevisiae* genes, previously classified by comparisons to other organisms, with each of the 13 other yeast species partially sequenced in this project. Because the term 'orphan' has a functional connotation, we propose to designate 'maverick' the set of *S. cerevisiae* genes that do not share homologues in phylogenetic groups other than the Ascomycetes. We show that a majority of the 'maverick' genes have homologues among the other yeast species and define a set of 1892 genes that, from sequence comparisons, appear 'Ascomycetes-specific'. Therefore, in addition to the 3759 genes conserved in other phyla, the *S. cerevisiae* genome must contain 5651 actual genes.

2. Materials and methods

2.1. Databases

Comparisons of *S. cerevisiae* protein sequences were done against: (i) Unigene [13] (Mmuniq and Hsuniq containing, respectively, 96 349 and 91 244 entries); (ii) a non-redundant merged GenBank/EMBL database constructed at Institut Pasteur on November 30, 1999, and comprising 1 837 469 entries; (iii) our 'filtered' SwissProt database defined in [14] containing 58 365 entries and (iv) the genomes of *Caenorhabditis elegans, Schizosaccharomyces pombe*, the six completely sequenced archaeal species and 16 bacterial species as described in [14], plus *Deinococcus radiodurans* [15].

2.2. Comparison algorithms and selected limits of significance

All sequence comparisons were done using *blastp* (version 2) or *tblastn* (version 2) [16] and were performed using the script *blastallgenomes* (see [14]), with the *SEG* filter and the *pam250* substitution matrix [17,18]. Following a previous work using randomized sequences [19], limits of significance for *blast* comparisons were found at 10^{-9} for Hsuniq and Mmuniq sequences, 10^{-5} for *C. elegans, Bacillus subtilis, Borrelia burgdorferi, Mycobacterium tuberculosis, Methanococcus jannaschii* and *Pyrococcus abyssii*, 10^{-3} for *S. pombe*, *Treponema pallidum, Pyrococcus horikoshii, Rickettsia prowazeckii* and *D. radiodurans*, 10^{-2} for *Chlamydia trachomatis, Chlamydia pneumoniae* and *Aeropirum pernix*, and 10^{-4} for all other genomes. All *tblastn* matches of *S. cerevisiae* sequences against Unigene and the non-redundant merged GenBank/EMBL database were individually inspected to eliminate the contaminations by *S. cerevisiae* sequences and by vectors containing *S. cerevisiae* sequences in numerous database entries (human, mouse, rat, *C. elegans*). Comparisons were validated or rejected on a case by case basis. Results of comparisons against the non-redundant GenBank/EMBL database and against our 'filtered' SwissProt database were sorted according to whether the organism giving a match is an Ascomycete or not.

Homologies between *S. cerevisiae* gene products and our original random sequence tag (RST) sequences from the 13 other yeast species were taken from Table 1 of [12].

3. Results and discussion

3.1. Establishing the list of 'maverick' genes of S. cerevisiae

The set of 6213 protein sequences predicted from the *S. cerevisiae* genome [20] was first compared with the collections of protein sequences predicted from other completely sequenced organisms in order to examine how the overall proportion of *S. cerevisiae* genes having homologues in other phylogenetic groups evolves when additional organisms are analyzed.

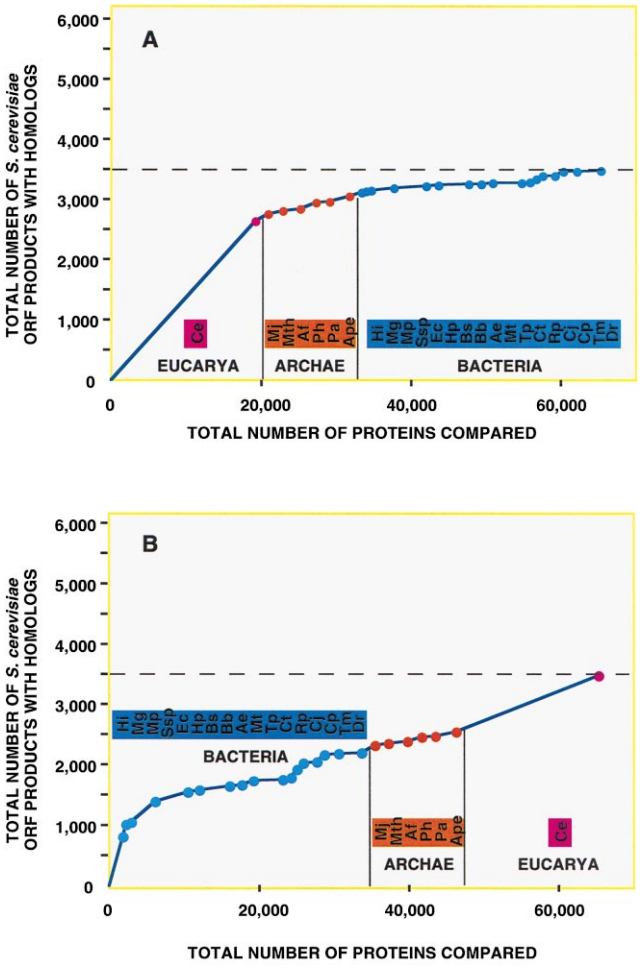


Fig. 1. Cumulative results of *blastp* comparisons of predicted ORF products from *S. cerevisiae* with other completely sequenced organisms. The figures represent the total number of *S. cerevisiae* ORF products (ordinate) having at least one homologue in serial comparisons with predicted proteins of other organisms (abscissa). Organisms are indicated by the following code: *Aquifex aeolicus*: Ae, *B. subtilis*: Bs, *B. burgdorferi*: Bb, *Campylobacter jejuni*: Cj, *C. pneumoniae*: Cp, *C. trachomatis*: Ct, *D. radiodurans*: Dr, *Escherichia coli*: Ec, *Haemophilus influenzae*: Hi, *Helicobacter pylori*: Hp, *M. tuberculosis*: Mt, *Mycoplasma genitalium*: Mg, *Mycoplasma pneumoniae*: Mp, *R. prowazeckii*: Rp, *Synechocystis* sp.: Ssp, *Thermotoga maritima*: Tm, *T. pallidum*: Tp, *A. pernix* K1: Ape, *A. fulgidus*: Af, *Methanobacterium thermoautotrophicum*: Mth, *M. jannaschii*: Mj, *P. abyssi*: Pa, *P. horikoshii*: Ph, *C. elegans*: Ce. Organisms were compared to *S. cerevisiae* in the order given by their code (left to right), starting with *C. elegans* (A) or *H. influenzae* (B). Data sources and total protein sequences of each organism are indicated in Table 1 of [14], except for *D. radiodurans* (3117 proteins, [15]). Note the absence of Ascomycetes or other fungi in these comparisons.

