Genomic Exploration of the Hemiascomycetous Yeasts: 12. *Kluyveromyces marxianus* var. *marxianus*

Bertrand Llorente[a,*], Alain Malpertuy[a], Gaëlle Blandin[a], François Artiguenave[b], Patrick Wincker[b], Bernard Dujon[a]

[a] *Unité de Génétique Moléculaire des Levures, (URA 2171 du CNRS and UFR 927 Université Pierre et Marie Curie), Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France*
[b] *Génoscope-Centre National de Séquençage, 2 rue Gaston Crémieux P.O. Box 191, 91006 Evry Cedex, France*

Received 3 November 2000; accepted 9 November 2000

First published online 27 November 2000

Edited by Horst Feldmann

Abstract As part of the comparative genomics project 'GENOLEVURES', we studied the *Kluyveromyces marxianus* var. *marxianus* strain CBS712 using a partial random sequencing strategy. With a 0.2× genome equivalent coverage, we identified ca. 1300 novel genes encoding proteins, some containing spliceosomal introns with consensus splice sites identical to those of *Saccharomyces cerevisiae*, 28 tRNA genes, the whole rDNA repeat, and retrotransposons of the Ty1/2 family of *S. cerevisiae* with diverged Long Terminal Repeats. Functional classification of the *K. marxianus* genes, as well as the analysis of the paralogous gene families revealed few differences with respect to *S. cerevisiae*. Only 42 *K. marxianus* identified genes are without detectable homolog in the baker's yeast. However, we identified several genetic rearrangements between these two yeast species.

Key words: Intron; Retrotransposon; Genomic library; tRNA; Synteny; *Saccharomyces cerevisiae*

1. Introduction

Kluyveromyces marxianus is a homothallic hemiascomycetous species usually encountered on cheese and other dairy products [1], and occasionally in human infections such as oesophagitis and vaginitis [2,3]. Because of their ability to mate and produce fertile hybrids [4], distinction between various species of the genus *Kluyveromyces* has been ascertained only recently thanks to DNA reassociation studies [5,6] and confirmed by the analysis of their rDNA-based phylogeny [7]. *K. marxianus* is the only inulin-assimilating *Kluyveromyces* species that does not assimilate or ferment α-glucoside and grows well at 37°C. Intraspecific polymorphism in the *K. marxianus* species is very high, and numerous synonyms thus exist, such as *Candida pseudotropicalis* and *Candida kefyr*. *C. kefyr* is the anamorph of this species. The chromosome number of *K. marxianus* isolates varies from 6 to 12, but most of the strains contain eight chromosomes. The type strain *K. marxianus* var. *marxianus* analyzed here contains 10 chromosomes with an estimated genome size of 14 Mb [8].

Molecular genetic studies of *K. marxianus* allowed the cloning of a few genes of biotechnological interest, such as the genes encoding the inulinase [9], the β-glucosidase [10], the pectin-degrading endopolygalacturonase Epg1p [11], the three glyceraldehyde-3-phosphate dehydrogenases Gap1p, Gap2p, Gap3p [12] and a pyruvate decarboxylase [9].

The present analysis of 2491 Random Sequenced Tags (RSTs) totaling 2363 kb or 17% of the entire genome of *K. marxianus* allowed the identification of ca. 1300 novel nuclear genes encoding proteins, 28 novel nuclear tRNA genes and a novel class of retrotransposon belonging to the *copia* family.

2. Materials and methods

Construction of the genomic library of the *K. marxianus* type strain CBS712 was as described in [13]. Average size of the inserts was tested on 95 clones randomly picked and is 3.5 kb with a standard deviation of 1.6 kb. A total of 1304 inserts were sequenced by GENOSCOPE, corresponding to a total of 2493 RSTs and 2.36 Mb (with 0.62% of ambiguous bases) [14].

