FEBS Letters 487 (2000) 76–81 FEBS 24377

Genomic Exploration of the Hemiascomycetous Yeasts: 13. *Pichia angusta*

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

[a] *Unité de Génétique Moléculaire des Levures (URA 2171 CNRS, UFR 927 Univ. P. and M. Curie), Département des Biotechnologies, Institut Pasteur, 25 rue du Dr Roux, F-75724 Paris Cedex 15, France*
[b] *Genoscope, 2 rue Gaston Crémieux, F-91000 Evry, France*

Received 3 November 2000; accepted 9 November 2000

First published online 29 November 2000

Edited by Horst Feldmann

Abstract As part of a comparative genomics project on 13 hemiascomycetous yeasts, the *Pichia angusta* type strain was studied using a partial random sequencing strategy. With coverage of 0.5 genome equivalents, about 2500 novel protein-coding genes were identified, probably corresponding to more than half of the *P. angusta* protein-coding genes, 6% of which do not have homologs in *Saccharomyces cerevisiae*. Some of them contain one or two introns, on average three times shorter than those in *S. cerevisiae*. We also identified 28 tRNA genes, a few retrotransposons similar to *Ty5* of *S. cerevisiae*, solo long terminal repeats, the whole ribosomal DNA cluster, and segments of mitochondrial DNA. The *P. angusta* sequences were deposited in EMBL under the accession numbers AL430961 to AL436044.

Key words: Genomic library; Retrotransposon; Intron

1. Introduction

Pichia angusta (formally designated *Hansenula polymorpha*) is a methylotrophic yeast able to utilize methanol as the sole carbon and energy source [1] and is well studied for structure, function and biogenesis of the peroxisomes (for reviews see [2–4]). In *P. angusta*, the methanol metabolism requires peroxisomal enzymes such as methanol oxidase or dihydroxyacetone synthase that can constitute up to 60–80% of the total protein mass of the cell. The use of promoter elements, controlling the high level of expression of these genes, has led to important commercial applications and methylotrophic yeasts have thus become prime factories for production of recombinant proteins (for reviews see [5,6]).

P. angusta is homothallic and ascospores can be easily isolated. Development of transformation procedures with autonomously replicating or integrative plasmids has facilitated genetic studies [7]. Prior to this project, about 100 gene sequences of *P. angusta* were available in public databases. They mostly correspond to genes involved in nitrogen or methanol metabolism or in the biogenesis and degradation of peroxisomes.

The present study describes the results of the analysis of 5082 random sequenced tags (RSTs) of the *P. angusta* type strain, revealing a total of about 2500 novel genes.

2. Materials and methods

2.1. Library construction

A total of 120 μg of DNA, extracted from the *P. angusta* strain CBS 4732 according to [8], was nebulized according to [9]. The 2–5 kb long fragments were gel-purified and end-filled using T4 DNA polymerase (Gibco BRL) and Klenow (Pharmacia). A total of 6 μg of final DNA was obtained after phenol-chloroform extraction. The pBAM3 vector (derivative from pBluescript KS) was digested with *Sma*I, dephosphorylated with Calf Intestinal Phosphatase (CIP, N.E. Biolabs) following the recommendations of the manufacturer and gel-purified. Linearized vector (100 ng) and *P. angusta* DNA fragments (100 ng) were ligated overnight at 16°C (with 7.2 U of a 6 U/μl solution of T4 DNA ligase N.E. Biolabs) and phenol-chloroform extracted.

Bacteria *E. coli* DH10B cells (ElectroMAX™ DH10B™, Gibco BRL) electroporated with 1/10 of the ligation mix following the recommendations of the manufacturer and plated on LB medium containing ampicillin (100 μg/ml), X-Gal (40 μg/ml) and IPTG (40 μg/ml) gave approximately 25 000 primary clones with inserts (white colonies).

2.2. Quality control and storage of the DNA library

Bacterial plasmids from 96 white colonies randomly picked were extracted by alkaline lysis minipreparation and digested with *Pvu*II to verify the presence and measure the size of the insert. No empty vector was found and the average size of the inserts was 3.8 kb with a standard deviation of 0.8 kb. Thereafter, 3478 white bacterial clones were randomly picked and grown overnight at 37°C under agitation in 96 well microtiter plates (two wells were left empty in each microplate for contamination control) containing 200 μl of SOC with ampicillin. Microtiter plates were then triplicated by transferring 50 μl aliquots of each bacterial culture to 50 μl of glycerol 80%. The plates were stored at −80°C until sequencing.