Fig. 1A shows that, after comparison with *C. elegans*, 3583 genes of *S. cerevisiae* remain without homologues (58% of total, a proportion very similar to that published in [21]). As expected, this number decreases when additional organisms are included in the comparisons because *S. cerevisiae*

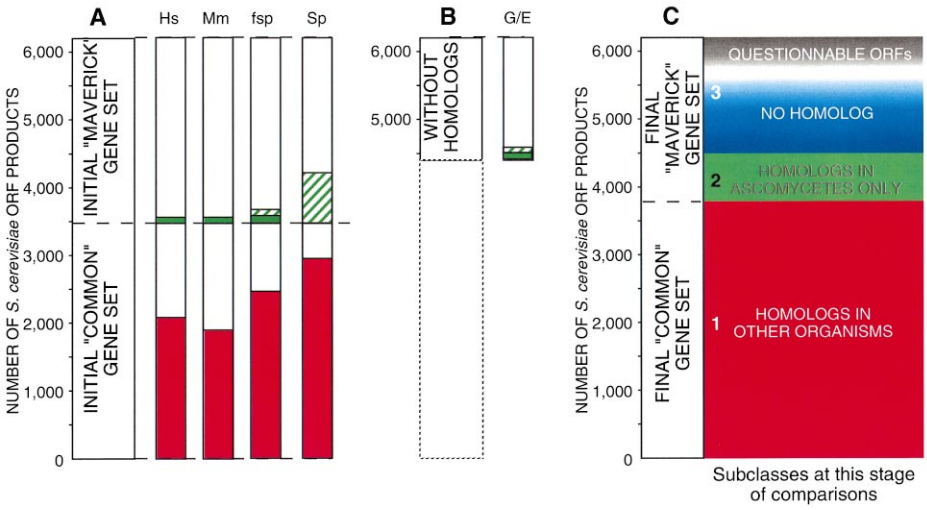


Fig. 2. Establishing the classification of *S. cerevisiae* ORFs prior to the comparisons with the 13 other yeast species. A: Vertical bars represent the number of *S. cerevisiae* ORF products of our initial 'common' (red) or 'maverick' (green) gene sets (see text and Fig. 1) having homologues when compared to: Hs: Hsuniq (91 244 entries), Mm: Mmuniq (96 349 entries), fsp: our 'filtered SwissProt' database (58 365 entries, see [14]), or Sp: the partial *S. pombe* sequence (3955 proteins). Striped bars indicate homology to Ascomycetes only, filled bars indicate homology to other organisms. B: Vertical bars represent the number of *S. cerevisiae* ORF products that have homologues when compared to our non-redundant merged GenBank/EMBL (G/E 1 837 469 entries). Note that the *tblastn* comparisons were done only for the 1919 ORFs that remained without homologues after the previous comparisons. Striped bars indicate homology to Ascomycetes only, filled bars indicate homology to other organisms. C: Classification of the *S. cerevisiae* ORF products resulting from all previous comparisons. The 'common' set contains 3759 ORFs, the 'maverick' set contains the remaining 2454 ORFs. ORFs of the 'maverick' set can be subdivided into those having homologues in Ascomycetes (subclass 2: 728 ORFs) and those having no homologue (subclass 3: 1726 ORFs). Note that among subclass 3 exist 313 ORFs regarded as 'questionable' ones, but the limit between actual and questionable ORFs is imprecise (color gradients).

genes that failed to have homologues in *C. elegans* may have homologues in *Archaea* or *Bacteria*. But when all 24 organisms were compared, a total of 2743 *S. cerevisiae* genes (44% of total) still remained without homologue in any other organisms. Depending on the order of comparisons, some species contribute more than others in providing novel homologues. For example, in the order of comparisons used in Fig. 1A, *P. horikoshii* contributes 107 novel homologues while *Archaeoglobus fulgidus*, although of comparable genome size, contributes only 39. Similarly, *M. tuberculosis*, although one of the largest bacterial genomes, contributes only four novel gene homologues compared to 61 or 62 for *C. trachomatis* or *C. pneumoniae* of much smaller genome sizes.

As the shape of such cumulative curves must vary according to the order of the comparisons, we have repeated the calculation starting with the *Bacteria*, then the *Archaea*, and placing *C. elegans* at the last position (Fig. 1B). When all 23 prokaryotic species were compared, there remained 3670 genes of *S. cerevisiae* without homologues. Of these, only 927 show homology when *C. elegans* is now included in the comparisons (compared to 2630, above). Thus, the two curves point out that, when more and more genomes are added for sequential comparisons with a given organism, less and less new genes find homologues. Finally, we obtained a total of 3470 genes of *S. cerevisiae* having at least one homologue in prokaryotic organisms or *C. elegans*. This defines our 'common' gene set. The remaining 2743 genes define our initial 'maverick' gene set.

The *S. cerevisiae* genes were then systematically compared with human, mouse and rat sequences retrieved from Unigene [13], as well with our 'filtered SwissProt' sequence collection [14]. Fig. 2 shows that the collection of 91 244 human gene sequences provides homologues to only 92 genes from the 'maverick' set (3.3% of the set) compared to 2069 genes from the 'common' set (59.6% of the set). Similarly, the collection of 96 349 rodent gene sequences gives homologues to only 73 of the 'maverick' genes (2.7%), compared to 1890 genes of the 'common' set (54.5%). Note that when the human and rodent sequences are taken together, they provide homologues to only 110 'maverick' genes, indicating that the two collections largely overlap despite the fact that they do not represent complete genome sequences. Thus, the vast majority of genes that are conserved between *S. cerevisiae* and human, mouse or rat belong to the class of genes also conserved in other organisms (prokaryotes or *C. elegans*). Addition of a large number of mammalian genes has not significantly reduced the number of 'maverick' *S. cerevisiae* genes. Similarly, when SwissProt was used for comparison, homologues were found to only 211 genes of the 'maverick' set (7.7%) compared to 2450 genes of the 'common' set (70.6%). Interestingly, a majority of the homologues to the 'maverick' set are provided by sequences from fungal species present in SwissProt. If one ignores those, only 92 genes of the 'maverick' set have homologues in SwissProt.

Taking into account the above three series of comparisons, there remained 2421 *S. cerevisiae* genes in our 'maverick' set, a relative decrease of only 11.7% compared to the previous figure (or 8.3% if one ignores the contribution of the Ascomycetes present in SwissProt). This already suggested to us that addition of novel organisms for comparisons should not significantly reduce the number of 'maverick' genes in *S. cerevisiae* as long as the novel organisms are not Ascomycetes. In more general terms, the set of genes conserved across major phylogenetic limits has essentially been described with the few organisms sequenced today, while the sets of 'phylum-specific' genes largely remain to be explored.

