Genomic Exploration of the Hemiascomycetous Yeasts: 17. *Yarrowia lipolytica*

Serge Casaregola[a,*], Cécile Neuvéglise[a], Andrée Lépingle[a], Elisabeth Bon[a], Chantal Feynerol[a], François Artiguenave[b], Patrick Wincker[b], Claude Gaillardin[a]

[a] *Collection de Levures d'Intérêt Biotechnologique, Laboratoire de Génétique Moléculaire et Cellulaire, INRA UMR216, CNRS URA1925, INA-PG, F-78850 Thiverval-Grignon, France*

[b] *Génoscope, Centre National de Séquençage, 2 rue Gaston Crémieux, BP191, F-91057 Evry Cedex, France*

Received 3 November 2000; accepted 9 November 2000

First published online 29 November 2000

Edited by Horst Feldmann

Abstract A total of 4940 random sequence tags of the dimorphic yeast *Yarrowia lipolytica*, totalling 4.9 Mb, were analyzed. BLASTX comparisons revealed at least 1229 novel *Y. lipolytica* genes 1083 genes having homology with *Saccharomyces cerevisiae* genes and 146 with genes from various other genomes. This confirms the rapid sequence evolution assumed for *Y. lipolytica*. Functional analysis of newly discovered genes revealed that several enzymatic activities were increased compared to *S. cerevisiae*, in particular, transport activities, ion homeostasis, and various metabolism pathways. Most of the mitochondrial genes were identified in contigs spanning more than 47 kb. Matches to retrotransposons were observed, including a *S. cerevisiae* Ty3 and a LINE element. The sequences have been deposited with EMBL under the accession numbers AL409956–AL414895.

Key words: Non-conventional yeast; Transposable element; Mitochondrial DNA; Functional classification

1. Introduction

Yarrowia lipolytica, formerly known as *Candida*, *Endomycopsis* or *Saccharomycopsis lipolytica*, is one of the most extensively studied non-conventional yeasts for it is quite different from other intensively studied yeasts like *Saccharomyces cerevisiae* or *Schizosaccharomyces pombe*. Its ecological niche encompasses lipid-rich food like margarine, olive oil, cheese, as well as sewage and oil plants. *Y. lipolytica* is non-pathogenic to human and has been approved for several GRAS industrial processes. It is the only species recognized within the *Yarrowia* genus [1].

Y. lipolytica is dimorphic. It either forms yeast-like cells or true mycelium depending on certain conditions (specific nutrients, pH, etc.). Little is known on the genetic regulation of the process leading to dimorphism. *Y. lipolytica* is an obligate aerobe that is able to use several unusual carbon sources like paraffins, various alcohols, and acetate. *Y. lipolytica* secretes large amounts of various metabolites and enzymes, which is one of its characteristic features. It has been used for the production of citric acid. When grown on rich medium, it naturally secretes an alkaline protease at neutral pH or an acidic protease at low pH. Secreted RNase, phosphatase, lipase and esterase activities were also detected under diverse growth conditions [2]. Among these, at least three lipase genes have been cloned recently [3]. Several genes involved in the first step of secretion were isolated including several members of the signal recognition particle. The translocation of secretory proteins into the endoplasmic reticulum (ER) in *Y. lipolytica* is mainly co-translational like in higher eukaryotes, whereas it is a post-translational process in *S. cerevisiae* [4]. The combination of a strong inducible promoter and the ability of *Y. lipolytica* to secrete proteins in large amounts led to the set-up of a very efficient system for the secretion of heterologous proteins [5].

Y. lipolytica is heterothallic. Both *MATA* and *MATB* genes have been cloned. Unlike *S. cerevisiae*, there are no silent cassettes. All but one natural isolates were found to be haploids [6], suggesting that this yeast has a stable haploid life cycle. Natural strains of *Y. lipolytica* display unusual features including low mating frequency, low fertility of hybrids, irregular meiotic segregation and mitotic haploidization. Inbreeding programs have improved mating frequency and fertility of hybrids, allowing tetrad analysis and the construction of a genetic map [7].

Electrophoretic karyotypes were obtained for natural isolates and laboratory strains. The reference strain for genome structure studies, E150, harbors six chromosomes ranging from 2.6 to 4.9 Mb in size. The size of the genome was estimated at 21–22 Mb, which is much larger than those of *S. cerevisiae* and *S. pombe* (both around 13 Mb). Karyotypes revealed an important chromosome length polymorphism between strains of different origin, consistent with the poor fertility of hybrids. Yet, a linkage map built with 43 markers indicated that overall chromosome structure has been conserved in various isolates, indicating that the genome size is nearly constant in laboratory strains from various origins [8].