2.3. Characteristics of the sequences and calculation of the genome size

A total of 5082 sequences were produced by GENOSCOPE [10], corresponding to a total amount of sequences of ca. 4.9 Mb (4.7 Mb correspond to nuclear DNA, with an average GC content of 47.6%). Strain CBS 4732 of *P. angusta* is likely to be diploid (Serge Casaregola, personal communication). Based on the contig assembly theory [11], the genome size was estimated in this work to be ca. 9 Mb (calculation was performed with a set of RSTs taking into account only one sequence per clone), half of which were sequenced in the present project.

3. Results and discussion

3.1. Nuclear ribosomal DNA

A total of 103 *P. angusta* RSTs were assembled into a single

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

0014-5793/00/$20.00
PII: S0014-5793(00)02284-5

contig whose consensus sequence is highly similar to the repeating unit of *Saccharomyces cerevisiae* rDNA. The rDNA repeats of *S. cerevisiae* are 9.1 kb in length and organized in a single cluster on chromosome 12 [12,13]. We found that the *P. angusta* repeating unit has a size of 8.1 kb, with an identical organization to the *S. cerevisiae* one. The sequences of *P. angusta* encoding 18S, 5.8S, 25S and 5S rRNAs are respectively 94.6, 96, 90 and 97% identical to those of *S. cerevisiae* and were used to construct the phylogenetic tree [14]. The internal transcribed spacers have substantially diverged between *P. angusta* and *S. cerevisiae*. Yet, a sequence of 20 nucleotides around the transcription initiation site for the 35S rRNA and two short segments of 20 and 15 nucleotides located, respectively, 281 and 230 bases downstream of this site, are identical in the two species.

Given the number of RSTs sequences identified to encode rRNA and the estimated size of *P. angusta* genome, the total number of rDNA repeats is only 25, which is very much less than what is observed in *S. cerevisiae* [15] or *Kluyveromyces thermotolerans* [16] but is similar to values obtained for the other species of this project.

3.2. Transposable elements

Retrotransposons are ubiquitous elements among eukaryotic genomes and play an important role in shaping genome evolution. In the *S. cerevisiae* genome, 331 LTR (long terminal repeat) elements have been described [17], of which 51 are part of full-length retrotransposable elements, belonging to five distinct retrotransposon families, designated *Ty1* through *Ty5* [18].

In *P. angusta*, retrotransposons were searched for by comparing sequence translation products with the *Ty* proteins of *S. cerevisiae*. Three contigs, unequal in size and numbers of RSTs, show significant similarities with *Ty* elements of *S. cerevisiae*. The larger one contains 33 RSTs that overlap a complete retrotransposon sequence of 4.2 kb flanked by two identical LTRs of 320 nucleotides each, completely different from those of *S. cerevisiae*. The presence of a unique ORF and the absence of frameshift in this large contig suggest that the majority of retrotransposons in *P. angusta* are more closely related to *Ty5* of *S. cerevisiae*, with the reverse transcriptase and RNaseH domains more conserved than the integrase domain. In addition, a few solo-*Ty5* LTRs were identified.

Given the estimated genome size and the number of identified retrotransposons, *P. angusta* strain CBS 4732 is estimated to contain a total of 15–20 copies of transposable elements of the *Ty5* family. This figure, however, may be an under-estimate because some tags may contain part of retrotransposon ORFs not identified.

Interestingly, another repeated sequence of about 250 nucleotides was identified in 18 independent *P. angusta* RSTs and shows no significant similarity to the *S. cerevisiae* chromosomes or any other DNA sequences in public databases. It is also preferentially associated (on the same insert or the same RST) with retrotransposons or LTRs previously identified above but is not immediately adjacent to a recognizable retrotransposon ORF. In addition, one extremity of this repeated sequence shows 95% of identity over 37 bases with the beginning of the *Ty5* LTRs identified in this study. We propose that this sequence belongs to another class of LTRs and may be associated with transposable elements not identified by comparison with *S. cerevisiae* or may represent the remnant of ancient retrotransposon excisions. RST BB0AA014A08TP1 presents an interesting example of probable integration of one of this putative LTR into a *Ty5* LTR.