We then compared *S. cerevisiae* with *S. pombe*. Although only a partial set of 3955 *S. pombe* sequences were available when this work was done (see [14]), this comparison was interesting in that it was the first yeast species included. A total of 3644 *S. cerevisiae* genes have homologues in the subset of *S. pombe* genes used (Fig. 2) but, interestingly, 742 of them now fall in the 'maverick' set (20% of the total, compared to less than 4% for the mammalian sequences), despite the very remote phylogenetic relationship between the two yeast species. This suggested to us that at least a certain fraction of the 'maverick' genes of *S. cerevisiae* may be 'yeast-specific'.

Before examining this hypothesis, we wanted to ensure that homologues to our 'maverick' set were not missed. The remaining 1919 *S. cerevisiae* genes without significant homologues after all above comparisons were, therefore, compared against the non-redundant GenBank/EMBL database using *tblastn* (results were filtered from contaminating *S. cerevisiae* sequences in such database entries, see Section 2). A total of 226 genes were found to have homologues previously overlooked. However, 132 of them show homologues only in yeast or other fungal species. The remaining 94 genes have homologues in other eukaryotic phyla (more than 90% of the cases) or in *Bacteria* or *Archaea* (Fig. 2B).

Combining all previous results, we ended up with a classification of the *S. cerevisiae* genes containing 3759 genes in the 'common' set, i.e. with homologues in organisms other than Ascomycetes, and 2454 'maverick' genes (Fig. 2C). The 'maverick' set contains 728 genes known to have homologues in Ascomycetes before this sequencing program started (mostly *S. pombe* sequences, but also a variety of piecemeal sequences from the different ascomycetous subclasses) and 1726 genes without homologue. Note that this last category contains 313 genes labelled as 'questionable' in yeast databases based on their size, composition, or the fact that they largely overlap other genes (291 cases).

3.2. The nature of the 'maverick' genes and the total number of actual genes in *S. cerevisiae*

Given the previous classification of *S. cerevisiae* ORFs, it was interesting to examine their homologues after comparison with the RSTs from the 13 different yeast species of this program [22–34]. Results are detailed for every gene and for each yeast species in Table 1 of [12]. A numerical analysis of these results is given by Table 1 where it can be seen that, when all 13 yeast species are considered together, homologues are found to 3680 *S. cerevisiae* genes of the 'common' set (97.9% of total) and to 1712 genes of the 'maverick' set (69.8% of total). This global result is interesting for three reasons. First, it shows that a majority of the 'maverick' genes have homologues in other yeast species whereas they had no clearcut homologues in other groups of organisms. Second, it shows that the fraction of 'maverick' genes with homologues in other yeasts is significantly lower than the fraction of 'common' genes in the same situation. Third, it suggests that the genome of *S. cerevisiae* contains a minimum of 5471 actual protein-coding genes (3759+1712) that are now individually identified from sequence conservation.

But the last two points need a more detailed numerical analysis due to the presence of an unknown number of questionable ORFs among the 'maverick' set. If one now examines separately (Table 1, right part) the two previously defined subclasses of the 'maverick' set, it appears that a large number (1017) of the genes having homologues in the other yeast species originate from the previous subclass 3 (note that 91 of them were previously regarded as 'questionable' in databases, illustrating our previous difficulty to decide between actual genes and spurious ORFs). Yet, the proportion of such genes in subclass 3 (58.9%) is much lower than that in subclass 2 (95.5%, a figure which is very close to that observed for the 'common' set) consistent with subclass 3 containing only a subset of actual genes. If one assumes that the propor-

Table 1
Total number of *S. cerevisiae* genes having homologues in each of the other yeast species sequenced in this program

	Genome coverage	All ORFs total = 6213		'Common' set total = 3759		'Maverick' set total = 2454		'Maverick' set			
		total	image	total	image	total	image	subclass 2 total = 728		subclass 3 total = 1726	
								total	image	total	image
S. bayanus	**0.4×**	**2887**	**46.5**	**1990**	**52.9**	**897**	**36.6**	**366**	**50.3**	**531**	**30.6**
S. exiguus	0.2×	1600	25.8	1206	32.1	394	16.1	202	27.7	192	11.1
S. servazzi	0.2×	1535	24.7	1159	30.8	376	15.3	180	24.7	196	11.3
Z. rouxii	**0.4×**	**2427**	**39.1**	**1833**	**48.8**	**594**	**24.2**	**305**	**41.9**	**289**	**16.6**
S. kluyveri	0.2×	1562	25.1	1175	31.3	387	15.8	195	26.8	192	11.1
K. thermotolerans	0.2×	1575	25.4	1142	30.4	433	17.6	201	27.6	232	13.4
K. lactis	**0.4×**	**2597**	**41.7**	**1954**	**52.0**	**643**	**26.2**	**339**	**46.6**	**304**	**17.5**
K. marxianus	0.2×	1546	24.9	1195	31.8	351	14.3	172	23.6	179	10.3
P. angusta	**0.4×**	**2502**	**40.3**	**2069**	**55.0**	**433**	**17.6**	**270**	**37.1**	**163**	**9.4**
D. hansenii	0.2×	1290	20.8	1046	27.8	244	9.9	148	20.3	96	5.5
P. sorbitophila	**0.4×**	**1593**	**25.6**	**1384**	**36.8**	**209**	**8.5**	**152**	**20.9**	**57**	**3.3**
C. tropicalis	0.2×	1130	18.2	959	25.5	171	7.0	122	16.8	49	2.8
Y. lipolytica	**0.4×**	**1225**	**19.7**	**1059**	**28.2**	**166**	**6.8**	**110**	**15.1**	**56**	**3.2**
Altogether		**5392**	**86.8**	**3680**	**97.9**	**1712**	**69.8**	**695**	**95.5**	**1017**	**58.6**

The left part of the table indicates, for each yeast species against which sequence comparisons were made (see [22–34]), the total number of the 6213 predicted *S. cerevisiae* ORFs that have at least one validated homologue (o or oo). Data compiled from Table 1 of [12]. 'Image' indicates the % of the total *S. cerevisiae* ORFs having homologues in the sequenced set of RSTs (note that some yeast species were sequenced at ca. 0.4× genome coverage (bold lines) compared to only 0.2× coverage for others, see [12]). The last line (altogether) indicates the same calculations when data from all 13 species are considered together. The central and right parts of the table give the break up of the same data according to the previous classification of the *S. cerevisiae* ORFs (see Section 3.1 and Fig. 2). Note that the 'image' of *S. cerevisiae* is larger for ORFs of the 'common' set compared to ORFs of the 'maverick' set, and also differs between the subclasses of the 'maverick' set due to the presence of questionable ORFs among them.