Y. lipolytica chromosomal origins of replication are not able to sustain plasmid replication without a centromeric sequence [9]. No sequence homology was found neither between centromeres or origins of replication within *Y. lipolytica* nor with those from other organisms. Ribosomal RNA gene clusters were found to be scattered on most of the chromosomes. In the reference strain, six clusters with sizes varying between 170 and 610 kb are located on at least four chromosomes [8]. In addition, up to five different types of rDNA units were

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

PII: S0014-5793(00)02288-2

identified in a single variant strain [10]. The 5S RNA gene is not present in the rDNA unit but gene copies are scattered throughout the genome. A retrotransposon, Ylt1, belonging to the Ty3/*gypsy* group has been identified [11]. It is bounded by unusually long (714 bp) long terminal repeats (LTRs). A recent survey of the GenEMBL database indicated that more than 100 nuclear genes have been cloned and sequenced. Also, more than 200 sequences carrying putative promoters are present in databases.

Several features thus bring *Y. lipolytica* phylogenetically close to higher eukaryotes: dispersion of the rDNA clusters and of the 5S RNA genes, small nuclear RNA sizes, the protein secretion process, and, in particular, the signal recognition particle 7S RNA. Phylogenetic analysis based on the comparison of the 18S rDNA and 26S rDNA indicates that *Y. lipolytica* sequences clearly diverged from those of other yeasts [12,13].

2. Materials and methods

2.1. Yeast strain

The strain W29 (CLIB89, CBS7504), a wild haploid isolate from French sewage, was used in this study. W29 is one of the parental strains of the French inbred lines [14] and its electrophoretic karyotype is very similar to that of the reference strain [8]. It was shown to be free of Ylt1 retrotransposons (Juretzek et al., in press).

2.2. Y. lipolytica genomic DNA library

The genomic DNA library (3264 clones) was constructed essentially as described [15].

2.3. Nucleic acid sequences

A total of 4940 random sequence tags (RSTs) with an average size of 995 bp were obtained [16]. Both ends of 2284 inserts were sequenced (over 92.4% of all RSTs). 372 inserts had only one end sequenced. The average size of the inserts of the library was found to be 3.91 kb (standard deviation, 1.2 kb). 34 clones had overlapping RSTs.

Sequence assembly was performed on the traces to increase the probability of generating the largest number of contigs with the programs phred (version 0.980904.c) and phrap (version 0.960731) [17,18]. Sequence of the vector pBAM3 was hidden from the RST library using cross-match (minscore 12; minmatch 20). Assembly was performed with phrap using a minscore of 14 and a minmatch of 30. Contigs were visualized with Consed, version 7 [19]. Since phrap does not perform assembly by creating a consensus sequence, additional bases were therefore left in the sequences. As no sequence editing was performed throughout this study, contigs may contain inaccurate bases that overestimate the size of the assemblies. A total of 2617 RSTs were included in 685 contigs. A large majority of the contigs (91.5%) were composed of two RSTs (510) and three RSTs (117). The contigs were subsequently used to define repeated sequences within the nuclear genome and extra-chromosomal sequences. Annotations were performed as described in [20].

3. Results and discussion

3.1. Ribosomal DNA

BLASTN comparison revealed matches with five contigs. Contig 843 carries all the genes of an entire rDNA unit. Contigs 839 and 840 carry an entire non-transcribed spacer (NTS) of 2899 and 3792 bp, respectively, confirming the variability between units is due to differences in length of the NTS [21]. Contigs 776 and 827 (three RSTs each) carry the junction between the rDNA unit (NTS region) and the adjacent chromosomal regions. We found direct repeats (an 11 bp sequence CTTGACGAGGC [21]) present five times in the contigs 839 and 843 and 14 times in contig 840.

A comparison of the *Y. lipolytica* 5S RNA gene sequence to the RST library unambiguously identified 10 RSTs sharing 94–97% identity. Along with *S. pombe* [22], *Y. lipolytica* is one of the few yeasts where 5S RNA genes are not included in the rDNA clusters [23]. Our data lead to an estimate of 40 dispersed copies of the 5S RNA genes.

3.2. Transposable elements

A retrotransposon with unusual features, Ylt1, was identified in *Y. lipolytica* and more than 30 copies are present in the strains studied [11]. Other strains, however, including W29, are devoid of this transposon. Here, we confirm and extend this finding since no match to this retrotransposon was found in our entire set of RSTs totalling 4.9 Mb. Further comparison analysis revealed that two contigs, 371 (two RSTs) and 665 (three RSTs), have significant matches with the *S. cerevisiae* Ty3B protein (32% identity over 137 amino acids (aa) and 33% over 370 aa, respectively). Contig 371 matches a short region of the reverse transcriptase functional domain and contig 665 matches the RNase H and the integrase functional domains. An extended search with Gproteome database [20] revealed that a third contig, 834 (10 reads), matches the LTR-less retrotransposon LINE [24,25]. We observed 22% aa identity over 738 aa of the mouse L1md Orf2 spanning the reverse transcriptase domain. This is the first LINE-like element discovered in a yeast.