3.3. Mitochondrial DNA

P. angusta RSTs were also compared to the complete mitochondrial DNA sequence from *S. cerevisiae* (see [19]), then translated and compared to all mitochondrial translation products of five Ascomycetes species: *S. cerevisiae*, *Allomyces macrogynus*, *Pichia canadensis*, *Podospora anserina* and *Schizosaccharomyces pombe* (accession numbers AJ011856, U41288, D31785, X55026 and X54421). These sets of proteins include standard mitochondrial gene products, group I, group II intron translation products and genes encoding the subunits of the NADH dehydrogenase complex which are present in the mitochondrial DNA of a wide range of fungal species including *A. macrogynus*, *P. canadensis*, *P. anserina* or *P. angusta* but not in the *S. cerevisiae* or the *S. pombe* mitochondrial genomes.

The mitochondrial genome of *P. angusta* was not completely sequenced but a total of 126 RSTs, distributed in 11 contigs plus two singletons, were found to be similar to the mitochondrial genes coding for cytochrome oxidase (subunits 1, 2 and 3), apocytochrome *b*, NADH dehydrogenase (subunits 1, 4, 5 and 6) and ATPase (subunits 8 and 9). The sequences encoding cytochrome oxidase and apocytochrome *b* subunits are associated with intronic ORFs from group I introns. No RST showed significant similarity with *S. cerevisiae* mitochondrial tRNA or rDNA genes.

3.4. Plasmids and autonomously replicating sequences

No natural plasmid is known to exist in *P. angusta*. All RSTs were, nevertheless, compared to the complete sequence of the *S. cerevisiae* 2-micron plasmid but no similarity was found. In addition, no RST was similar to the autonomously replicating sequence HARS36 of the *P. angusta* DL-1 strain [20].

3.5. Identification of protein-coding genes with orthologs in *S. cerevisiae*

The majority of the genes in our sequenced inserts were identified by comparison to the *S. cerevisiae* genome (see [19]): 3175 RSTs (62.4% of total) contain at least one gene or part of a gene having significant similarity with 2502 distinct *S. cerevisiae* ORF products (40.8% of total). The average percent of amino acid identity in *blastx* alignments between the two species is 48.3% [19] which is consistent with other values encountered for the other yeast species of this project considering their position on the phylogenetic tree [14].

In total, we identified a minimum of 2320 and a maximum of 2699 genes of *P. angusta*, the uncertainty about gene number being due to the existence of several RSTs matching non-overlapping parts of the same gene (see [19]). If we assume that the *P. angusta* genome is about 9 Mb in size and that the gene density is similar to that of *S. cerevisiae*, more than 50% of the total *P. angusta* genes were identified. This result fully justifies the choice of our strategy for the rapid identification of a large number of genes in a yeast genome. An exhaustive list of the *S. cerevisiae* genes having homologs in *P. angusta* is available at http://cbi.genopole-bordeaux.fr/genolevures.

Table 1
List of homologs of the *P. angusta* RSTs that do not have a homolog in the *S. cerevisiae* genome

