Genomic Exploration of the Hemiascomycetous Yeasts: 15. *Pichia sorbitophila*

Jacky de Montigny[a], Catherine Spehner[a], Jean-Luc Souciet[a], Fredj Tekaia[b], Bernard Dujon[b], Patrick Wincker[c], François Artiguenave[c], Serge Potier[a,*]

[a] *Laboratoire de Génétique et Microbiologie, UPRES-A 7010 ULP/CNRS, Institut de Botanique, 28 rue Goethe, F-67000 Strasbourg, France*
[b] *Unité de Génétique Moléculaire des Levures, Institut Pasteur/CNRS, département des Biotechnologies, Institut Pasteur, 25 rue du Docteur Roux, F-75724 Paris Cedex 15, France*
[c] *Genoscope, Centre National de Séquençage, 2 rue Gaston Crémieux, P.O. Box 191, F-Evry Cedex, France*

Received 3 November 2000; accepted 9 November 2000

First published online 29 November 2000

Edited by Horst Feldmann

Abstract This paper reports the genomic analysis of the strain CBS7064 of *Pichia sorbitophila*, a homothallic diploid yeast. We sequenced 4829 random sequence tags of about 1 kb and compared them to the *Saccharomyces cerevisiae* gene products. Approximately 1300 nuclear genes, 22 tRNAs, the rDNA locus, and six mitochondrial genes have been identified. The analysis of the rDNA genes has permitted to classify this organism close to the *Candida* species. Accession numbers from AL414896 to AL419724 at EMBL databank.

Key words: Comparative genomics; Random sequence tag; Hemiascomycete; *Saccharomyces cerevisiae*

1. Introduction

The hemiascomycete *Pichia sorbitophila* was isolated as a contaminant of a 70% sorbitol solution [1]. This organism has been described as a homothallic diploid [2]. It has not yet been classified in any phylogenetic tree. Interest in *P. sorbitophila* arose from its physiological properties distinct from *Saccharomyces cerevisiae*. In fact, this yeast is highly resistant to osmotic stress in general and to salt stress (4 M NaCl) in particular. These properties seem to be due to its ability to actively co-transport with protons glycerol, the major solute this yeast accumulates, in order to maintain osmotic balance [3,4].

Little is known about its genomic content. Recent approaches by pulsed field gel electrophoresis led us to propose that this yeast contains seven chromosomes with sizes ranging from 1 to 2.9 Mbp corresponding to an estimated total of 13.9 Mbp [5]. Two genes have been sequenced thus far: the *TIM10 S. cerevisiae* homolog, a putative component of the mitochondrial protein import system [6] and the equivalent to the *LYS2* gene [7]. Three other ORFs with unknown function are present in the EMBL/DDJB databank (accession number Y18559).

This paper reports the genomic analysis of strain CBS7064 by exploring the sequence of 4829 random sequence tags (RSTs) of about 1 kb and comparing them to the yeast gene products as described in [8]. Approximately 1300 new nuclear gene products, 22 tRNAs, the rDNA locus and six mitochondrial proteins were identified. The sequence of the rDNA has permitted to propose a phylogenetic position of this organism close to the *Candida* species [9].

2. Materials and methods

2.1. Strain

The *P. sorbitophila* strain used is CBS7064.

2.2. DNA extraction, library construction, sequencing, RST analysis and annotations

The protocols and the modifications described in [8,10–12] were followed with a minor modification for genomic DNA extraction: the DNA was recovered with a sterile glass rod instead of a classical isopropanol precipitation.

3. Results and discussion

3.1. Characterization of the RST

4829 RSTs which represent a total of 4 036 247 nucleotides were sequenced by Genoscope with less than 1.5% ambiguity. The RSTs have an average length of 836 nucleotides ranging from 40 to 1209 nucleotides (standard deviation, 189 nt). A total of 2159 inserts representing 4318 RSTs (95% of all RSTs) were sequenced on both ends and 90 had overlapping sequences. The remaining 511 RSTs (5%) were sequenced only at one end.