3. Results and discussion

3.1. Contigs

Contigs were assembled as in [15]. The largest contig contains 173 sequences, is 9.3 kb in length and corresponds to rDNA repeats. The second one contains 34 sequences, is 6.5 kb in length and corresponds to retrotransposon sequences. The other contigs contain respectively seven sequences (two contigs), six sequences (two contigs), five sequences (three contigs), four sequences (11 contigs), three sequences (68 contigs) and two sequences (291 contigs). 1415 sequences are not part of any contig. Most of the contigs contain at least one gene, and the largest one without any identified gene is 2.6 kb.

3.2. Nuclear ribosomal DNA

All the rDNA sequences of *K. marxianus* are clustered in a 9.3 kb long contig. Comparison of our sequences to the 18S rDNA sequence previously reported [7] revealed only four differences out of 1796 nt. Alignments of 18S, 5.8S, 25S and 5S genes of *S. cerevisiae* and *K. marxianus* show more than 94% of nucleotidic identity. The order of these elements is the same in *K. marxianus* as in *S. cerevisiae*. ITS1 and 2 (Internal Transcribed Spacers), NTS1 and 2 (Non-Transcribed Spacers) and 5′-ETS and 3′-ETS (External Transcribed Spacers) are more divergent. Interestingly, the sequence around the initiation site of the 35S transcript is conserved.

A total of 88 inserts were included into the rDNA contig. For three other inserts, only one out of the two RSTs dis-

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

0014-5793/00/$20.00
PII: S0014-5793(00)02283-3

played similarity to rDNA. Such inserts probably overlap extremities of clusters of rDNA transcription units. This small number of inserts containing rDNA only at one extremity indicate that the rDNA transcription units are probably clustered in a unique locus in the *K. marxianus* genome. Given a genome size of 14×10^6 bp, the estimated number of rDNA repeated units is 106.

3.3. tRNAs genes

We identified 28 tRNA genes from *K. marxianus* by sequence similarity to the complete collection of *S. cerevisiae* tRNA genes. Nucleotide sequence conservation is high. All but one possess the same anticodon as the *S. cerevisiae* homologous tRNA gene. The UUC anticodon of tE(GAA) tRNA in *S. cerevisiae* appears as CUC in *K. marxianus*. In the latter case, an aspartic acid is expected to be charged according to the wobble rule. Comparison of the anticodon stem of all the *K. marxianus* tRNA genes with respect to *S. cerevisiae* revealed four cases of compensatory mutations and one case of either non-compensatory mutation or, most probably, a sequence error. A total of 18 different anticodons were found, which correspond to 14 different amino acids. Despite the low number of total tRNA genes identified by this procedure, no obvious bias concerning the proportions of the different tRNA families was detectable with respect to the pool of tRNA genes from *S. cerevisiae*.

Four out of the 28 tRNAs genes possess an intron. In three cases the *S. cerevisiae* tRNA genes also contain an intron, but intron sequence conservation between the two species was observed only once. In the fourth case, tR(CGT), the *S. cerevisiae* homolog does not possess an intron. On the opposite, a *K. marxianus* tRNA gene (RST AZ0AA014D12T1) does not contain an intron while the *S. cerevisiae* homolog, tF(TTC)i tRNA, does. We conclude that neither the presence nor the sequence of tRNA introns are under selective pressure in either species.

Finally, we observed a case of a direct tandem duplication of a tRNA gene, with a 253 nucleotides long intergenic region, a phenomenon absent from *S. cerevisiae*.

3.4. Retrotransposons

We searched for *K. marxianus* retrotransposons by similarity to *S. cerevisiae* Ty1, 2, 3, 4 and 5 elements. 34 RSTs form a 6.5 kb contig whose consensus sequence is very similar to ORF B of Ty1 and Ty2, but poorly similar to ORFA of Ty1 and Ty2. The levels of similarity to *S. cerevisiae* Ty1 and Ty2 ORFs are nearly the same. We identified putative 385 bp long LTRs present at both ends of the contig, but poorly similar to the *S. cerevisiae* Ty1 and Ty2 LTRs. *K. marxianus* Ty is 5.9 kb in length including both LTRs, the remaining 0.6 kb of the 6.5 kb contig corresponding to adjacent sequences. As in the cases of Ty1 and Ty2 of *S. cerevisiae*, a +1 frameshift is found near position 1500. The CTTAGGC heptamer [16] is also found in this region which strongly suggests that the +1 frameshifting mechanism described in *S. cerevisiae* also exists in *K. marxianus*.