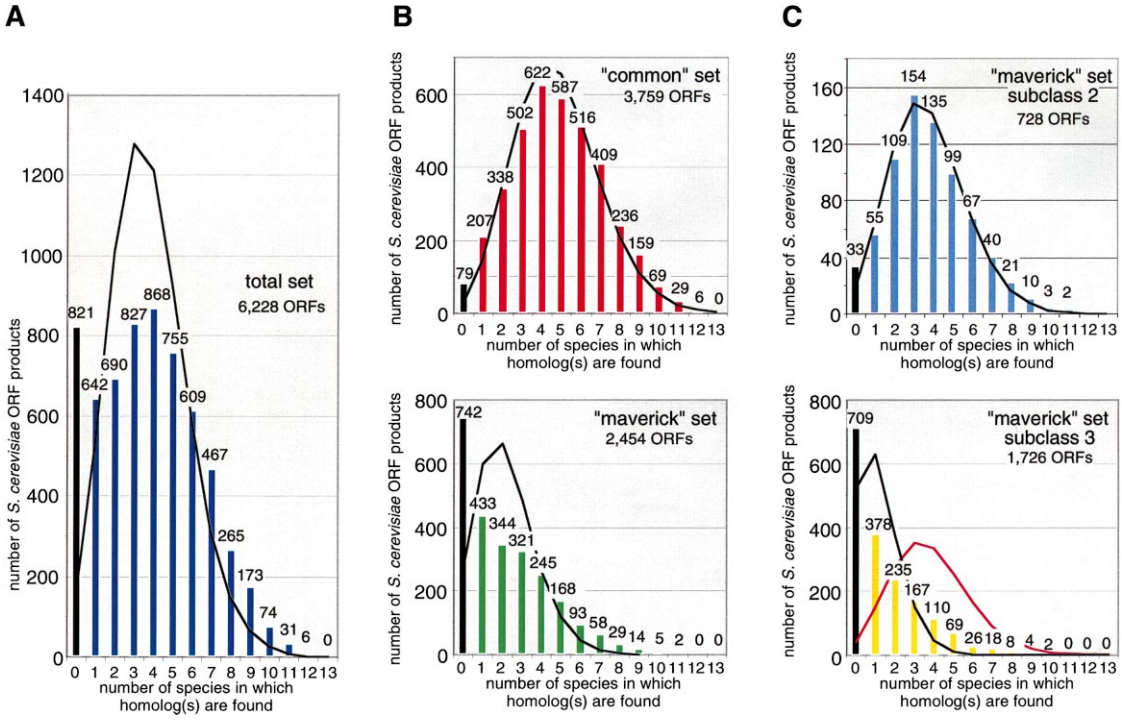


Fig. 3. Classification of the *S. cerevisiae* ORFs according to the total number of yeast species in which homologue(s) were found. A: Distribution for the entire set of predicted ORFs (6213) plus the novel ORFs discovered from this program (see [20]). Note that the latter have homologues in 1–4 species. Vertical bars represent the number of ORFs having homologues in a given number of yeast species (abscissa). Data computed from Table 1 of [12] and including all validated homologies ('o' and 'oo'). Average number of yeast species in which homologue(s) were found is 3.8. The black curve represents values of a Poisson distribution of the same average. B: Similar representation for ORFs of the 'common' set and those of the 'maverick' set, separately. Note that the novel ORFs discovered in this program are not included. Average numbers of yeast species in which homologue(s) were found are 4.8 and 2.2, respectively. C: Similar representation for ORFs of the two subclasses of the 'maverick' set (defined in Fig. 2). Average numbers of yeast species in which homologue(s) were found are 3.8 and 1.2 for subclasses 2 and 3, respectively. The red curve (lower panel) represents values of a Poisson distribution with a mean of 3.8, as in the upper panel.

tion of actual genes having homologues in other yeast species should be the same in the two subclasses of 'maverick' genes, then the number of actual genes in the subclass 3 should be 1065 only. By difference, the total number of questionable ORFs is 661, and the total number of actual protein-coding genes in *S. cerevisiae* should be 5552.

Beyond simple statistics, the most interesting aspect of this project is that it allows us to identify which of the *S. cerevisiae* ORFs are conserved or not conserved in other yeast species and in which species conservation is found. Such data can be found in Table 1 of [12] which, for each *S. cerevisiae* ORF, gives the list of other yeast species in which homologues were found and the degree of sequence divergence between the gene products. This table, however, has to be interpreted with caution because an absence of homologue in one yeast species may simply result from the fact that our sequencing coverage is limited to 0.2–0.4 genome equivalents. The statistical distribution of the number of yeast species providing homologues to each of the predicted *S. cerevisiae* ORFs was, therefore, analyzed (Fig. 3). When the entire set of predicted ORFs are considered (Fig. 1A), an average of 3.8 species give homologues to each *S. cerevisiae* ORF, with a distribution ranging from 0 to 12 species. It is clear, however, that the distribution is not random. Compared to a Poisson distribution, we observe a large excess of ORFs that have no homologue (821 found, 139 expected) and of ORFs with homologues in numerous species (seven and above) with a corresponding deficit of ORFs having homologues in numbers of species close to the mean. The origin of the non-randomness becomes explicit when the 'common' set and the 'maverick' set are considered separately (Fig. 3B). While the distribution for the 'common' set nearly follows a Poisson law with yet a slight excess of the zero class (79 found for 31 predicted), the distribution for the 'maverick' set is highly biased with 742 found in the zero class compared to 272 expected, pointing out the burden of the questionable ORFs within this set. The same conclusion is reached when one considers separately the distributions for the two subclasses of the 'maverick' set defined in Fig. 2. ORFs known to have homologues in Ascomycetes prior to the present sequencing program (subclass 2) show a nearly random distribution while ORFs without homologues prior to this program show a biased distribution. The Poisson curve is itself biased by the excessive weight of the zero class due to the presence of a significant proportion of questionable ORFs that reduces the mean to 1.2 compared to 3.8 for the subclass 2. If one uses a mean of 3.8 as in subclass 2, only 38 ORFs should fall in the zero class, compared to the 709 found. From this calculation, we estimate that subclass 3 contains 671 questionable ORFs and 1055 actual protein-coding genes (of which 1017 are identified by homology). Thus, the total number of actual protein-coding genes in *S. cerevisiae* should be 5542, a figure very close to the above estimate of 5552 (mean value is 5547). The problem is to identify the remaining real genes not found in other species for statistical reasons from the spurious ORFs and from the real genes that may be specific to *S. cerevisiae*. Of the 709 genes without homologue in