The DNA library sent to Genoscope contained 3936 clones representing a potential of 7872 RSTs; 4829 sequences representing 61.3% of the total expected RSTs were obtained and 3044 RSTs (38.6%) could not be sequenced. The experimental determination by *Eco*R1 digestion of the insert size from 96 random clones revealed that 28 plasmids (30%) were empty or had lost part of the insert or plasmid sequences including the hybridization site of the sequencing primer. These plasmid rearrangements pointing out the instability of the *P. sorbitophila* DNA in *Escherichia coli* explain why around 40% of the sequences could not be obtained. This instability increased after three or four successive cultures even on selective media.

The *Eco*R1 analysis showed that the 68 remaining plasmids carry inserts between 1 and 9 kb. The average size is 4.4 kb with a standard deviation of 1.7 kb. These results are in agreement with the expected size of the library inserts.

*Corresponding author. Fax: (33)-3-88 35 84 84.
E-mail: potier@gem.u-strasbg.fr

0014-5793/00/$20.00
PII: S0014-5793(00)02286-9

A total of 2937 RSTs could be grouped in 970 contigs of 2–123 RSTs with a maximum size of 8.42 kb. The majority of the contigs (63%) were composed of two RSTs. The largest contig of 123 RSTs corresponded to rDNA sequences.

3.2. Proteins with orthologues in S. cerevisiae

Open reading frames have been identified in the six possible frames, translated into putative proteins, compared to *S. cerevisiae* database as described in [8] and validated after examination of the sequence position on RSTs. The BLATSX results on the 4829 RSTs showed that 1506 translated RSTs (31.2%) had only one unambiguous match with a *S. cerevisiae* protein, 60 (1.2%) had at least two unambiguous 'o' matches, 373 (7.7%) had an ambiguous 'oo' match and 26 (0.05%) had composite matches 'o' and 'oo'. Finally, 2864 RSTs representing 59.2% were not validated including 223 open reading frames which did not show any match.

The unambiguous 'o' matches identified the *P. sorbitophila* homologs for 1070 *S. cerevisiae* ORFs. The ambiguous 'oo' RSTs analysis revealed 523 additional sequences which belong to 220 gene families. The percentages of amino acid identity/similarity were 51 and 67, respectively.

The position of the 1070 *S. cerevisiae* orthologues of *P. sorbitophila* were plotted onto the *S. cerevisiae* chromosomes [9]. The matches are equally distributed on each chromosome except for the right arm of chromosome I for which no match was found. The random distribution of ORFs along the chromosomes indicates that the library used in this work as well as the number of random RSTs examined are representative of the *P. sorbitophila* genome. The absence of matches with the right arm of chromosome I could be interpreted as an absence of this region in the *P. sorbitophila* genome and of recent extension of *S. cerevisiae* chromosome I.

3.3. tRNAs

All RST matching with a tRNA sequence were annotated as 'oo' since they all belong to families. The comparison with the 42 tRNA families in *S. cerevisiae* defined 19 families in *P. sorbitophila*. The minimal number of tRNA genes for a given family was obtained by considering the number of different contigs matching with the same tRNA gene. The number of 23 different tRNA genes could be assessed by this method. Considering that the library covers one third of the genome, the putative number of tRNA genes in *P. sorbitophila* should amount to 70 which is a very low compared to the 274 tRNA genes found in *S. cerevisiae* (Table 1).

Several interesting features emerged. First, two RSTs were associated to two tRNA genes, tRNA-Asn/tRNA-Thr and tRNA-Val/tRNA-Gly. Second, a *P. sorbitophila* tRNA-Trp gene had an intron in *S. cerevisiae* which was absent in *P. sorbitophila*. Third, a tRNA gene with an AAG anticodon not existent in *S. cerevisiae* was identified in *P. sorbitophila*. It corresponds to a tRNA-Leu which reveals 85% identity with the tRNA-Leu of *S. cerevisiae* carrying the GAG anticodon. Further, a tRNA-Glu gene was identified with a CUC anticodon which reveals 84% identity with a tRNA-Glu gene carrying a UUC anticodon but no significant identity with the one carrying the CUC anticodon. This difference in the anticodon sequences most likely is not due to a sequencing error because two different RSTs carrying the same tRNA had an identical sequence.