We also identified a few other sequences containing parts of putative Ty elements that could not be included into the contig described above. The RST AZ0AA010D01T1 is very similar to both the LTR and ORF A of a novel Ty but contains a small deletion of ca. 20 nucleotides in the 5′ part of ORF A. Two RSTs contain a LTR flanked by ORF B on their 5′ side and ORF A on their 3′ side suggesting tandem repeats pertinent to Ty elements. However, both sequences display a deletion, because one lacks the 5′ part of ORF A and the other the 3′ part of ORF B. Two additional RSTs contain solo LTRs, sharing 91% of nucleotides identical to LTRs from the contig. At last, three other RSTs contain signatures of LTRs that display internal deletions with respect to the contig ones.

RST AZ0AA003C01D1 that contains such a degenerated solo LTR, is located ca. 80 nucleotides upstream of a tRNA gene similar to tT(ACA) tRNA from *S. cerevisiae*. As in the case of *S. cerevisiae*, this might suggest interactions between some *K. marxianus* retrotransposons and tRNA genes as part of the integration mechanism.

Finally, we identified three other RSTs with similarity to *S. cerevisiae* Ty 1 and 2 but different from the *K. marxianus* Ty contig as described above. This heterogeneity of the retro-

Table 1
Characteristics of *K. marxianus* spliceosomal introns identified by comparison to *S. cerevisiae*

K. marxianus RST	3′ Exon 1	5′ Intron	S1	Branch point	S2	3′ Intron	S	*S. cerevisiae* homologs (intron size)
AZ0AA009C09D1	AAG	GTATGT	42	TACTAAC	32	TAG	81	*YMR225c*[a] (147)
AZ0AA001F01T1	GAA	GTATGT	383	TACTAAC	33	TAG	423	*YNL096c*[a] (345), *YOR096w*[a] (401)
AZ0AA005E01D1	GCG	GTATGT	523	TACTAAC	33	TAG	563	*YNR053c* (531)
AZ0AA008G04D1	ACT	GTATGT	296	TACTAAC	61	TAG	364	*YKL081w* (326)
AZ0AA006A09T1	ACT	GTATGT	300	TACTAAC	68	TAG	375	*YKL081w* (326)
AZ0AA013B12D1	GCG	GTATGT	429	TACTAAC	28	TAG	464	*YDR450w*[a] (435), *YML026c*[a] (401)
AZ0AA005F12D1	TGG	GTATGT	381	TACTAAC	47	TAG	435	*YIL018w*[a] (400), *YFR031ca*[a] (147)
AZ0AA006F01D1	AAC	GTATGT	330	TACTAAC	19	TAG	356	*YLR061w*[a] (389)
AZ0AA010F06T1	CGG	GTATGT	59	TACTAAC	12	TAG	80	*YKL006ca* (141)
AZ0AA012E05T1	CTT	GTATGT	152	TACTAAC	48	TAG	207	*YIL106w* (85)
XAZ0AA002E06T1	AAA	GTATGT	68	TACTAAC	218	TAG	293	*YFL034cb* (114)
AZ0AA003A02D1	nd	nd	nd	TACTAAC	42	TAG	nd	*YAL030w* (113)
AZ0AA004E04T1	nd	nd	nd	TACTAAC	29	TAG	nd	*YGR118w*[a] (320), *YPR132w*[a] (365)
AZ0AA007C03D1	nd	nd	nd	TACTAAC	34	TAG	nd	*YDR064w*[a] (539)
AZ0AA007D03D1	nd	nd	nd	TACTAAC	42	TAG	nd	*YLR287ca*[a] (430), *YOR182c*[a] (411)
AZ0AA007F12D1	nd	nd	nd	TACTAAC	51	TAG	nd	*YGL033w* (70)
AZ0AA010C06D1	nd	nd	nd	TACTAAC	37	TAG	nd	*YDR500c*[a] (389), *YLR185w*[a] (359)
AZ0AA010F06D1	nd	nd	nd	TACTAAC	51	TAG	nd	*YHL001w*[a](398), *YKL006w*[a] (398)