The search for LTRs that flanked the putative Ty3-like elements turned out to be negative. Surprisingly, however, we found a sequence in the 3′ flanking region of the *YLSQS1* gene that matches one group of 10 RSTs displaying 95–89% identity over 277–273 bp starting with TGTTG and ending with CAATA, 5 bp sequences characteristic of LTRs. In addition, five of these putative LTRs were found to be bounded by the same 5 bp that probably account for a duplication of the target site, consistent with these sequences being solo LTRs. The remaining five sequences may be associated to retrotransposons but no sequence homology with known retrotransposon open reading frames (ORFs) was detected within these RSTs.

3.3. Mitochondrial DNA (mtDNA)

Six contigs (807 RSTs) spanning 47 181 bp belong to mtDNA. The following genes were unambiguously identified: *COX1*, *COX2*, *COX3* (cytochrome *c* oxidase subunits 1, 2 and 3), *COB* (apocytochrome *b*), *ATP9*, *ATP8* and *ATP6* (ATP synthase subunits 9, 8 and 6). Several subunits of the NADH dehydrogenase complex 1, ND1, ND2, ND3, ND4, ND4L, ND5 and ND6, were identified. At least five introns were found in the *COX1* gene and in the *COB* gene. Unlike the *S. cerevisiae COX3* gene, the *Y. lipolytica* homologue carries one intron. In addition, introns were found in two of the NADH dehydrogenase subunits, ND1 and ND5. Matches to intronic ORFs were also found, in particular by comparison with the *Pichia canadensis* mtDNA sequence. However, these intronic ORFs could not be assigned due to their close resemblance. Ribosomal RNA genes and tRNA genes were found to be poorly conserved and their assignment to the chromosomal map has to await further analysis. We deduced that the (G+C) content of mtDNA is 24.7% vs. 48.4% for nuclear DNA excluding rDNA sequences. A detailed description of the *Y. lipolytica* mitochondrial genome will be presented in Kerscher et al. (manuscript in preparation).

3.4. Annotations of Y. lipolytica RSTs

Each of the 4940 RSTs, except those corresponding to rDNA and mtDNA (984 RSTs), was systematically compared to various databases defined in [20].

The minimal number of genes homologous to *S. cerevisiae* genes in *Y. lipolytica*, determined as described in [15,26], amounts to at least 1083, the maximal number being 1177. Further comparison of the RSTs with other genomes and the SwissProt database [20] led to the identification of 146 (min) to 167 (max) genes. The overall number of newly identified genes amounts to at least 1229. Among these, 48 correspond to already known *Y. lipolytica* genes present in databases. Compared to the large number of genes identified in the other yeasts studied in this project, the comparably low figure of 1187 newly identified nuclear genes in *Y. lipolytica* for some 5000 RSTs confirmed that *Y. lipolytica* is more distant from *S. cerevisiae* than the other species of this project. Assuming that the genomes of *S. cerevisiae* and *Y. lipolytica* are similar in gene number, we consider that we have identified about 20% of the genes of *Y. lipolytica*, not including mitochondrial and tRNA genes.

We detected 42 nuclear tRNA genes corresponding to 13 aa. The presence of introns is conserved in each case when compared to *S. cerevisiae* but their sequences and sizes vary. Members of large tRNA gene families of *S. cerevisiae* tend to be more frequently observed in *Y. lipolytica* with notable exceptions (tD(GAC); tK(AAG); tR(AGA)).

3.5. Orthologues with no equivalent in S. cerevisiae

From a general comparison of the RSTs with genomes other than that of *S. cerevisiae* [20], we defined at least 146 new genes that are not found in *S. cerevisiae* (Table 1). Only two orthologues were found in Archaea and 11 orthologues in Bacteria. Among eukaryotes, we found 89 matches with yeast proteins (75 *S. pombe* and five *Y. lipolytica* paralogues), eight matches with fungal proteins and 41 matches with higher eukaryotic proteins (23 with *Caenorhabditis elegans*).

The identified genes are mostly involved in metabolism, like transcriptional regulators, and in transport, consistent with the observed over-representation of the latter type of functions (see below). Orthologues reflecting the known properties of *Y. lipolytica* on fatty acid metabolism were found in novel activities, such as a lipase and a fatty acid desaturase, in addition to the paralogues of the *Y. lipolytica LIP2* gene. Several proteases were also detected.