Determination of the genetic code carried out in this project [8] pointed out that in *P. sorbitophila* the CUG codon is not used to specify serine instead of leucine. Similar observations were previously made in several *Candida* species [13,14] and *Debaryomyces hansenii* [15]. Unfortunately, we have not been able to identify the corresponding tRNA gene suggesting that it is present at a low copy number.

3.4. rDNA

A total of 161 RSTs matched significantly with genomic rDNA sequences of *S. cerevisiae*, 123 of these RSTs constitute a contig of 8368 nucleotides. This contig reveals a repeat of 8235 nucleotides which contains sequences encoding 25S, 18S,

Table 1
Nuclear tRNA genes detected in *P. sorbitophila*

	S. cerevisiae anticodon	*P. sorbitophila* anticodon	Intron *P. sorbitophila.*	Intron *S. cerevisiae*	Number of *S. cerevisiae* tRNA genes in the corresponding family	Number of RSTs matching with the tRNA gene	Number of contigs matching with the tRNA gene	Minimal number of *P. sorbitophila* tRNA
tRNA-ALA2	UGC	UGC			5	1	1	1
tRNA-CYS	GCA	GCA			4	1	1	1
tRNA-ASP	GUC	GUC			15	3	1	1
tRNA-GLU3	UUC	**CUC**			13	2	1	1
tRNA-GLY2	UCC	UCC			3	2	1	1
tRNA-GLY1	GCC	GCC			16	3	2	2
tRNA-HIS2	GUG	GUG			7	2	1	1
tRNA-LYS 2	UUU	UUU	27	23	7	3	1	1
tRNA-LEU5	GAG	**AAG**			1	1	0	1
tRNA-LEU4	UAA	UAA			7	2	1	1
tRNA-ASN	GUU	GUU			10	5	2	2
tRNA-GLN	UUG	UUG			1	4	1	2
tRNA-GLN	UUG	UUG			7	1	1	1
tRNA-SER2	AGA	AGA			11	4	1	1
tRNA-THR	UGU	UGU			1	2	1	1
tRNA-THR1	AGU	AGU			11	2	2	2
tRNA-VAL1	AAC	AAC			13	4	1	1
tRNA-TRP	CCA	CCA	0	34	6	1	1	1
tRNA-TYR	GUA	GUA	15	14	8	2	1	1
								Total: 23

Table 2
Non-*S. cerevisiae* homologs of *P. sorbitophila*

Organism	Accession number	Functional comments
Ascomycetes		
Aspergillus niger	Q92406	NADH ubiquinone oxido-reductase 51 kDa subunit
C. albicans	P12461	thymidylate synthase (EC 2.1.1.45)
C. albicans	P43062	G_1/S-specific cyclin CLN2
C. albicans	P46599	serine threonine protein kinase STE7 homolog
C. albicans	P56553	cell growth protein
C. albicans	Q04802	glucosamine 6 phosphate isomerase (EC 5.3.1.10)
C. tropicalis	Q12600	SIS2 protein (halotolerance protein HAL3)
Emericella nidulans	P48777	purine permease
N. crassa	P24918	NADH ubiquinone oxidoreductase precursor 78 kDa
N. crassa	P40915	NADH-ubiquinone oxido-reductase 24 kDa subunit
N. crassa	P48467	kinesin heavy chain
Pichia angusta	P12807	peroxisomal copper amine oxidase (EC 1.4.3.6)
P. canadensis	P48917	NADH-ubiquinone oxido-reductase chain 4 (EC 1.6.5.3)
Schizosaccharomyces pombe	T41665	putative dipeptidase
S. pombe	Q10088	putative agmatinase precursor (EC 3.5.3.11)
S. pombe	Q10449	hypothetical 57.2 kDa sp protein
Other eukaryotes		
Rattus norvegicus	P14408	fumarate hydratase mitochondrial precursor

5.8S and 5S in the same organization as in *S. cerevisiae,* sharing 89, 92, 94 and 89% of identity, respectively. In contrast, the intergenic regions within the rDNA locus show no similarity with those of *S. cerevisiae* [9].

