Genomic Exploration of the Hemiascomycetous Yeasts: 16. *Candida tropicalis*

Gaëlle Blandin[a], Odile Ozier-Kalogeropoulos[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 27 November 2000

Edited by Horst Feldmann

Abstract The genome of the diploid hemiascomycetous yeast *Candida tropicalis*, an opportunistic human pathogen and an important organism for industrial applications, was explored by the analysis of 2541 Random Sequenced Tags (RSTs) covering about 20% of its genome. Comparison of these sequences with *Saccharomyces cerevisiae* and other species permitted the identification and the analysis of a total of more than 1000 novel genetic elements of *C. tropicalis*. Moreover, the present study confirms that in *C. tropicalis*, the rare CUG codon is read as a serine and not a leucine. The sequences have been deposited at EMBL with the accession numbers AL438875–AL441602.

Key words: Genetic code; Diploid species; Intron; rDNA

1. Introduction

Several species of the fungal group *Candida*, including *Candida tropicalis*, are pathogenic for humans [1]. In recent years, infections by these yeasts have become a significant health problem associated with the spread of AIDS, the increased use of the immunosuppressive therapy and the increase of nosocomial fungal infections [2]. *C. tropicalis* has also important industrial applications. In the chemical industry, its capacity to assimilate *n*-alkanes is used for the production of long-chain dicarboxylic acids and for the preparation of polyamide, polyester and perfume [3]. In the food industry, *C. tropicalis* is used for its capacity to transform xylose into xylitol, a five-carbon sugar alcohol able to replace sucrose [4]. This species has also been extensively studied for the induction of peroxisomal enzymes involved in the utilization of *n*-alkanes [5], and because of the non-universality of its genetic code. As in most species of the genus *Candida*, the universal leucine CUG codon in *C. tropicalis* is read as serine [6,7].

It is well established that *C. tropicalis* is a diploid yeast without sexual reproduction [3,8], but other characteristics of this yeast, such as the number of its chromosomes and its genome size, are not precisely known. Analysis by pulsed-field gel electrophoresis of different strains revealed 10–12 chromosomes and the calculation of its diploid genome size varies between 20 and 31 Mb depending on the authors [8,9].

At the beginning of our study, about 100 sequences of *C. tropicalis* were present in public databases corresponding to 60 different genes. We present here the analysis of 2541 Random Sequenced Tags (RSTs) of the genome of this yeast and report the discovery of more than 1000 novel genes.

2. Materials and methods

The *C. tropicalis* (type strain CBS 94) random genomic library was constructed as described by Blandin et al. [10] and is composed of 1786 different clones. The average size of the inserts, determined on a sample of 96 white bacterial clones randomly picked, is 3.5 kb (standard deviation: 0.9 kb). A total of 2541 sequences were produced by Genoscope [11].

3. Results and discussion

3.1. Characteristics of the sequences

Altogether 2.4 Mb were sequenced. A total of 2.3 Mb correspond to nuclear DNA and the remaining fraction to mitochondrial elements. The guanine-cytosine content of the nuclear sequences is 34.9% which is significantly lower than 44%, the value obtained in a previously study effectuated on a smaller sample of sequences [12].

The sequences were assembled into 309 contigs as described in [13]. The largest contigs contain respectively 214 sequences identified as ribosomal DNA, 37 and 24 sequences corresponding to mitochondrial DNA, and 15 sequences probably bearing Long Terminal Repeat elements (LTRs). The other 305 contigs are constituted of 2–8 RSTs each and 1558 sequences were not included in any contig. The fact that *C. tropicalis* is diploid without sexual reproduction with an interallelic divergence of 7%, as previously determined for two different genes [8], could biased part of the assembly.

The haploid genome size equivalent of the strain CBS 94 was estimated from the contig distribution as described in [14]. Assembly was done using only one sequence per cloned fragment (1359 RSTs) and haploid genome size was found at 11 Mb. This value is close to the calculation of Kamiryo et al. [8] on the strain pK233 and is about 25% less than the average of the values obtained by Doi et al. [9] on three strains (NUM37, NUM267 and NUM400). Based on these data, the 2.3 Mb sequenced in this work represent about 20% of the haploid genome equivalent.