It must be stressed that a large part of the orthologues found here by comparing our sequences with Gproteome exist in *S. cerevisiae* but could not be detected because of sequence divergence (discussed in [26]). Of interest, in addition to orthologues of several cell cycle proteins, orthologues of *BIMB* involved in sister chromatid separation, and the mitosis regulator *BIME* (both in *Emericella nidulans*), as well as the *S. pombe* orthologue *CUT1*, were found, suggesting that a subset of proteins involved in cell cycle have diverged considerably and can only be detected by comparison with ascomycetous organisms. We nevertheless found genes that have not been shown to exist in *S. cerevisiae*. The case of the *S. pombe MEI2* gene, also found in the yeast *Pichia angusta* during this project [27], involved in the regulation of meiosis, is intriguing as it was thought to be specific of the *S. pombe* cell cycle [28]. A functional homologue of this gene was recently discovered in *Arabidopsis thaliana* but its role in the plant has not been described [29].

Other genes encoding, for example, the mitochondrial NADH dehydrogenase complex 1 subunits, β-glucosidase, or aryl sulfatase (from the bacterium *Pseudomonas aeruginosa*) were found in *Y. lipolytica* as well as in several species studied in this project. This indicates that these genes are very likely present in all organism but were recently lost in *S. cerevisiae* and closely related species (see [26]). The presence of genes encoding resistance to antibiotics like the homologue of the pristinamycin synthase of the bacterium *Streptomyces pristinaespiralis* may suggest lateral transfer.

3.6. Duplicated genes

Greater than 40% of the *S. cerevisiae* genes are members of families [30]. We found 45 orthologues of *S. cerevisiae* and four orthologues of other organisms that occurred at least twice in *Y. lipolytica*. When orthologues of *S. cerevisiae* singletons are considered, genes involved in metabolism were the most frequent. In particular, three copies of an urea transport protein were detected, consistent with *Y. lipolytica* being one of the rare hemiascomycete urease$^+$ strains. Three copies of a *S. cerevisiae* ORF similar to an acylglycerol lipase and two copies of a succinate coenzyme A (CoA) ligase were found. A peptide transporter was also present in three copies. The class of transporters and membrane proteins are well represented when orthologues of *S. cerevisiae* members of gene families are considered: allantoate permeases, glutathione transporter, proteins involved in iron uptake, voltage-gated chloride channel protein, ATP-binding cassette transporter, members of the major facilitator superfamily, and multidrug resistance proteins. In addition, a great deal of proteases were found; overall, 10 orthologues of proteases corresponding to five different families in *S. cerevisiae* were present in our search. Unlike transporters which can be found in several families that contain a large number of paralogues, proteases are less frequent and belong to smaller families in *S. cerevisiae*. We were further able to detect seven copies of alcohol dehydrogenases, three of which had been cloned. In the set of genes not present in *S. cerevisiae*, we detected two lipase genes, one being the known *LIP2* gene [3] and the other one representing a newly discovered gene. We also found a paralogue of the much studied alkaline extracellular protease [31].

3.7. Functional classification of the newly identified genes

We examined the functions of the *Y. lipolytica* gene orthologues to *S. cerevisiae* according to the MIPS functional catalog modified by Gaillardin et al. [26] and estimated the expected number of *Y. lipolytica* new genes per functional class [26]. When the expected number of *Y. lipolytica* genes was lower than that of *S. cerevisiae*, discrepancies were not analyzed, since this may reflect a lower degree of sequence conservation between orthologues that could result in a biased interpretation. By contrast, an over-representation of genes within a group of functions might reflect differences in physiology between the two yeasts.

In fact, in some cases over-representation resulted in a number of *Y. lipolytica* orthologues exceeding the number of genes in *S. cerevisiae*. For the 'biogenesis of peroxisomes' class, we found five genes in our search vs. only five genes existing overall in *S. cerevisiae*. In addition to three of the known *Y. lipolytica* genes, *PEX1*, *PEX2* and *PEX17*, we detected two

Table 1
Potential functions encoded by *Y. lipolytica* RSTs having no validated homologues in the genome of *S. cerevisiae*