3.5. Transposons

Transposable elements are ubiquitous in eukaryotes. In *S. cerevisiae*, 52 full-length retrotransposable elements were found as well as 268 solo long terminal repeats [16], which had been classified into five types: Ty1, 2, 4 and 5, which belong to the copia-like family and Ty3 which is a member of the gypsy-like family [17].

All RSTs and contigs were compared with the *S. cerevisiae* Ty element sequences but no significant matches were obtained. Comparisons of all *P. sorbitophila* putative ORFs with the protein sequences TyA and TyB did not reveal significant similarities. Since in *S. cerevisiae* the majority of the Ty elements are located in the vicinity of tRNA genes [18], the RSTs containing tRNA genes were carefully checked for retrotransposons though without success.

These results could therefore mean that transposon-like elements are present at a very low copy number in *P. sorbitophila*. Indeed, as we have identified roughly one third of the genome, one would expect to find a significant number of RSTs corresponding to Ty-like elements if their number would be in the same range as in *S. cerevisiae*.

At least three other explanations should be considered. The first were that Ty-like elements are present in *P. sorbitophila* but have considerably diverged from those in *S. cerevisiae*. The second would have to imply that *P. sorbitophila* does not contain retrotransposons at all and therefore constitutes a notable exception among eukaryotes. The third assumption would imply a cloning bias, due for instance to instability of *P. sorbitophila* transposons in *E. coli*.

3.6. Mitochondrial DNA

Mitochondrial DNA of *P. sorbitophila* was not comprised in a unique large contig. The BLASTX comparison of the 4829 RSTs to the *S. cerevisiae* mitochondrial protein database revealed 10 RSTs matching significantly and corresponding to the mitochondrial mRNA maturases BI2, BI3 and BI4, the cytochrome *c* oxidase subunit II and the membrane-associated ATPase subunit 9. No mitochondrial tRNA or rDNA genes could be identified by a BLASTN analysis.

If only one RST of an insert matches with a mitochondrial sequence, while the second RST from the same insert shows no match, one may presume that the second RST also monitors the *P. sorbitophila* mitochondrial genome. This argument was used when analyzing the contigs including the 10 RSTs identified as belonging to mitochondrial DNA. By this procedure we have identified 32 further putative mitochondrial RSTs. The total of 42 RSTs were included in eight contigs representing 7.8 kb of mitochondrial DNA. This by excess evaluation estimates the size of the entire mitochondrial genome to be around 24 kb since one third of the genome was covered by the library. This is clearly less than the 86 kb of the *S. cerevisiae* mitochondrial genome [19].

This notion together with the obvious absence of tRNA and rRNA genes suggests that the mitochondrial DNA from *P. sorbitophila* is largely different from the *S. cerevisiae* one. None the less, we cannot exclude that the underrepresentation of mitochondrial RSTs could be due to differential mitochondrial DNA extraction or to instability of some plasmids in the *E. coli* library as discussed above.

3.7. Proteins with no orthologues in S. cerevisiae and phylogenetic classification

The occurrence of yeast specific genes not found in *S. cerevisiae* has already been described for several yeast species [20,21]. We therefore performed a general comparison of the *P. sorbitophila* RSTs with several entirely sequenced genomes and the SwissProt database depleted of the *S. cerevisiae* genes and their homologs [8].