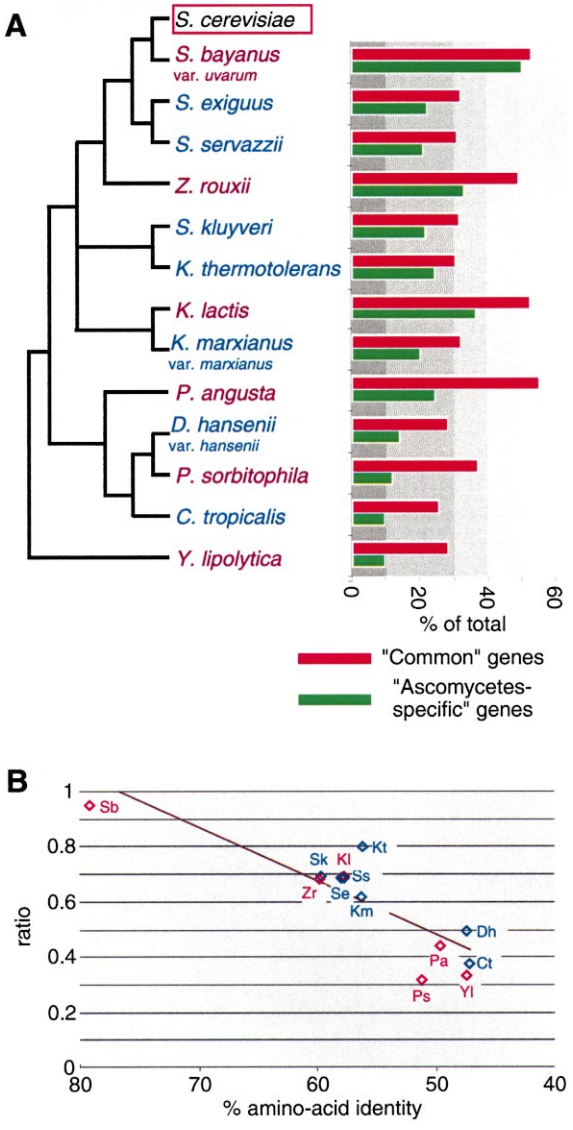


Fig. 4. Ascomycetes-specific genes in the *S. cerevisiae* genome. A: The figure represents, for each yeast species, the proportion of *S. cerevisiae* genes of the 'common' set (total 3759) and of the 'Ascomycetes-specific' set (total 1892) that have homologues. The cladogram is taken from [12]. Species sequenced at 0.4× genome equivalents are in red, those sequenced at 0.2× genome equivalents are in blue. B: For each yeast species is plotted the ratio of the above proportions ('Ascomycetes-specific'/'common', ordinate) as a function of sequence divergence from *S. cerevisiae* (mean amino acid identity, abscissa).

any of our other yeast species, 92 are functionally characterized in *S. cerevisiae*. This figure is in good agreement with the number of genes found in the other yeast species that are absent from *S. cerevisiae*. Thus the total number of *S. cerevisiae* genes should be increased by at least 54 (92−38).

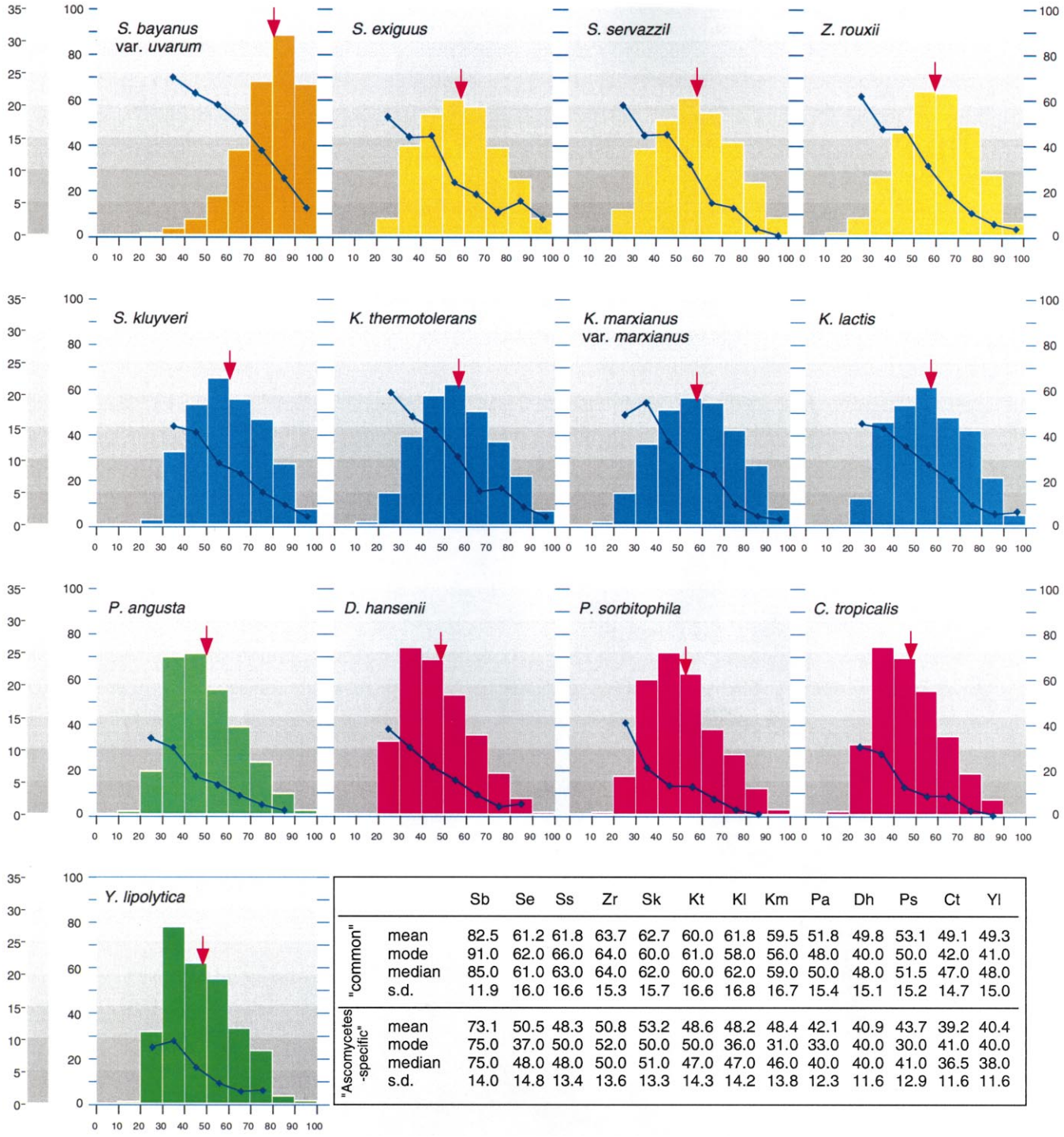
In conclusion, after this large comparative sequencing program of 13 other yeast species [22–34] and considering the 50 novel genes identified in this work (see [20]), the genome of *S. cerevisiae* appears to contain 5651 active genes of which ca. 32% have homologues only in other Ascomycetes or, for most of them so far, only in other Hemiascomycetes.