The 5′ splice site is identified by the three terminal nucleotides of the upstream exon (3′ exon 1) and the first six nucleotides of the intron (5′ intron). S1 is the length of the interval from the first position of the intron to the beginning of the TACTAAC box. S2 is the length of the interval from the end of the TACTAAC box to the last position of the intron. The 3′ splice site is identified by the last three nucleotides of the intron (3′ intron). S: *K. marxianus* intron length. All sizes are in nucleotides. nd: not determined.
[a]Ribosomal proteins encoding genes. Note that the *YMR225c* product is mitochondrial.

Table 2
S. cerevisiae gene families of at least seven members having no homologs in the *K. marxianus* RSTs

S. cerevisiae gene family		Number of homologs expected in *K. marxianus*	Functional comments
name	size		
P16.2.f7.1	7	1.5	P-type ATPases involved in Na^+ efflux
P7.4.f7.1	7	1.5	glucosidases; maltases: subtelomeric location.
P8.5.f5.1	8	1.7	aspartyl proteinase of the periplasmic space; pepsin barrier; subtelomeric location.
P26.1.f7.1	9	1.9	subtelomeric location.
P9.2.f9.1	9	1.9	mannosyltransferases
P9.4.f8.1	9	1.9	subtelomeric location.
P11.2.f7.1	11	2.3	*FLO* family; subtelomeric location.
P26.1.f13.1	17	3.6	cell wall biogenesis and architecture; subtelomeric location.
P24.1.f23.1	23	4.8	seripauperin family; subtelomeric location.

The number of *K. marxianus* genes expected in the RSTs was estimated considering that the minimal number of *K. marxianus* gene per *S. cerevisiae* gene is 0.21 (1301 *K. marxianus* ORFs out of 6213 *S. cerevisiae* ORFs). Gene contents of these families are presented in [15].

transposon sequences could reveal either inactive copies that accumulated mutations or the existence of several classes of retrotransposons in *K. marxianus*.

In conclusion, *K. marxianus* possesses at least two classes of retrotransposons related to the Ty1 and Ty2 families of *S. cerevisiae*. Given a genome size of 14×10^6 bp, the most frequent class must be represented by ca. 34 full length retrotransposons in the entire genome and ca. 30 solo LTRs. No Ty 3, 4 or 5 homologs could be detected.

3.5. Mitochondrial sequences

Only two sequences similar to the *S. cerevisiae* 21S mtRNA were encountered. Such an under-representation was not observed in the two other *Kluyveromyces* species studied in this project [17,18], and likely results from loss of mtDNA during the DNA extraction procedure [13].

3.6. Genetic code and codon usage

The analysis presented in [15] confirms that *K. marxianus* utilizes the universal genetic code. RSCU values (Relative Synonymous Codon Usage [19]) were computed for *K. marxianus* and compared to the *S. cerevisiae* ones. Interestingly, differences between RSCU values of each codon pair corresponding to F, Y, N and K, namely TTT/TTC, TAT/TAC, AAT/AAC and AAA/AAG, were found to be inverted between *S. cerevisiae* and *K. marxianus*. In these four cases, the codon ending by a T or a A has the lower RSCU value in *K. marxianus*, and conversely in *S. cerevisiae*.