Kingdom	Species	Accession number	Gene name	Function
Archaea	*Archaeoglobus fulgidus*	AF0367	*oxlT-2*	Oxalate/formate antiporter
	Pyrococcus horikoshii	PH0203	*pho03*	Maltose/maltodextrin transport ATP-binding protein
Bacteria	*Bacillus stearothermophilus*	Q53389	*amaB*	*N*-Carbamyl-l-amino acid amidohydrolase
	Campylobacter jejunii	Cj1199		1-Aminocyclopropane-1-carboxylate oxidase
	Escherichia coli	ECyjaB	*ECjaB*	Hypothetical protein
	E. coli	ECb1327		Hypothetical protein
	P. aeruginosa	P51691	*atsA*	Arylsulfatase
	Lactococcus lactis (subsp. *cremoris*)	O87765	*pcp*	Pyroglutamyl-peptidase I
	Rhodococcus sp. (strain IGTS8)	P54995	*soxA*	Dibenzothiophene desulfurization enzyme A
	Mycobacterium tuberculosis	MTRv3342		SAM-dependent methyltransferase
	M. tuberculosis	MTRv3049c		Predicted flavoproteins involved in potassium transport
	M. tuberculosis	MTRv0014c		Similarity to serine/threonine protein kinases active-site signature
	S. pristinaespiralis	P54991	*snaA*	Pristinamycin IIa synthase subunit A
Asco	*Candida albicans*	P43062	*CLN2*	G1/S-specific cyclin
	Kluyveromyces lactis	P49374	*HGT1*	Glucose transporter high-affinity
	Kluyveromyces marxianus	Q07288	*ADH1*	Alcohol dehydrogenase I
	Saccharomycopsis fibuligera	P22507	*BGL2*	β-D-Glucoside glucohydrolase
	Y. lipolytica	P09230	*XPR2B*	Alkaline extracellular protease precursor
	Y. lipolytica	AJ012632	*LIP2*	Triacylglycerol lipase
	Y. lipolytica	AJ012632	*LIP2*	Triacylglycerol lipase
	Y. lipolytica	Q99155	*PEX2*	Peroxisomal assembly protein – peroxin
	Y. lipolytica	P87200	*PEX17*	Peroxisomal membrane protein pex17
	Aspergillus niger	Q12556	*AO-I*	Copper amine oxidase 1
	A. niger	O74180	*AOX1*	Alternative oxidase precursor
	E. nidulans	P33144	*BIMB*	Cell division-associated protein BimB
	E. nidulans	P24686	*BIME*	Negative regulator of mitosis
	Mycosphaerella graminicola	O42764	*HPPD*	4-Hydroxyphenylpyruvate dioxygenase
	Neurospora crassa	P38680	*MTR*	Amino acid transport system protein
	Penicillium camembertii	P25234	*MDLA*	Mono- and diacylglycerol lipase precursor
	Podospora anserina	P20681	*COI*	Cytochrome *c* oxidase polypeptide I
	S. pombe	U2AG_SCHPO	*U2AG*	Splicing factor U2Af
	S. pombe	BC12C2.02C	*STE16*	Necessary for sexual differentiation and meiosis
	S. pombe	SPCC757.05C		Putative acetylornithine deacetylase
	S. pombe	SPCC736.13		Hypothetical protein
	S. pombe	SPCC70.08C		Probable methyltransferase
	S. pombe	SPCC663.01C		Sap2 family putative cell cycle-dependent phosphatase associated protein
	S. pombe	SPCC645.13		Hypothetical protein
	S. pombe	SPCC4G3.12C		Hypothetical protein
	S. pombe	SPCC4G3.11		Hypothetical protein
	S. pombe	SPCC4G3.07C		Hypothetical protein
	S. pombe	SPCC320.08		Hypothetical protein
	S. pombe	SPCC320.08		Hypothetical protein
	S. pombe	SPCC306.04C		Set domain protein; transcriptional silencing
	S. pombe	SPCC1919.14C		Putative transcription factor TFIIIB component
	S. pombe	SPCC1840.03		Importin β-subunit
	S. pombe	SPCC18.03		Putative cystine-rich transcriptional regulator
	S. pombe	SPCC1450.07C		Putative D-amino acid oxidase
	S. pombe	SPCC1322.17C		Similarity to hypothetical proteins
	S. pombe	SPCC126.14		Putative mRNA splicing factor
	S. pombe	SPCC1259.12C		Similarity to human RANBPM
	S. pombe	SPBC947.07		Possible involvement in ribosome biosynthesis Surf-like
	S. pombe	SPBC651.09C		Conserved hypothetical protein
	S. pombe	SPBC4B4.10C		Apoptosis-specific protein homologue
	S. pombe	SPBC23G7.13C		Putative urea active transporter
	S. pombe	SPBC19C7.04C		Hypothetical protein
	S. pombe	SPBC17D11.04C		Putative transcriptional regulator, PHD finger protein
	S. pombe	SPBC15D4.02		Zinc finger protein
	S. pombe	SPBC15C4.06C		Zinc finger C3HC4 type protein
	S. pombe	SPBC146.08C		Hypothetical protein
	S. pombe	SPBC13G1.07		Hypothetical protein
	S. pombe	SPBC13G1.04C		Hypothetical protein
	S. pombe	SPBC1271.10C		Putative MSF transporter
	S. pombe	SPAC8C9.14		Putative heat shock transcription factor
	S. pombe	SPAC6B12.07C		Hypothetical zinc finger protein
	S. pombe	SPAC637.13C		Hypothetical protein
	S. pombe	SPAC4F10.06		Hypothetical protein
	S. pombe	SPAC3H8.08C		Putative transcriptional regulatory protein
	S. pombe	SPAC3H1.11		Hypothetical zinc finger protein
	S. pombe	SPAC3A11.08		Cullin homologue
	S. pombe	SPAC2F7.16C		Putative phospholipase D1