The total number of protein-coding genes in *S. cerevisiae* deduced here is much higher than that predicted by Kowalczuk and collaborators (4800) based on theoretical calculations [10] but is extremely similar to that predicted by Zhang and Wang (5645) based on other theoretical calculations [35]. In contrast to these authors, our estimate is based on experimental data. Yet, it may also be subject to discussion because the mere conservation of a sequence in an other yeast species does not, in itself, prove the presence of an active gene. We have, therefore, examined the distribution of the *S. cerevisiae* homologues in the different other yeast species. The rationale is that conservation in several species, especially the most distant ones, is a better argument for an active gene than conservation in a single species, especially the less distant one, *Saccharomyces bayanus* var. *uvarum*. For 175 of the 384 'maverick' ORFs having only one homologue, that homologue is in *S. bayanus* var. *uvarum*, compared to only 29, 34, 22, two and nine for *Z. rouxii*, *Kluyveromyces lactis*, *P. angusta*, *P. sorbitophila* and *Yarrowia lipolytica*, the other species sequenced at ca. 0.4 genome equivalents. A similar trend, though less pronounced, is observed for the 'maverick' ORFs having two or even three homologues. This difference is easily interpretable in classical terms by the increasing phylogenetic divergence of the other species with respect to *S. cerevisiae*. But it remains conceivable that some of the homologies found do not correspond to active genes. This phenomenon, however, must be limited to only a small fraction of the genes that we have considered here as 'Ascomycetes-specific'. Even if we considered as dubious all homologues found only in *S. bayanus* var. *uvarum* (175), or those found only in *S. bayanus* var. *uvarum* plus *S. exiguus* (16), or even those found only in the previous two species plus *S. servazzi* (1), then the total number of actual ORFs in *S. cerevisiae* could not be less than 5459. Even if we considered as dubious in addition all homologues found in *S. bayanus* var. *uvarum* plus any one of the 16 other species (129), or in *S. bayanus* var. *uvarum* plus any two other species (155), then the total number of actual ORFs in *S. cerevisiae* could not be less than 5175. Such hypotheses become extremely difficult to defend, unless one argues that sequence conservation is basically meaningless in terms of active function.

3.3. 'Ascomycetes-specific' genes

Results of Table 1 of [12] need to be analyzed in more detail because the presence of homologues to *S. cerevisiae* genes in each of the 13 other yeast species studied should also depend upon its phylogenetic distance to *S. cerevisiae*. Table 1 illustrates the phenomenon. Of the six yeast species sequenced at ca. 0.4× genome coverage, *S. bayanus* var. *uvarum* shows homologues to 46% of the *S. cerevisiae* ORFs, while *Z. rouxii*, *K. lactis* and *P. angusta* show homologues to ca. 40% of them, and the two most distantly related species, *P. sorbitophila* and *Y. lipolytica*, to only 26% and 20%, respectively. A similar effect is observed for the seven yeast species sequenced at ca. 0.2× genome coverage with *S. exiguus*, *S. servazzi*, *S. kluyveri*, *K. thermotolerans* and *Kluyveromyces marxianus* var. *marxianus* showing homologues to ca. 25% of the *S. cerevisiae* ORFs, and the two more distantly related species, *Debaryomyces hansenii* and *Candida tropicalis*, to only 21% and 18%, respectively.

All above figures are relative to the total number of predicted *S. cerevisiae* ORFs. If one now considers separately the 'common' set and the 'maverick' set, clearcut differences emerge. They are due, for one part, to the artefactual presence



		Sb	Se	Ss	Zr	Sk	Kt	Kl	Km	Pa	Dh	Ps	Ct	Yl
"common"	mean	82.5	61.2	61.8	63.7	62.7	60.0	61.8	59.5	51.8	49.8	53.1	49.1	49.3
	mode	91.0	62.0	66.0	64.0	60.0	61.0	58.0	56.0	48.0	40.0	50.0	42.0	41.0
	median	85.0	61.0	63.0	64.0	62.0	60.0	62.0	59.0	50.0	48.0	51.5	47.0	48.0
	s.d.	11.9	16.0	16.6	15.3	15.7	16.6	16.8	16.7	15.4	15.1	15.2	14.7	15.0
"Ascomycetes -specific"	mean	73.1	50.5	48.3	50.8	53.2	48.6	48.2	48.4	42.1	40.9	43.7	39.2	40.4
	mode	75.0	37.0	50.0	52.0	50.0	50.0	36.0	31.0	33.0	40.0	30.0	41.0	40.0
	median	75.0	48.0	48.0	50.0	51.0	47.0	47.0	46.0	40.0	40.0	41.0	36.5	38.0
	s.d.	14.0	14.8	13.4	13.6	13.3	14.3	14.2	13.8	12.3	11.6	12.9	11.6	11.6

Fig. 5. Distributions of amino acid identities in validated *blast* alignments between *S. cerevisiae* ORF products and their homologues in each of the other yeast species. For each yeast species, the distribution of amino acid identities with *S. cerevisiae* ('o' alignments only) is shown by the histograms (class intervals of 10%). The mean is indicated by the red arrow. To facilitate comparisons between species, data were normalized to the total number of alignments for each species and expressed in % (gray scale). The proportion (in %) of 'Ascomycetes-specific' genes within each class of the histogram is indicated by the curve. Classes containing less than 10 genes were ignored. Data computed from Table 1 of [12]. To help comparisons between species, the mean, modal and median values (in %) and the standard deviations (S.D.) are given in the insert. Data were computed separately for the 'common' genes and for the 'Ascomycetes-specific' genes.

of over 600 questionable ORFs in the 'maverick' set and, for the other part, to the actual sequence divergence within the Hemiascomycetes realm. We are now in a position to estimate the latter. Assuming, as before, that the proportion of genes having homologues in the subclass 3 of the 'maverick' set should be the same as that in its subclass 2, it becomes clear that the proportion of 'Ascomycetes-specific' genes of *S. cerevisiae* having homologues in the other yeast species decreases when the phylogenetic distance of that species to *S. cerevisiae* increases. This is not surprising in itself but the

interesting observation is that the relationship with the phylogenetic distance is more accentuated for the 'Ascomycetes-specific' genes than for the genes of the 'common' set (Fig. 4). In other words, 'Ascomycetes-specific' genes represent a class of genes that tend to be more sensitive to evolutionary divergence than the average. Species variations are, however, observed. *K. thermotolerans*, for example, shows a higher than average proportion of homologues to the 'Ascomycetes-specific' genes while the opposite is true for *P. sorbitophila*.