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

PII: S0014-5793(00)02287-0

3.2. Nuclear ribosomal DNA

A total of 214 *C. tropicalis* RSTs were assembled into a single contig whose consensus sequence is highly similar to the *Saccharomyces cerevisiae* rDNA repeat unit. As in *S. cerevisiae*, the *C. tropicalis* genes coding for the 35S ribosomal precursor RNA and the 5S RNAs are in opposite orientation and our data are consistent with a single cluster of rDNA repeat. The repeating unit is longer in *C. tropicalis* than in *S. cerevisiae* (10.7 and 9.1 kb, respectively [15]) and its structure is apparently different because 5S rRNA genes seem tandemly repeated. The 25S, 18S, 5.8S and 5S rRNA genes of *C. tropicalis* are respectively 94.7, 95, 95.6 and 91.4% identical to those of *S. cerevisiae* and were used to construct a phylogenetic tree [16]. Moreover, coding sequences excluded, the other regions of the repeated unit (internal transcribed spacers and transcription initiation site) have diverged. From the number of RSTs obtained, the rDNA region is estimated to contain about 80 repeats.

3.3. tRNA genes

C. tropicalis tRNA genes were searched for by comparison with the 42 families of *S. cerevisiae* tRNA genes [19]. A total of 22 complete or partial sequences of putative tRNA genes were thus identified. Their anticodons are identical to those of the *S. cerevisiae* homologs with one exception where nevertheless the change is synonymous considering the genetic code of *C. tropicalis*.

Three of the 22 identified tRNA genes contain introns whose sizes and sequences have diverged from the introns of their *S. cerevisiae* homologs. In addition, two tandemly duplicated tRNA genes with intron were found homologous to a *S. cerevisiae* tRNA gene having no intron. These two genes are in the same orientation and interspersed by a short region of 18 nucleotides (RST XBD0AA002D11D1). In addition, they are located 14 nucleotides downstream from a LTR element (see below). At the opposite, three *C. tropicalis* tRNA genes without intron were identified by similarity with tRNA genes of *S. cerevisiae* that possess an intron.

3.4. Transposable elements

C. tropicalis transposable elements were first searched for by comparing the translation products of the RSTs with the

Table 1
Analysis of *C. tropicalis* RSTs having no validated homolog in the genome of *S. cerevisiae*: list of homologs present in other organisms

Organism	AC	Functional comments, gene name or ORF number
Bacteria		
Methylophilus methylotrophus	Q50228	formamidase (EC 3.5.1.49), FMDA
Rhizobium sp. (strain NGR234)	P55441	hypothetical monooxygenase, Y4FC
Thermotoga maritima	AAD35670	putative lipopolysaccharide biosynthesis protein BplA, TM0585
Archaea		
Pyrococcus horikoshii	BAA30703	hypothetical ferripyochelin binding protein, PH1591
Eukaryotes		
Ascomycetes		
Candida albicans	P53709	DNA repair protein, RAD14
	P28875	zinc finger protein 1, CZF1
	P46599	serine/threonine protein kinase STE7 homolog (EC 2.7.1.-)
	P78588	probable ferric reductase transmembrane component, FREL
	O74713	high-affinity glucose transporter, HGT1
	P87219	sorbitol utilization protein, SOU1
	P87024	β-glucan synthesis-associated protein, SKN1
	O13433	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase 1 (EC 3.1.4.11), PLC-1
	P46596	opaque-phase-specific protein OP4 precursor, OPS4
	P46589	adherence factor (adhesion and aggregation mediating surface antigen), ADF1
	O13368	agglutinin-like protein ALA1 precursor
	P46590	agglutinin-like protein 1 precursor, ALS1
	O74623	agglutinin-like protein 3 precursor, ALS3
	O74660	agglutinin-like protein 4 precursor, ALS4
Pichia jadinii	P78609	uricase (EC 1.7.3.3)
Pichia stipitis	P22144	D-xylulose reductase (EC 1.1.1.9), XYL2
Yarrowia lipolytica	P30887	acid phosphatase precursor (EC 3.1.3.2), PHO2
Saccharomycopsis fibuligera	P22506	β-glucosidase 1 precursor (EC 3.2.1.21), BGL1
S. pombe	CAA18296	putative fatty acid desaturase, SPBC3B8.07C
	CAB39904	putative transcription factor, SPCC645.08C
	Q10088	putative agmatinase precursor (EC 3.5.3.11), SPAC11D3
	AAD37449	stress-activated MAP kinase interacting protein, SIN1
	CAA16996	putative agmatinase precursor (EC 3.5.3.11), SPBC8E4.03
Haematonectria haematococca	P52958	cutinase transcription factor 1 α, CT1A
Neurospora crassa	Q02854	NADH-ubiquinone oxidoreductase (CI-21 kDa) (EC 1.6.5.3)
Other eukaryotes		
Euglena gracilis	P30397	hypothetical chloroplastic 64.3 kDa protein in RPS3 3′ region, ORF516
Arabidopsis thaliana	P42744	auxin-resistance protein, AXR1
Picea abies	Q08632	short-chain type dehydrogenase/reductase, SDR1
Drosophila melanogaster	Q94535	splicing factor U2AF 38 kDa subunit
Gallus gallus	Q90980	cyclic nucleotide-gated channel, ROD photoreceptor, α subunit CNG3
Rattus norvegicus	Q64536	[pyruvate dehydrogenase(lipoamide)] kinase isozyme 2, mitochondrial precursor (EC 2.7.1.99), PDK2
Homo sapiens	Q15647	thyroid receptor interacting protein 15, TRIP15