Table 1 (*continued*)

Kingdom	Species	Accession number	Gene name	Function
	S. pombe	SPAC2F3.13C		Probable queuine tRNA-ribosyltransferase
	S. pombe	SPAC2F3.08		Putative sucrose carrier
	S. pombe	SPAC2C4.17C		Similarity to hypothetical proteins
	S. pombe	SPAC27D7.08C		Hypothetical protein
	S. pombe	SPAC26H5.04		Hypothetical protein
	S. pombe	SPAC26F1.08C		Hypothetical protein
	S. pombe	SPAC25G10.01		Hypothetical protein
	S. pombe	SPAC24H6.13		Putative major facilitator superfamily protein
	S. pombe	SPAC24C9.05C		Hypothetical protein
	S. pombe	SPAC24C9.05C		Hypothetical protein, membrane protein
	S. pombe	SPAC23A1.16		Hypothetical protein
	S. pombe	SPAC22G7.02		Hypothetical protein
	S. pombe	SPAC22G7.01C		Hypothetical protein
	S. pombe	SPAC22F3.13		Hypothetical protein
	S. pombe	SPAC22E12.02		RNA-binding protein
	S. pombe	SPAC1F7.11C		Putative transcriptional regulatory protein
	S. pombe	SPAC19G12.01C		Hypothetical protein
	S. pombe	SPAC18B11.03C		Hypothetical protein
	S. pombe	SPAC17H9.16		Putative mitochondrial import receptor subunit
	S. pombe	SPAC17A5.16		Hypothetical protein
	S. pombe	SPAC1327.01C		Hypothetical protein
	S. pombe	SPAC12B10.16C		Hypothetical protein
	S. pombe	SPAC12B10.16C		Hypothetical protein, membrane protein
	S. pombe	SPAC12B10.03		WD repeat protein
	S. pombe	SPAC10F6.11C		Hypothetical protein
	S. pombe	SPSIN1	*SIN1*	Stress-activated map kinase interacting protein
	S. pombe	SPRAD1	*RAD1*	DNA repair protein
	S. pombe	SPPHP5	*PHP5*	Hap5p homologue
	S. pombe	SPPAC2	*PAC2*	cAMP-independent regulatory protein Pac2
	S. pombe	SPMOE1	*MOE1*	Negative regulator for microtubule dynamics
	S. pombe	SPMEI2	*MEI2*	Mei2 protein
	S. pombe	SPGAP1	*GAP1*	GTPase-activating protein
	S. pombe	SPCUT1	*CUT1*	Homologue to cell division-associated protein BimB
	S. pombe	SPCDC17	*CDC17*	DNA ligase
	S. pombe	SPCDB4	*CDB4*	Curved DNA-binding protein
Other eukarya	*C. elegans*	CEZK945.10		Protein required for males to locate the hermaphrodite vulva
	C. elegans	CEZK84.1		Similarity to human mucin
	C. elegans	CEY102A5A.1		Protein of unknown function
	C. elegans	CEW03G1.7		Putative acid sphingomyelinase
	C. elegans	CET09B4.10		Protein of unknown function
	C. elegans	CER151.6		Similarity to *S. cerevisiae* Der1p involved in degradation of misfolded soluble proteins in the ER
	C. elegans	CER02F11.3		Hypothetical protein
	C. elegans	CEM03C11.1		Similarity to human and *Drosophila melanogaster* cAMP-dependent kinases
	C. elegans	CEK10H10.3		Protein of unknown function
	C. elegans	CEK09E4.3		Protein of unknown function
	C. elegans	CEF55A11.3		Similarity to *S. cerevisiae* Hrd1p, required for ER degradation of misfolded lumenal and integral membrane proteins
	C. elegans	CEF48E3.3		Strong similarity to *D. melanogaster* UGT, UDP-glucose-glycoprotein glucosyltransferase
	C. elegans	CEF45H11.2		Member of the ubiquitin family, protein synthesis ribosome associated
	C. elegans	CEF45D3.5		Protein degradation in the ER
	C. elegans	CEF38E1.9		Strong similarity to human MPDU1
	C. elegans	CEF27E11.1		Putative orthologue of human SLC28A2 protein
	C. elegans	CEF22D6.4		Protein of unknown function
	C. elegans	CEF02A9.5		Similar to human and *Drosophila* propionyl-CoA carboxylases
	C. elegans	CED2096.4		Predicted in amino acid metabolism
	C. elegans	CEC50B8.3		Protein of unknown function
	C. elegans	CEC41C4.7		Putative cystinosin
	C. elegans	CEC17G10.8	*fat3*	Fatty acid desaturase
	D. melanogaster	P18173	*GLD*	Glucose dehydrogenase
	Homo sapiens	Q15393	*KIAA0017*	Hypothetical protein
	H. sapiens	Q15166	*PON3*	Arylesterase
	H. sapiens	Q13216	*CKN1*	Cockayne syndrome WD repeat protein Csa
	H. sapiens	Q02817	*MUC2*	Mucin 2 precursor
	H. sapiens	P78381	*UGALT*	Galactose translocator
	H. sapiens	P35579	*MYH9*	Myosin, heavy polypeptide 9
	H. sapiens	P09661	*SNRPA1*	U2 small nuclear ribonucleoprotein A′