3.4. Sequence divergence in the 'Ascomycetes-specific' genes

To try to quantify this trend, we have examined the percent of amino acid identities in the sequence alignments between the *S. cerevisiae* gene products and their homologues in the other yeast species (Fig. 5). Despite the fact that such figures may be an imperfect representation of the actual divergence between two given genes (because some alignments concern gene fragments, not complete gene sequences, and because some segments may diverge more than others), their distribution gives a precise estimate of the degree of divergence between the species because the number of genes studied is large enough (ca. 20 000 genes in total from the 13 species). As expected, the distributions are broad (from ca. 20% to 100% identities, with standard deviations of 11–16%), indicating that some genes are more conserved than others. Yet, the distributions differ for the different yeast species. The mean values vary from ca. 79% for *S. bayanus* var. *uvarum* to ca. 47% for the more distant species *D. hansenii* var. *hansenii*, *C. tropicalis* and *Y. lipolytica*, with a number of intermediate species showing figures around 55–60%. The distributions of amino acid identities are, therefore, in general agreement with the phylogenetic distances between *S. cerevisiae* and the other yeasts based on rDNA sequences (see [12]). Note, however, that the closest yeast species studied, *S. bayanus* var. *uvarum*, show an average of ca. 21% amino acid divergence with *S. cerevisiae*, indicating that the two species are more distant than generally believed.

Fig. 5 also shows that the proportion of 'Ascomycetes-specific' genes is not equal in the various classes of identities. This phenomenon is remarkably constant for all species studied, declining from ca. 40–60% in the gene classes of low sequence conservation to less than 10% in the gene classes of high sequence conservation. Thus 'Ascomycetes-specific' genes are, on average, less conserved than other genes. This is quantitatively illustrated when one compares the classical parameters of the distributions of the 'Ascomycetes-specific' genes to those of the 'common' set of genes (Fig. 5, insert). A similar conclusion is reached if one considers the codon bias index which is significantly less pronounced for the 'Ascomycetes-specific' genes than for the other genes (data not shown).

The tendency for rapid sequence divergence may explain the existence of some 'Ascomycetes-specific' genes simply by the fact that the sequences of their homologues in other phylogenetic groups fall below our selected threshold to validate the sequence alignments as significant of homology. Specific examples have been recognized in our set of 'Ascomycetes-specific' genes that correspond to this situation. For example, a few *S. cerevisiae* genes encoding known mitoribosomal proteins or known transcription factors appear as 'Ascomycetes-specific' in our classification although it is clear that equivalent functions must exist in other phyla. But despite the fact that our homology criteria have been rather conservative, we do not believe that this situation affects a large number of cases because there exists a number of 'Ascomycetes-specific' genes which are highly conserved within our different yeast species (over 50% amino acid identities) but are not present elsewhere. Similarly, some genes present in some yeast species are absent in others.

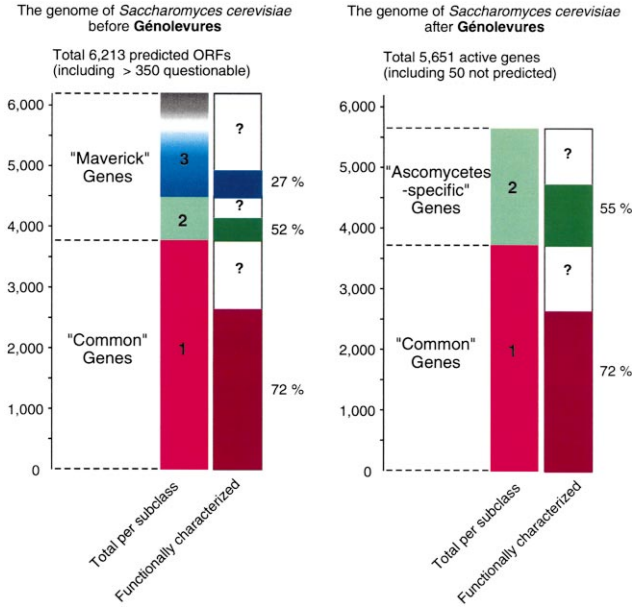


Fig. 6. ORFs and protein-coding genes of *S. cerevisiae*. Left panel: vertical bars represent the total number of ORFs predicted from the *S. cerevisiae* genome sequence (left) and the corresponding number of them which are functionally characterized (right). The total set is subdivided (as in Fig. 2) from sequence comparisons done prior to this work. The % of functionally characterized genes within each subclass is indicated on the right. Only ORFs with gene names are considered here as functionally characterized. Assigned functions based solely on sequence similarities are not taken into account. The question marks symbolize the fraction of each subclass that remains to be functionally characterized. Right panel: vertical bars represent the total number of protein-coding genes of *S. cerevisiae* as deduced from the sequence comparisons with the 13 other Hemiascomycetous species done in this work (left) and the corresponding number of them which are functionally characterized (right). The % of functionally characterized genes within each subclass is indicated on the right.

3.5. Functions of the 'Ascomycetes-specific' genes

The most interesting question is thus to determine to which extent the 'Ascomycetes-specific' genes may represent specific functional categories for yeasts, or perhaps fungi in general, that would not have equivalents in other phyla. This problem is examined in details in [36] using the available functional classification of the *S. cerevisiae* genes. If one only considers the functional categories containing at least 15 genes (for statistical significance) and that are either over- or under-represented at least two times in the 'Ascomycetes-specific' genes, a few interesting cases emerge. For example, 'Ascomycetes-specific' genes are over-represented among the genes involved in cell wall biosynthesis, a situation that can easily be interpreted because the cell wall is a characteristic feature of fungi, distinct in composition from the cellulosic wall of plant cells. It plays an essential role in protecting the yeast cells from osmotic variations of rapidly changing environment in natural conditions and is also a major morphogenetic component in-

volved in budding, mating or filamentation. In the pathogenic yeast *Candida albicans*, the cell wall contains receptors to the laminin or the fibronectin of the host [37,38] offering a target of choice for antifungal research.

Another interesting case is the over-representation of 'Ascomycetes-specific' genes in the functional category involved in pheromone response, a key element for speciation because it is an essential step before mating. The absence of homologues to *YDR461w* and *YNL145w* (*MFA1* and *MFA2*) in the 13 other yeast species is striking, particularly in view of the fact that homologues to *YPL187w* and *YGL089c* (*MFα1* and *MFα2*) were found in many species (*S. bayanus* 74% of identity, *S. exiguus* 58% of identity, *Z. rouxii* 65% of identity, *K. lactis* 48% of identity and *Y. lipolytica* 38% of identity) and that the receptor of the pheromone ***a*** is found in *K. thermotolerans*, *K. lactis* and *P. sorbitophila*. Diversity of the pheromone response pathway must play a role in limiting interspecific mating.