We retained 17 RSTs whereof 16 matched to ascomycete genes and to *Candida albicans* or *Neurospora crassa* in particular (Table 2). Three *P. sorbitophila* RSTs showed significant matches with NADH ubiquinone oxidoreductase subunit genes of *N. crassa* and *Pichia canadensis* but not with those of *S. cerevisiae*. For three RSTs, the matches were significantly higher with *CLN7* and *HST7* of *C. albicans* and *HAL3* of *Candida tropicalis* than with the equivalent genes of *S. cerevisiae*.

These observations are consistent with the phylogenetic position of *P. sorbitophila* closer to the *Candida* species as suggested by the cladogram established from 18S and 25S rRNA sequence comparisons [9].

Acknowledgements: We thank Marie-Laure Straub and Yves Tourrette for technical assistance. This work was supported by Grant No 11-0926-99 from BRG (Bureau des Ressources Génétiques). The yeast genetics department of the UPRES-A 7010 is a member of the 'Génopole Alsace-Lorraine'. B.D. is a member of Institut Universitaire de France.

References

[1] Rodrigues de Miranda, L., Appel, K.R. and Seyfarth, H. (1980) Antonie Van Leeuwenhoek 46, 157–159.
[2] Soares de Oliveira, R.P. (1995) Thesis Universidade Do Minho (Braga, Portugal).
[3] Lages, F. and Lucas, C. (1995) Yeast 11, 111–119.
[4] Oliveira, R.P., Lages, F. and Lucas, C. (1996) FEMS Microbiol. Lett. 142, 147–153.
[5] Sychrova, H., Braun, V., Potier, S. and Souciet, J.L., Yeast, in press.
[6] Kayingo, G., Potier, S., Hohmann, S. and Prior, B.A. (2000) Yeast 16, 589–596.
[7] Bleykasten-Grosshans, C., Prior, C. and Potier, S. (2000) Yeast, in press.
[8] Tekaia, F., Blandin, G., Malpertuy, A., Llorente, B., Durrens, P. et al., FEBS Lett. 487, 17–30 (this issue).
[9] Souciet, J.L., Aigle, M., Artiguenave, F., Blandin, G. Bolotin-Fukuhara, M. et al., FEBS Lett. 487, 3–12 (this issue).
[10] Blandin, G., Llorente, B., Malpertuy, A., Wincker, P., Artiguenave, F. and Dujon, B. (2000) FEBS Lett. 487, 31–36 (this issue).
[11] Artiguenave, F., Wincker, P., Brottier, P., Duprat, S., Jovelin, F. et al., FEBS Lett. 487, 13–16 (this issue).
[12] de Montigny, J., Straub, M.L., Potier, S., Tekaia, F., Dujon, B. et al., FEBS Lett. 487, 52–55 (this issue).
[13] Kawaguchi, Y., Honda, H., Taniguchi-Morimura, J. and Iwasaki, S. (1989) Nature 341, 164–166.
[14] Suzuki, T., Ueda, T., Yokogawa, T., Nishikawa, K. and Watanabe, K. (1994) Nucleic Acids Res 22, 115–123.
[15] Lépingle, A., Casaregola, S., Bon, E., Neuvéglise, C., Nguyen, V.H. et al., FEBS Lett. 487, 82–86 (this issue).
[16] Hani, J. and Feldmann, H. (1998) Nucleic Acids Res. 26, 689–696.
[17] Jordan, I.K. and McDonald, J.F. (1999) Genetics 151, 1341–1345.
[18] Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A. and Voytas, D.F. (1998) Genome Res. 8, 464–478.
[19] Foury, F., Roganti, T., Lecrenier, N. and Purnelle, B. (1998) FEBS Lett. 440, 325–331.
[20] Ozier-Kalogeropoulos, O., Malpertuy, A., Boyer, J., Tekaia, F. and Dujon, B. (1998) Nucleic Acids Res. 23, 5511–5524.
[21] Barth, G. and Gaillardin, C. (1997) FEMS Microbiol. Rev. 19, 219–237.