Organism	AC	Functional comments
Bacteria		
Bacillus stearothermophilus	Q45515	D-hydantoinase, dihydropyrimidinase HYDA
Bacillus subtilis	CAB15988	YxeK; similar to monooxygenase
Rhodococcus sp. *IGTS8*	P54995	dibenzothiophene desulfurization enzyme A SOXA
Rhodococcus erythropolis	O05691†	non-heme haloperoxidase THCF
Escherichia coli	AAC75307†	putative racemase b2247
	AAA69296	sensor protein BarA
	AAC75226	yeiN; unknown function
Helicobacter pylori	AAD07376†	hypothetical protein HP0310
Methylophilus methylotrophus	Q50228	formamidase FMDA
Mycobacterium tuberculosis	CAB02180	Rv1399c; probable lipase lipH
	CAB02385	Rv0773c; putative γ-glutamyl transpeptidase ggtA
Pseudomonas aeruginosa	P51691†	arylsulfatase ARS
Pseudomonas sp. *NS671*	Q01262	hydantoin utilization protein A HYUA
	Q01264	hydantoin utilization protein C HYUC
Synechocystis sp.	BAA17668	hypothetical protein, sll1773
Archaea		
Archeoglobus fulgidus	AAB89829	putative membrane protein AF1420
Ascomycetes		
P. angusta	Q00925	peroxisomal matrix protein PER1 precursor (peroxin-8)
	P78723	peroxisomal membrane protein PER10 (peroxin-14)
	AAD52811	PEX1 (peroxin-1)
	P04841	methanol oxidase MOX, AOX.
	CAA92206	nitrite reductase
Pichia pastoris	Q01964	peroxisomal protein PER6
	Q92448	6-phosphofructokinase
Pichia stipitis	P50167	D-arabinitol 2-dehydrogenase ARDH
Candida albicans	O74248	putative polyamine transporter GPT1
	P33181	probable sucrose utilization protein SUC1
	P53705	integrin α chain-like protein INT1
	P87218[2]	sorbitol utilization protein SOU2
Candida boidinii	P14293	peroxisomal membrane protein B PMP20
	P21245	peroxisomal membrane protein PMP47A
Candida maltosa	Q00673	probable NADH-ubiquinone oxidoreductase (CI-31KD)
Kluyveromyces lactis	P07921	lactose permease LACP
Kluyveromyces marxianus	P07337[5–6]	β-glucosidase precursor BGLS
	P30887	acid phosphatase precursor PHO2
S. pombe	AAA02871	cell division control protein 18 CDC18
	AAA35330	double-strand-break repair protein RAD21
	AAB05993	UDP-Glc:glycoprotein glucosyltransferase GPT1
	AL109738	hypothetical protein CAB52163; SPAC8F11.02C
	BAA77269	ALP4
	CAA22658	putative carboxylesterase-lipase family member
	CAA92304	hypothetical protein; SPAC11D3.03c
	CAB11645	hypothetical protein; SPAC19A8.09
	CAB11668	putative vacuolar protein; SPAC23H4.14
	CAB39802	conserved hypothetical protein; SPBC1778.07
	CAB40174	putative D-amino acid oxidase; SPCC1450.07c
	O13770	hypothetical protein; SPAC17A5.08
	O13783	hypothetical protein; SPAC17G6.05C
	O13790	putative cell division control protein; SPAC17G6.12
	O13924	hypothetical protein; SPAC23C4.03
	O42653	hypothetical protein; SPAC10F6.14C
	O42798	ribonuclease H1 RNH1
	O42887	arginase family protein; SPAC8E4.03
	O42947	hypothetical protein; SPBC16H5.12C
	O43000	hypothetical protein; SPBC2G2.01C
	O43005	β-adaptin; SPBC2G2.06C
	O43029[1–2]	fructosyl amine; SPBC354.15
	O59674	putative mitochondrial carrier; SPBC29A3.11C
	O59715	hypothetical protein; SPBC3B8.07C
	O59832	putative dipeptidase; SPCC965.12
	O60064	putative mannose-1-phosphate guanyl transferase; SPBC13G1.02
	O60075	hypothetical protein; SPCC1494.01
	O74334	hypothetical protein; SPBC1685.14C
	O74395	major facilitator superfamily protein; SPBC4F6.09
	O74397	putative asparagine synthase; SPBC4F6.11C
	O74746	putative DNA J domain containing protein
	O74916[2]	putative acetylornithine deacetylase; SPCC757.05C
	O74923	putative membrane transport protein; SPCC757.13
	O74925	putative vacuolar membrane protein; SPCC790.02

Table 1 (*continued*)