3.7. K. marxianus protein coding ORFs

We identified a total of 1301 to 1441 *K. marxianus* protein coding genes on the basis of similarity to *S. cerevisiae* protein coding genes. The determination on the number comes from the fact that several RSTs can be similar to non-overlapping

Table 3
S. cerevisiae genes having several *K. marxianus* homologs

S. cerevisiae ORF	Family	Size	*K. marxianus* homologs		*S. cerevisiae* functional comments
			min	max	
(a)					
YEL046c (*GLY1*)	singleton	1	2	2	threonine aldolase, required for glycine biosynthesis
YJL199c	singleton	1	2	2	unknown function
YKL080w (*VMA5*)	singleton	1	2	2	vacuolar H(+)-ATPase (V-ATPase) hydrophilic subunit (subunit C)
YKL215c	singleton	1	2	3	similar to *Pseudomonas* hydantoinases hyuA-hyuB
YKL217w (*JEN1*)	singleton	1	2	2	lactate-proton symporter
YOL060c (*AMI3*)	singleton	1	2	2	protein required for normal mitochondrial structure
YJL212c (*OPT1*)	P2.1.f2.1	2	2.5	3.8	member of the oligopeptide transporter (OPT) family
YPR194c	P2.1.f2.1	2	1.5	2.3	member of the OPT family
YLR117c (*SFY3*)	P2.264.f2.1	2	2	3	pre-mRNA splicing factor
(b)					
YBR025c	P2.382.f2.1	2	1.2	1.2	member of the GTP-binding protein family
YCR010c	P3.53.f3.1	3	1.5	1.5	unknown function
YNR002c	P3.53.f3.1	3	1.5	1.5	unknown function
YBL064c	P3.76.f3.1	3	2	2	mitochondrial thiol peroxidase of the 1-Cys family
YBR245c	P17.1.f16.1	17	2	3	similar to SNF2/SWI2 DNA binding regulatory protein
YER047c (*SAP1*)	P21.1.f17.1	21	2	3	similar to Yta6p of 26s proteasome
YGR040w (*KSS1*)	P108.1.f12.1	12	2	2	Ser/Thr protein kinase of the MAP kinase family
YHR050w (*SMF2*)	P3.43.f3.1	3	2	3	probable manganese transporter
YJR152w (*DAL5*)	P10.2.f3.1	3	2	2	allantoate permease
YKL049c (*CSE4*)	P3.62.f3.1	3	2	2	similar to histone H3
YNL125c	P4.39.f4.1	4	2	2	similar to mammalian monocarboxylate transporters, member of the monocarboxylate porter (MCP) family of the MFS

Min and max determinations of the numbers of *K. marxianus* homologs of *S. cerevisiae* genes and gene families are presented in [27].
(a) Cases where this leads to an over-representation of the corresponding *K. marxianus* gene families with respect to the *S. cerevisiae* homologs.
(b) Cases where this does not lead to an over-representation of the corresponding *K. marxianus* gene families with respect to the *S. cerevisiae* homologs.

parts of the same *S. cerevisiae* gene [15]. A total of 1546 different genes from *S. cerevisiae* were classified in the 'o' (1105) and 'oo' (475) matches (or both, 34). These genes are distributed randomly on the 16 *S. cerevisiae* chromosomes (data not shown). The mean percent of identity between sequences from the two species is 61.1 and the median is 58; which corresponds to 75% (mean) and 74% (median) of amino acid similarity (data not shown).

3.8. Spliceosomal introns

We only searched for spliceosomal introns in the 46 *K. marxianus* genes that are homologous to 57 *S. cerevisiae* genes containing spliceosomal introns as defined in Yeast Intron Database YIDB [20]. In this set of *K. marxianus* genes, the following consensus were searched: GTATGT for the 5′ splice site; TACTAAC for the branch point and YAG for the 3′ splice site. A total of 18 *K. marxianus* intron-containing genes were identified (Table 1). The complete sequences of introns were determined in 11 cases showing that, on average, intron length is similar in the two species despite individual variations. Intron positions are conserved between *S. cerevisiae* and *K. marxianus* genes. Sixteen other *K. marxianus* genes do not contain introns with the above consensus sequences at similar positions to *S. cerevisiae* genes or elsewhere. The sequence similar to the region overlapping the intron site in *S. cerevisiae* was not available for the remaining 12 or 13 genes. In these cases, no intron was detected elsewhere in the available sequence. These results suggest that intron splicing in *K. marxianus* uses the same consensus sequences as in *S. cerevisiae*.