It is also interesting to note the over-representation of 'Ascomycetes-specific' genes in four functional categories classified as involved in the regulation of lipid, fatty acid and sterol biosynthesis, the regulation of amino acid metabolism, the regulation of nitrogen and sulphur utilization, and the regulation of carbohydrate utilization. Interestingly, most of the 'Ascomycetes-specific' genes of these categories are transcription factors which tend to be poorly conserved. A rapid evolution of transcription factors may induce considerable changes in the cell physiology and differentiate between the yeast species.

Other interesting examples are discussed in [36].

4. Concluding remarks

The extensive sequence comparisons permitted by this sequencing program on a homogeneous collection of yeasts have considerably improved our interpretation of the *S. cerevisiae* genome itself. Not only the total number of actual genes is now more precisely defined but nearly all of the *S. cerevisiae* protein-coding sequences can now be associated with a number of homologues from other yeast species of various evolutionary distances. As can be seen on Fig. 6, nearly half of the 'Ascomycetes-specific' genes remain to be functionally characterized. Conservation or divergence of their sequences in other yeast species may offer interesting clues for future functional studies.

Acknowledgements: We thank our colleagues from the Unité de Génétique moléculaire des levures for fruitful discussions and H. Feldmann and A. Goffeau for careful reading of the manuscript. B.D. is a member of Institut Universitaire de France.

References

[1] Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H.L. and Coulson, A.R. et al. (1981) Nature 290, 457–465.
[2] Chomyn, A., Mariottini, P., Cleeter, M.W.J., Ragan, C.I. and Matsuno-Yagi, A. et al. (1985) Nature 314, 592–597.
[3] Dujon, B. (1981) in: The Molecular Biology of the Yeast *Saccharomyces* (Strathern, J.N., Jones, E.W. and Broach, J.R., Eds.), pp. 505–635, Cold Spring Harbor Laboratory Press, NY.
[4] Goffeau, A. (2000) FEBS Lett. 480, 37–41.
[5] Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W. and Dujon, B. et al. (1996) Science 274, 563–567.
[6] Goffeau, A. et al. (1997) Nature 387, 1–105.
[7] Oliver, S.G. et al. (1992) Nature 357, 38–46.
[8] Dujon, B. (1996) Trends Genet. 12, 263–270.
[9] Mackiewicz, P., Kowalczuk, M., Gierlik, A., Dudek, M.R. and Cebrat, S. (1999) Nucleic Acids Res. 27, 3503–3509.
[10] Kowalczuk, M., Mackiewicz, P., Gierlik, A., Dudek, M.R. and Cebrat, S. (1999) Yeast 15, 1031–1034.
[11] http://websvr.mips.biochem.mpg.de/proj/eurofan.index.html.
[12] Souciet, J.L. et al. (2000) FEBS Lett. 487, 3–12 (this issue).
[13] Boguski, M.S. and Schuler, G.D. (1995) Nat. Genet. 10, 369–371.
[14] Tekaia, F., Blandin, G., Malpertuy, A., Llorente, B., Durrens, P. et al. (2000) FEBS Lett. 487, 17–30 (this issue).
[15] White, O. et al. (1999) Science 286, 1571–1577.
[16] Altschul, S.F. (1991) J. Mol. Biol. 219, 555–565.
[17] Altschul, S.F., Madden, T.L., Schöffer, A.A., Zheng Zhang, J.Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[18] Wootton, J.C. and Federhen, S. (1993) Comput. Chem. 17, 149–163.
[19] Tekaia, F. and Dujon, B. (1999) J. Mol. Evol. 49, 591–600.
[20] Blandin, G. et al. (2000) FEBS Lett. 487, 31–36 (this issue).
[21] Chervitz, S.A., Aravind, L., Sherlock, G., Ball, C.A. and Koonin, E.V. et al. (1998) Science 282, 2022–2028.
[22] Bon, E., Neuvéglise, C., Casaregola, S., Artiguenave, F., Winckler, P. et al. (2000) FEBS Lett. 487, 37–41 (this issue).
[23] Bon, E., Neuvéglise, C., Lépingle, A., Wincker, P., Artiguenave, F. et al. (2000) FEBS Lett. 487, 42–46 (this issue).
[24] Casaregola, S., Lépingle, A., Neuvéglise, C., Bon, E., Vang Nguyen, H. et al. (2000) FEBS Lett. 487, 47–51 (this issue).
[25] de Montigny, J., Straub, M.L., Potier, S., Tekaia, F., Dujon, B. et al. (2000) FEBS Lett. 487, 52–55 (this issue).
[26] Neuvéglise, C., Bon, E., Lépingle, A., Wincker, P., Artiguenave, F. et al. (2000) FEBS Lett. 487, 56–60 (this issue).
[27] Malpertuy, A., Llorente, B., Blandin, G., Artiguenave, F., Wincker, P. et al. (2000) FEBS Lett. 487, 61–65 (this issue).
[28] Bolotin-Fukuhara, M., Lemaire, M., Marmeisse, R., Montrocher, R., Termier, M. et al. (2000) FEBS Lett. 487, 66–70 (this issue).
[29] Llorente, B., Malpertuy, A., Blandin, G., Wincker, P., Artiguenave, F. et al. (2000) FEBS Lett. 487, 71–75 (this issue).
[30] Blandin, G., Llorente, B., Malpertuy, A., Wincker, P., Artiguenave, F. et al. (2000) FEBS Lett. 487, 76–81 (this issue).
[31] Lépingle, A., Casaregola, S., Bon, E., Neuvéglise, C., Vang Nguyen, H. et al. (2000) FEBS Lett. 487, 82–86 (this issue).
[32] de Montigny, J., Spehner, C., Souciet, J.L., Tekaia, F., Dujon, B. et al. (2000) FEBS Lett. 487, 87–90 (this issue).
[33] Blandin, G., Ozier-Kalogeropoulos, O., Wincker, P., Artiguenave, F. and Dujon, B. (2000) FEBS Lett. 487, 91–94 (this issue).
[34] Casaregola, S., Neuvéglise, C., Lépingle, A., Bon, E., Feynerol, C. et al. (2000) FEBS Lett. 487, 95–100 (this issue).
[35] Zang, C.T. and Wang, J. (2000) Nuc. Acid Res. 28, 2804–2814.
[36] Gaillardin, C., Duchateu Nguyen, G., Tekaia, F., Llorente, B. et al. (2000) FEBS Lett. 487, 134–149 (this issue).
[37] Bouchara, X. et al. (1990) Infect. Immun. 58, 48–54.
[38] Calderone, R.A (1991) Microbiol. Rev. 55, 1–20.