Organism	AC	Functional comments
	O74979	hypothetical protein; SPCC1827.07C
	O94269	possible ubiquitin carboxyl-terminal hydrolase; SPBP8B7.21
	O94389	hypothetical protein
	O94401	putative lectin precursor; SPCC126.08C
	O94431	putative class V pyridoxal phosphate dependent aminotransferase
	O94606	hypothetical protein; SPCC622.19
	P08965	SPME12
	P40902	sexual differentiation process protein ISP7
	P78771	SPP78771
	P78795	probable eukaryotic translation initiation; SPTIF35
	P78894	SPP78894
	P87122	hypothetical protein; SPAC3A12.06C
	P87216	VIP1 protein
	Q09709	hypothetical protein; SPAC18B11.02C
	Q09731	hypothetical Trp–Asp repeats containing protein; SPAC31A2.14
	Q09875	hypothetical protein; SPAC12G12.12
	Q09929	hypothetical protein; SPAC21E11.07
	Q10088[2]	putative agmatinase precursor; SPAC11D3.09
	Q10199	hypothetical protein; SPBC11C11.02
	Q10301[1–2]	hypothetical protein; SPAC22H10.08
	Q10341	hypothetical protein; SPAC19G10.13
	Q12381	SPPRP1 or ZER1+ pre-mRNA splicing factor
	Q92341[2–3]	hypothetical protein; SPAC1F8.03C
Aspergillus niger	P41751	aldehyde dehydrogenase DHAL
	Q12556	copper amine oxidase 1 AMO1
	Q92406	NADH-ubiquinone oxidoreductase (CI-51kD)
	P24918	NADH-ubiquinone oxidoreductase (CI-78kD)
	Q12644	NADH-ubiquinone oxidoreductase (CI-23kD)
Neurospora crassa	P19968	NADH-ubiquinone oxidoreductase (CI-21kD)
Emericella nidulans	P48777	purine permease UAPC
	Q07307[5–6]	uric acid-xanthine permease UAPA
Nectria haematococca	P24552	D-amino acid oxidase OXDA
	P52958	cutinase transcription factor 1 α CT1A
Trigonopsis variabilis	Q99042[2–3]	D-amino acid oxidase OXDA
Other eukaryotes		
Arabidopsis thaliana	P46313‡	omega-6 fatty acid desaturase FD6E
Leishmania major	P22045	probable reductase P100
Caenorhabditis elegans	T21562	hypothetical protein F30A10.5
	T21688	hypothetical protein F33A8.1
	T22008	hypothetical protein F39H2.2
	T22460	hypothetical protein F49E2.1
	T32916	hypothetical protein K02F2.3
	T24290	hypothetical protein T01D3.5
	AAB52502	hypothetical protein T27A10.3
	T26280	hypothetical protein W08D2.4
Mus musculus	Q62203	spliceosome-associated protein 62 SP62
Rattus norvegicus	Q63707‡	dihydroorotate dehydrogenase precursor PYRD
	Q64536	[pyruvate dehydrogenase(lipoamide)] kinase PDK2
Homo sapiens	O75380‡	NADH-ubiquinone oxidoreductase (CI-13kD A)
	P17405	sphingomyelin phosphodiesterase precursor ASM
	P34059	*N*-acetylgalactosamine-6-sulfatase precursor GA6S
	Q04609	prostate-specific membrane antigen PSM
	Q16718	NADH-ubiquinone oxidoreductase (CI-13kD B)
	Q16739‡	ceramide glucosyltransferase CEGT

AC: accession number. Superscript (column 2) indicates the number of distinct genes in *P. angusta* matching the same entry (P87218, P07337, O43029, O74916, Q10088, Q10301, Q92341, Q07307 and Q99042). In all other cases, only one homolog is present in *P. angusta*. Note that some *P. angusta* genes share very significant alignments with bacterial or eukaryotic ORFs only and their best homologs in the different species are respectively annotated with † and ‡.

3.6. Identification of additional protein-coding genes without orthologs in S. cerevisiae

In order to identify possible additional genes of *P. angusta* not found in *S. cerevisiae*, comparisons with SwissProt, *S. pombe* and 23 fully sequenced proteomes from the three kingdoms were performed as described in [19]. All the RSTs were also compared with the 79 *P. angusta* non-redundant protein sequences present in GenBank on December 29th, 1999. Table 1 shows results of this analysis.

A minimum of 134 and a maximum of 140 novel genes of *P. angusta* have significant counterparts in other Ascomycetes in higher eukaryotes or in bacteria. Not unexpectedly, most *P. angusta* gene products identified in this analysis are homologous to proteins from Ascomycetes. A total of 16 *P. angusta* gene products, however, are more similar to bacterial or archaeal proteins than to eukaryotic proteins. Moreover, four of them are homologous to bacterial proteins only, respectively to *R. erythropolis* non-heme haloperoxidase (57%

Table 2
Characteristics of *P. angusta* spliceosomal introns identified by comparison to *S. cerevisiae* ones