Table 1 (*continued*)

Kingdom	Species	Accession number	Gene name	Function
	Leishmania amazonensis	P42865	*CRYZ*	Possible quinone oxidoreductase
	Mus musculus	Q60759	*GCDH*	Glutaryl-CoA dehydrogenase precursor
	M. musculus	Q04519	*SMPD1*	Sphingomyelinase
	M. musculus	P23949	*BRF2*	Butyrate response factor 2
	M. musculus	P22227	*REX-1*	Reduced expression-1 protein
	Plasmodium simium	Q03110	*CS*	Circumsporozoite protein precursor
	Rattus norvegicus	Q62871	*DNCI2*	Dynein intermediate chain 2
	R. norvegicus	Q63100	*DNCI1*	Dynein intermediate chain 1
	R. norvegicus	P70473	*PPP2R3*	2-Arylpropionyl-CoA epimerase
	Rhizopus niveus	P43231	*CARB*	Rhizopuspepsin 2 precursor
	Sus scrofa	P17403	*GLTP*	Glycolipid transfer protein

new peroxisomal genes. We found six orthologues of genes involved in 'other metabolism of amino acids' while five genes are present in *S. cerevisiae*. Likewise, the 'transport of nitrogen and sulfur' class with the 10 *Y. lipolytica* orthologues detected here exceed the eight genes of *S. cerevisiae*.

We also found an increased level for some functions though to a lesser extent, like 'nuclear biogenesis' (199% increase), 'carbohydrate transport' (185% increase), 'polynucleotide degradation' (214% increase). As expected from its known secretory ability, the class comprising the extracellular and secreted proteins was found over-represented in *Y. lipolytica* and amongst these, five proteases. 'β-Oxidation of fatty acids', a class shown to be increased in *Debaryomyces hansenii* [32], was also clearly increased in *Y. lipolytica* (256%). Overall, metabolism of lipids and fatty acids (class 01.06) is over-represented in *Y. lipolytica*, as expected. Except for sub-classes comprising the genes involved in the regulation of these pathways and undefined genes (sub-class 01.06.99), all the other sub-classes with genes involved in the biosynthesis, the breakdown and the utilization of lipids, indicated a dramatic increase up to four times the number of genes expected in *S. cerevisiae*. Incidentally, this result shows that the method we followed in this project is really indicative of the variation in the representation of some metabolic pathways compared to *S. cerevisiae*.

Another class of functions 'transport facilitation' (class 07) is over-represented in *Y. lipolytica*, in particular the 'ion transporter' (sub-class 07) was systematically higher. Over-representation of transport genes is also representative for *D. hansenii* [32].

Acknowledgements: We are very grateful to C. Caron (INRA, Jouy-en-Josas), J.-M. Vansteene (INRA, Grignon) and F. Tekaia (Institut Pasteur) for their help with the set-up of computing tools. We thank Prof. G. Barth for the communication of unpublished sequences and Prof. D. Ogrydziak for helpful discussion. E.B. was supported by the EEC scientific research Grant QLRI-1999-01333. Part of this work was supported by a BRG Grant (ressources génétiques des microorganismes no. 11-0926-99).