Table 4
List of homologs of the *K. marxianus* RSTs that do not have a validated homolog in *S. cerevisiae*

Organism	AC	Functional comments
Archaea		
Archaeoglobus fulgidus	AAB89774	mitochondrial benzodiazepine receptor/sensory transduction protein, Af1475
Bacteria		
Escherichia coli	C64974	hypothetical protein b2076
Bacillus subtilis	CAB15751	similar to ABC transporter, YwjA
	CAB15679[a]	unknown, ywnB
Campylobacter jejuni	CAB73453	putative iron/ascorbate-dependent oxidoreductase, Cj1199
Mycobacterium tuberculosis	O06222	UDP-*N*-acetylmuramoylalanine-D-glutamate ligase, murD, Rv2155c
Ascomycetes		
K. marxianus	P07337	β-glucosidase (EC 3.2.1.21)
	S34953	transcription initiation factor IIB
Kluyveromyces lactis	P00723	β-galactosidase (EC 3.2.1.23), LAC4
	P40418	regulatory protein, SWI6
	P08657[a]	lactose regulatory protein, LAC9
	P49374[a]	high affinity glucose transporter, HGT1
Saccharomyces kluyveri	Q02342[a]	cell division control protein 25, CDC25
	O13377	PET122 protein precursor
Pichia pastoris	Q01961	peroxisome assembly protein, PAS10
Debaryomyces occidentalis	P50505	high affinity potassium transporter, HAK1
Candida albicans	P87218	sorbitol utilization protein SOU2
	P87219	sorbitol utilization protein SOU1
Fusarium solani	P38364	pistatin demethylase (EC 1.14.-.-) of the cytochrome P450 family
Schizosaccharomyces pombe	O13965	hypothetical protein, SPAC24C9.05C
	Q09875	hypothetical protein, SPAC12G12.12
	CAA21255	conserved hypothetical protein, SPBC1709.16C
	CAA19111	acetamidase, SPCC550.07
	CAA21876[a]	putative major facilitator protein, SPCC622.20C
	CAA19168[a]	MFS transporter of unknown specificity, SPBC530.02
	CAB52163	hypothetical protein, SPAC8F11.02C
	Q09329	MLO2 protein
	CAA93795[a]	putative homoserine o-acetyltransferase, SPAC19G10.13
	CAB40174	putative D-amino acid oxidase, SPCC1450.07C
	CAA20729	MFS efflux transporter of unknown specificity, SPBC4F6.09
	CAA22658	similar to carboxylesterase-lipase, SPCC417.12
	CAB16270	hypothetical zinc finger protein, SPAC2F3.16
	P78771	
Other eukaryotes		
Arabidopsis thaliana	CAB79481	putative protein, At4g26260
	AAD23730	putative copper amine oxidase, At2g42490
	CAB80069[a]	metal-transporting P-type ATPase, At4g33520
	AAC64301	putative pirin protein, At2g43120
	AAD26955[a]	putative sugar transporter, At2g16130
Nicotiana tabacum	P49098[a]	cytochrome B5
Caenorhabditis elegans	AAF99922	hypothetical protein, F29B9.4
Ovis aries	P47843	glucose transporter type 3 protein
Rattus norvegicus	P97521[a]	mitochondrial carnitine/acyl carnitine carrier protein

AC: accession number.
In all cases, only one homolog is present in *K. marxianus*. Note that P07337 (β-glucosidase EC 3.2.1.21) comes from strain *ATCC12424* different from the *CBS712* strain used in this study. Note that the presence of a gene similar to *S34953* (transcription initiation factor IIB from *K. marxianus* var. *lactis*) results from the comparison to GenBank, this entry being absent from GPROTEOME.
[a]The *K. marxianus* homolog exhibits sequence similarity with a *S. cerevisiae* protein that has not been validated (see [15]).