P. angusta RST	3′ Exon 1	5′ Intron	S1	Branch point	S2	3′ Intron	S	*S. cerevisiae* homologs (intron size)
BB0AA006E10T1	TGG	GTA**A**GT	212	TACTAAC	3	CAG	222	*YBL099w* (none)
BB0AA012H05DP1	TGT	GTA**A**GT	35	TACTAAC	nd	nd	nd	*YMR004* (none)
BB0AA023H04TP1	TTG	GTA**A**GT	nd	nd	nd	TAG	41	*YJL166w* (none)
BB0AA002A05D1†	AAG	GTA**A**GT	122	TACTAAC	8	TAG	137	*YMR116c* (273)
BB0AA002C10D1	ATG	GTA**A**GT	41	**C**ACTAAC	8	CAG	56	*YDR447c** (314), *YML024w** (398)
BB0AA002E10D1†	CTC	GTA**G**GT	169	TACTAAC	9	TAG	185	*YNL301c** (432), *YOL120c** (447)
BB0AA002F02T1†	ACA	GTA**A**GT	nd	nd	nd	CAG	37	*YNL246w* (95)
BB0AA002G03D1	AAT	GTA**A**GT	nd	nd	nd	CAG	36	*YGL087c* (85)
BB0AA002H06T1	*AAG*	*GTATGT*	~278	TACTAAC	8	TWG	~293	*YDR381w* (766)
BB0AA004F01T1	CTC	GTA**A**GT	153	T-CTAAC	7	TAG	166	*YJL136c** (460), *YKR057w** (322)
BB0AA005E07D1	TTG	GTATGT	210	T**G**CTAAC	8	TAG	225	*YDR471w** (384), *YHR010w** (561)
BB0AA005G04D1	AAG	GTA**G**GT	30	TACTAAC	7	AAG	44	*YBR181c** (352), *YPL090c** (394)
BB0AA006F12T1	AGG	GTA**G**GT	nd	nd	nd	TAG	53	*YDL125c* (111)
BB0AA018D09DP1†	GAG	GTA**G**GT	73	TACTAAC	6	TAG	86	*YBL026w* (128)
BB0AA008E01T1†	TCT	GTA**A**GT	20	TA**G**CTAAC	6	CAG	34	*YPR028w* (133)
BB0AA009E10T1	nd	nd	nd	**A**ACTAAC	6	TAG	nd	YNR053c (531)
BB0AA010D06DP1†	AGA	GTA**A**GT	nd	nd	nd	CAG	35	*YDR092w* (268)
BB0AA010F11TP1	ACG	GTA**A**GT	27	TA**T**CTAAC	9	TAG	45	*YML085c* (116), *YML124c* (298)
BB0AA013C02DP1	AAG	GTA**C**GT	26	TACT**GAC**AAC	7	AAG	43	*YGL137w* (200)
BB0AA013G05TP1†	ATG	GTA**A**GT	20	T**G**CTAAC	4	TAG	31	*YLR078c* (89)
BB0AA015B10DP1	GAA	GTATGT	23	TACTAAC	8	CAG	38	*YDR500** (389), *YLR185w** (359)
BB0AA019G12DP1†	AAG	GTA**A**GT	52	**C**ACTAAC	7	CAG	66	*YBR191w** (388), *YPL079w** (421)
XBB0AA002E12T1†	ATG	GTA**A**GT	23	**A**ACTAAC	7	CAG	37	*YER074w** (466), *YIL069c** (409)
BB0AA026F04DP1	*CAG*	GTA**A**GT	~30	T**G**CTAAC	7	TAG	~44	*YDL083c** (432), *YMR143w** (544)
XBB0AA002E03D1	ATG	GTA**A**GT	173	TACTAAC	8	TAG	188	*YLR367w** (483)
BB0AA026C04TP1	nd	nd	nd	TACTAAC	5	TAG	nd	*YGL076c** (468), *YPL198w** (409): first intron
BB0AA026C04TP1	GCC	GTATGT	138	TACTAAC	8	TAG	153	*YGL076c** (459), *YPL198w** (407): second intron
BB0AA018F03DP1†	*ACC*	*GTA**A**GT*	nd	nd	nd	*AAG*	~45	*YDR424c* (80): first intron
BB0AA018F03DP1	*ATG*	*GTA**A**GT*	~19	T-CTAAC	7	CAG	~32	*YDR424c* (96): second intron

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: *P. angusta* intron length. All sizes are in nucleotides. nd: not determined. †Indicates that the *P. angusta* intron was found on different RSTs assembled into the same contig, only one RST being reported. Bold types identify positions not in agreement with the GTATGT and TACTAAC consensus of *S. cerevisiae*. Italics indicate an ambiguity in the consensus position. *Identifies ribosomal protein encoding genes. The last four lines report two *P. angusta* genes that contain two introns, as in their *S. cerevisiae* counterparts.

identity), *P. aeruginosa* arylsulfatase (60% identity), and *E. coli* or *H. pylori* genes of unknown function (respectively 37 and 62% of identity).