References

[1] Kurtzman, C.P. (1998) in: The Yeasts (Kurtzman, C.P. and Fell, J.W., Eds.), pp. 420–421, Elsevier, Amsterdam.
[2] Barth, G. and Gaillardin, C. (1997) FEMS Microbiol. Rev. 19, 219–237.
[3] Pignede, G., Wang, H., Fudalej, F., Gaillardin, C., Seman, M. and Nicaud, J.M. (2000) J. Bacteriol. 182, 2802–2810.
[4] Beckerich, J.-M., Boisramé, A. and Gaillardin, C. (1998) Int. Microbiol. 1, 123–130.
[5] Madzak, C., Blanchin-Roland, S., Cordero Otero, R.R. and Gaillardin, C. (1999) Microbiology 145, 75–87.
[6] Wickerham, L.J., Kurtzman, C.P. and Herman, A.I. (1970) Science 167, 1141.
[7] Ogrydziak, D., Bassel, J. and Mortimer, R. (1982) Mol. Gen. Genet. 188, 179–183.
[8] Casaregola, S., Feynerol, C., Diez, M., Fournier, P. and Gaillardin, C. (1997) Chromosoma 106, 380–390.
[9] Vernis, L., Abbas, A., Chasles, M., Gaillardin, C.M., Brun, C., Huberman, J.A. and Fournier, P. (1997) Mol. Cell. Biol. 17, 1995–2004.
[10] Fournier, P., Gaillardin, C., Persuy, M.A., Klootwijk, J. and van Heerikhuizen, H. (1986) Gene 42, 273–282.
[11] Schmid-Berger, N., Schmid, B. and Barth, G. (1994) J. Bacteriol. 176, 2477–2482.
[12] Keogh, R.S., Seoighe, C. and Wolfe, K.H. (1998) Yeast 14, 443–457.
[13] Kurtzman, C.P. and Robnett, C.J. (1998) Antonie Van Leeuwenhoek 73, 331–371.
[14] Barth, G. and Gaillardin, C. (1996) in: Non Conventional Yeasts in Biotechnology (Wolfe, K., Ed.), pp. 313–388, Springer, Berlin.
[15] Casaregola, S., Lépingle, A., Neuvéglise, C., Bon, E., Nguyen, H.V., Artiguenave, F., Wincker, P. and Gaillardin, C. (2000) FEBS Lett. 487, 47–51 (this issue).
[16] Artiguenave, F., Wincker, P., Brottier, P., Duprat, S., Jovelin, F., Scarpelli, C., Verdier, J., Vico, V., Weissenbach, J. and Saurin, W. (2000) FEBS Lett. 487, 13–16 (this issue).
[17] Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) Genome Res. 8, 175–185.
[18] Ewing, B. and Green, P. (1998) Genome Res. 8, 186–194.
[19] Gordon, D., Abajian, C. and Green, P. (1998) Genome Res. 8, 195–202.
[20] Tekaia, F., Blandin, G., Malpertuy, A., Llorente, B., Durrens, P. et al. (2000) FEBS Lett. 487, 17–30 (this issue).
[21] van Heerikhuizen, H., Ykema, A., Klootwijk, J., Gaillardin, C., Ballas, C. and Fournier, P. (1985) Gene 39, 213–222.
[22] Barnitz, J.T., Cramer, J.H., Rownd, R.H., Cooley, L. and Soll, D. (1982) FEBS Lett. 143, 129–132.
[23] Clare, J.J., Davidow, L.S., Gardner, D.C. and Oliver, S.G. (1986) Curr. Genet. 10, 449–452.
[24] Hattori, M., Kuhara, S., Takenaka, O. and Sakaki, Y. (1986) Nature 321, 625–628.
[25] Loeb, D.D., Padgett, R.W., Hardies, S.C., Shehee, W.R., Comer, M.B., Edgell, M.H. and Hutchison, C.A.d. (1986) Mol. Cell. Biol. 6, 168–182.
[26] Gaillardin, C., Duchateau-Nguyen, G., Tekaia, F., Llorente, B., Casaregola, S. et al. (2000) FEBS Lett. 487, 134–149 (this issue).
[27] Blandin, G., Llorente, B., Malpertuy, A., Wincker, P., Artiguenave, F. and Dujon, B. (2000) FEBS Lett. 487, 31–36 (this issue).
[28] Yamamoto, M. (1996) Cell. Struct. Funct. 21, 431–436.
[29] Hirayama, T., Ishida, C., Kuromori, T., Obata, S., Shimoda, C., Yamamoto, M., Shinozaki, K. and Ohto, C. (1997) FEBS Lett. 413, 16–20.
[30] Tekaia, F. and Dujon, B. (1999) J. Mol. Evol. 49, 591–600.
[31] Davidow, L.S., O'Donnell, M.M., Kaczmarek, F.S., Pereira, D.A., DeZeeuw, J.R. and Franke, A.E. (1987) J. Bacteriol. 169, 4621–4629.
[32] Lépingle, A., Casaregola, S., Bon, E., Neuvéglise, C., Nguyen, H.V., Wincker, P., Artiguenave, F. and Gaillardin, C. (2000) FEBS Lett. 487, 82–86 (this issue).