3.9. Conservation of 'Ascomycetes specific' genes

Among the 1546 genes from *S. cerevisiae* that possess an ortholog in *K. marxianus*, only 351 (22.7%) had homologs in Ascomycetes or no homolog at all, while such genes represented 40% of the *S. cerevisiae* genome. These 351 *S. cerevisiae* genes have now been classified as 'Ascomycete specific' genes. The under-representation of this class of genes in the *K. marxianus* RSTs is common to the 13 studied species and is discussed in [21].

3.10. Functional classification

In the absence of experimental data on the functions of the many *K. marxianus* genes that have been identified in this work, we have examined whether the distribution of their *S. cerevisiae* homologs in the defined functional classes was biased or not [22]. We found that, within the limits of statistical significance, no functional bias appears in *K. marxianus* compared to *S. cerevisiae*. A more complete analysis of this problem is presented in [22].

3.11. Paralogous gene families

The study of the *K. marxianus* paralogous gene families performed in [23] shows that the degree of gene redundancy in this species is similar to that of *S. cerevisiae*. However, some example of gene families that vary in size between these two species exist and are discussed below.

Of the nine *S. cerevisiae* families of seven members or more having no homologs in our *K. marxianus* data set (Table 2), one of them, P7.4.f7.1 encodes maltases [24] consistent with the inability of *K. marxianus* to ferment or assimilate maltose [1]. Members of this family are subtelomeric in *S. cerevisiae*, as is the case for seven of the nine families listed in Table 2. Their absence from *K. marxianus* (also noted for *Kluyveromyces thermotolerans*, see [18]) may indicate their recent amplification in *S. cerevisiae* (consistent with the low sequence divergence between members of such families) or loss in *Kluyveromyces* species.

Conversely, some gene families are over-represented in *K. marxianus* compared to *S. cerevisiae* (Table 3). Six singletons in *S. cerevisiae* have at least two homologs in *K. marxianus* and three families of two members (P2.1.f2.1 and P2.264.f2.1 encoding oligopeptide transporters [25] and proteins involved in RNA splicing [26]) have at least four and three homologs in *K. marxianus*, respectively.

Finally, the two direct tandem duplications of *S. cerevisiae* genes involving *YOR285w* and *YOR286w* on one side, and *YDR342c* and *YDR343c* on the other side, also exist in the same orientation in *K. marxianus*. Interestingly, *YMR169c* and *YMR170c* that reflect a direct duplication in *S. cerevisiae*, have homologs in reverse orientation in *K. marxianus*. The conservation of such tandem duplications is remarkable given the rather large phylogenic distance between *K. marxianus* and *S. cerevisiae*. It suggests that the duplications existed before the separation of the two species.

3.12. Identification of ORFs coding proteins by RSTs sequence comparison to non-S. cerevisiae proteins

In order to discover genes in *K. marxianus* that are absent from *S. cerevisiae*, we compared our data set to GPROTEOME as defined in [15]. Forty-two or 43 additional ORFs from *K. marxianus* were found (Table 4). Among these new ORFs, 21 encode proteins similar to proteins of known functions. Note the presence of the β-galactosidase and the β-glucosidase, two proteins absent from *S. cerevisiae*, that allow *K. marxianus* to use lactose and cellobiose as carbon sources. Also note the presence of eight putative transport proteins, whose similarity to *S. cerevisiae* transport proteins was considered as non-significant. None of the above genes was identified before in *K. marxianus* var. *marxianus CBS712*.

3.13. Synteny conservation and chromosomal maps

Out of 454 syntenic pairs of *K. marxianus* genes identified, 226 are also syntenic in *S. cerevisiae* (49.8%). In addition, we identified 37 transchromosomal series in *K. marxianus*. Such structures correspond to series of genes that are adjacent in *K. marxianus* but whose homologs are distributed in two different chromosomal regions in *S. cerevisiae*. As discussed in [27], such series may reflect the chromosomal map of the common ancestor of *S. cerevisiae* and *K. marxianus*.