3.7. Spliceosomal introns

The prediction of spliceosomal introns from our sequences is difficult because most genes were only partially sequenced. To circumvent this difficulty, we decided to only examine the presence of introns among the 89 genes of *P. angusta* whose *S. cerevisiae* homolog contains spliceosomal introns [21]. In such cases, 5′ and 3′ splice sites and branchpoint consensus sequences were searched in the *P. angusta* gene at positions equivalent to the *S. cerevisiae* orthologs. In 29 cases, this position was not available in our sequences and for 12 other genes, we are uncertain about the possible presence of an intron, because of low quality of alignments with the *S. cerevisiae* ortholog. From all other cases, we identified introns in 24 *P. angusta* genes (see Table 2), with splice sites and branchpoint sequences mostly consistent with those described in *S. cerevisiae* [21]. Moreover, during manual annotation of the alignments, we identified three genes (also represented in Table 2) of *P. angusta* that contain an intron whereas their counterparts in *S. cerevisiae* do not. In addition, the RSTs BB0AA026C04TP1 and BB0AA018F03DP1 contain genes similar to *YGL076c*, *YPL198w* and *YDR424c*, respectively, which are among the rare genes of *S. cerevisiae* containing

two introns. The homologous genes in *P. angusta* also contain two introns at equivalent positions.

Interestingly, the introns of *P. angusta* are significantly shorter (three times on average) than their *S. cerevisiae* counterparts. Short intron size is also apparent by measuring the distance between the branchpoint and the 3′ splice site which is as short as seven nucleotides on average in *P. angusta*, compared to 40 for *S. cerevisiae* [22].

Finally, among the 89 *P. angusta* genes examined, 24 genes do not possess an intron at the position as defined in their respective *S. cerevisiae* orthologs.

3.8. tRNA genes

P. angusta tRNA genes were searched for by comparison with the 52 families of *S. cerevisiae* tRNA genes [23]. A total of 28 complete or partial sequences of putative tRNA genes were identified. Assuming that our genomic library represents about 0.5 genome equivalents, the total number of tRNA genes identified is low when compared to the 274 tRNAs in *S. cerevisiae* but equivalent values were obtained for the majority of the 12 other yeasts inspected in this project.

In all but two cases, the anticodon is conserved between the *P. angusta* tRNA gene and its homolog in *S. cerevisiae*. Considering that the genetic code is universal in the two species [19], the two changes preserve the charged amino acid. Among the 28 recognized tRNA genes, only one contains an intron which has diverged in size and sequence from that of the *S. cerevisiae* homolog. On the contrary, four genes do not possess an intron when their counterparts in *S. cerevisiae* do. Finally, in *P. angusta*, two tRNA genes with anticodons AAC and CTC are tandemly duplicated (respectively in RSTs BB0AA001D07T1 and BB0AA008B01D1), interspersed by a short region of roughly 75 nucleotides, a situation not observed in *S. cerevisiae*.

3.9. Gene family expansion

Estimating gene redundancy is difficult in the absence of the complete sequence of the *P. angusta* genome and even more so because the genes that we identified were not fully sequenced. A comprehensive study of the evolution of structural gene families between *S. cerevisiae* and the 13 hemiascomycetous yeasts of the project is discussed in [24]. We want to point out here five cases of gene families clearly larger in *P. angusta* than in *S. cerevisiae*.

As a first example, the ORFs *YJR152w* (*DAL5*, allantoate transporter) and *YLR004c* belong to a three-member family (see [23] for a definition of families). These two genes are homologous to at least nine different genes of *P. angusta* (with alignments showing 28–41% of amino acid identity) and additional genes similar to these allantoate transporter genes of *S. cerevisiae* probably exist in the rest of the genome. Percentages of identity are not sufficient to permit the conclusion that these nine novel genes could have the same function as their orthologs. Nevertheless they may encode other putative transporters for different substrates, absent from the *S. cerevisiae* proteome. The third ORF of the allantoate transporter family, *YLL055w*, has no similarity with any *P. angusta* RST.

Other examples illustrating the expansion of gene families in *P. angusta* in comparison to *S. cerevisiae* are the genes encoding the mannosyltransferases *YBR015c* (*MNN2*) and *YJL186w* (*MNN5*), a family of two in *S. cerevisiae* with seven different homologs in *P. angusta*; or *YNL111c* (*CYB5*, cytochrome *b5*), *YNL202w* (*SPS19*, peroxisomal 2,4-dienoyl-CoA reductase, sporulation specific protein) and *YKL215c* (similar to hyuA and hyuB of *P. aeruginosa*), singletons in *S. cerevisiae* but with four or five homologs in *P. angusta*.

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 Moléculaire des Levures, especially Agnès Thierry for help in the construction of libraries, Fredj Tekaia and Alexis Harington for comments on the manuscript. B.D. is a member of Institut Universitaire de France.

References

[1] Kurtzman, C.P. (1998) in: The yeasts. A taxonomic study (P., K.C. and W., F.J., Eds.), pp. 273–352, Elsevier, Amsterdam.
[2] van der Klei, I.J. and Veenhuis, M. (1997) Trends Microbiol. 5, 502–509.
[3] Sulter, G.J., Harder, W. and Veenhuis, M. (1993) FEMS Microbiol. Rev. 11, 285–296.
[4] Veenhuis, M., van der Klei, I.J., Titorenko, V. and Harder, W. (1992) FEMS Microbiol. Lett. 79, 393–403.
[5] Faber, K.N., Harder, W., Ab, G. and Veenhuis, M. (1995) Yeast 11, 1331–1344.
[6] Hollenberg, C.P. and Gellissen, G. (1997) Curr. Opin. Biotechnol. 8, 554–560.
[7] Gellissen, G. and Hollenberg, C.P. (1997) Gene 190, 87–97.
[8] Johnston, J.R. (1988) in: Yeast, a practical approach (I., C. and H., D.J., Eds.), pp. 107–123, IRL, Oxford.
[9] Ozier-Kalogeropoulos, O., Malpertuy, A., Boyer, J., Tekaia, F. and Dujon, B. (1998) Nucleic Acids Res. 26, 5511–5524.
[10] Artiguenave, F. et al. (2000) FEBS Lett. 487, 13–16 (this issue).
[11] Lander, E.S. and Waterman, M.S. (1988) Genomics 2, 231–239.
[12] Johnston, M. et al. (1997) Nature 387, 87–90.
[13] Goffeau, A. et al. (1996) Science 274, 563–567.
[14] Souciet, J.-L. et al. (2000) FEBS Lett. 487, 3–12 (this issue).
[15] Rustchenko, E.P. and Sherman, F. (1994) Yeast 10, 1157–1171.
[16] Malpertuy, A., Llorente, B., Blandin, G., Artiguenave, F., Wincker, P. and Dujon, B. (2000) FEBS Lett. 487, 61–65 (this issue).
[17] Hani, J. and Feldmann, H. (1998) Nucleic Acids Res 26, 689–696.
[18] Goodwin, T.J. and Poulter, R.T. (2000) Genome Res. 10, 174–191.
[19] Tekaia, F. et al. (2000) FEBS Lett. 487, 17–30 (this issue).
[20] Sohn, J.H., Choi, E.S., Kang, H.A., Rhee, J.S. and Rhee, S.K. (1999) J. Bacteriol. 181, 1005–1013.
[21] Lopez, P.J. and Seraphin, B. (1999) RNA 5, 1135–1137.
[22] Spingola, M., Grate, L., Haussler, D. and Ares Jr., M. (1999) RNA 5, 221–234.
[23] Blandin, G., Tekaia, F., Durrens, P., Aigle, M., Bolotin-Fukuhara, M. et al. (2000) FEBS Lett. 487, 31–36 (this issue).
[24] Llorente, B., Durrens, P., Tekaia, F., Aigle, M., Artiguenave, F. et al. (2000) FEBS Lett. 487, 71–75 (this issue).