Acknowledgements: This work was supported in part by a BRG Grant (ressources génétiques des microorganismes, No 11-0926-99). We thank our colleagues of Unité de génétique moléculaires des levures, especially Cécile Fairhead for fruitful discussions and Fredj Tekaia for the use of bioinformatic tools. B.D. is a member of Institut Universitaire de France.

References

[1] Lachance, M.A. (1998) in: The yeasts. A taxonomic study (Kurtzman, C.P. and Fell, J.W., Eds.), pp. 227–247, Elsevier, Amsterdam.
[2] Listemann, H., Schulz, K.D., Wasmuth, R., Begemann, F. and Meigel, W. (1998) Mycoses 41, 343–344.
[3] Abu-Elteen, K.H., Abderrahman, S. and Sallal, A.K. (1997) Cytobios 90, 41–45.
[4] Johannsen, E. (1980) Antonie Van Leeuwenhoek 46, 177–189.
[5] Vaughan-Martini, A. and Martini, A. (1987) Int. J. Syst. Bacteriol. 37, 380–385.
[6] Fuson, G.B., Presley, H.L. and Phaff, H.J. (1987) Int. J. Syst. Bacteriol. 37, 371–379.
[7] Cai, J., Roberts, I.N. and Collins, M.D. (1996) Int. J. Syst. Bacteriol. 46, 542–549.
[8] Belloch, C., Barrio, E., Garcia, M.D. and Querol, A. (1998) Yeast 14, 1341–1354.
[9] Holloway, P. and Subden, R.E. (1993) Curr. Genet. 24, 274–277.
[10] Raynal, A. and Guerineau, M. (1984) Mol. Gen. Genet. 195, 108–115.
[11] Siekstele, R., Bartkeviciute, D. and Sasnauskas, K. (1999) Yeast 15, 311–322.
[12] Fernandes, P.A., Sena-Esteves, M. and Moradas-Ferreira, P. (1995) Yeast 11, 725–733.
[13] Blandin, G., Llorente, B., Malpertuy, A., Wincker, P., Artiguenave, F. and Dujon, B. (2000) FEBS Lett. 487, 76–81 (this issue).
[14] Artiguenave, F. et al. (2000) FEBS Lett. 487, 13–16 (this issue).
[15] Tekaia, F. et al. (2000) FEBS Lett. 487, 17–30 (this issue).
[16] Voytas, D.F. and Boeke, J.D. (1993) Trends Genet. 9, 421–427.
[17] Bolotin-Fukuhara, M. et al. (2000) FEBS Lett. 487, 66–70 (this issue).
[18] Malpertuy, A., Llorente, B., Blandin, G., Artiguenave, F., Wincker, P. and Dujon, B. (2000) FEBS Lett. 487, 61–65 (this issue).
[19] Lloyd, A.T. and Sharp, P.M. (1993) Yeast 9, 1219–1228.
[20] Lopez, P.J. and Seraphin, B. (2000) Nucleic Acids Res. 28, 85–86.
[21] Malpertuy, A. et al. (2000) FEBS Lett. 487, 113–121 (this issue).
[22] Gaillardin, C. et al. (2000) FEBS Lett. 487, 134–149 (this issue).
[23] Llorente, B. et al. (2000) FEBS Lett. 487, 122–133 (this issue).
[24] Tabata, S., Ide, T., Umemura, Y. and Torii, K. (1984) Biochim. Biophys. Acta 797, 231–238.
[25] Paulsen, I.T., Sliwinski, M.K., Nelissen, B., Goffeau, A. and Saier Jr., M.H. (1998) FEBS Lett. 430, 116–125.
[26] Chung, S., McLean, M.R. and Rymond, B.C. (1999) RNA 5, 1042–1054.
[27] Llorente, B. et al. (2000) FEBS Lett. 487, 101–112 (this issue).