AC: accession number. In all cases, only one homolog has been found in *C. tropicalis*, except for the *C. albicans* ALS3 gene (AC: O74623) matching with two distinct genes of *C. tropicalis*.

Table 2
Characteristics of *C. tropicalis* spliceosomal introns identified by comparison to *S. cerevisiae* ones

C. tropicalis RST	3′ Exon 1	5′ Intron	S1	Branch point	S2	3′ Intron	S	*S. cerevisiae* homologs (intron size)
BD0AA001G05TP1	AGA	GTATGT	289	TACTAAC	20	TAG	316	*YHL001w** (398), *YKL006w** (398)
BD0AA002C05DP1	GAA	GTATGT	193	TACTAAC	30	TAG	230	*YLR287ca** (430), *YOR182c** (411)
BD0AA002G12DP1	nd	nd	nd	TACTAAC	28	TAG	>354	*YHR141c** (441), *YNL162w** (482)
BD0AA006B01TP1	nd	GTATGT	223	TACTAAC	14	TAG	244	*YDR450w** (435), *YML026c** (401)
BD0AA016G11TP1	nd	nd	nd	TACTAAC	27	TAG	>79	*YLR448w** (384), *YML073c** (415)
BD0AA006A02TP1†	GAT	GTATGT	362	TACTAAC	14	TAG	383	*YPL143w** (525), *YOR234c** (527)
BD0AA003D01TP1	nd	nd	nd	TACTAAC	14	TAG	>76	*YAL003w* (366)
BD0AA008C11TP1	GCT	GTATGT	39	TACTAAC	13	TAG	59	*YER179w* (92)
BD0AA014A03DP1	nd	nd	nd	TACTAAC	19	TAG	>68	*YNR053c* (531)
XBD0AA002E05D1	nd	nd	nd	TACTAAC	11	TAG	>415	*YBR082c* (95), *YDR059c* (90)
BD0AA010C10DP1	CAG	GTATGT	41	**TCTAAC**	15	AAG	62	*YEL023c* (none)

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 and S2 indicate the length of the interval respectively comprised between the first position of the intron and the beginning of the branch point box and between the end of the branch point box and the last position of the intron. The 3′ splice site is identified by the last three nucleotides of the intron (3′ Intron). S: *C. tropicalis* intron length. All sizes are in nucleotides. nd: not determined. †Indicates that in this case, the *C. tropicalis* intron was found on two RSTs assembled into a contig, and that only one RST is reported. Bold types identify nucleotides of a sequence not in total agreement with the branch point consensus of *S. cerevisiae*. *Indicates ribosomal protein encoding genes.

proteins of *S. cerevisiae* Ty retrotransposons [18]. Only two RSTs, assembled in a 1.4 kb contig, contain an open reading frame showing weak similarity with a *Ty*3 protein of *S. cerevisiae*. Another 2.6 kb contig, containing 15 RSTs that remained unidentified after all comparisons with *S. cerevisiae* elements, has a structure that is consistent with the proposition that it represents a family of repeated elements. These elements could be LTRs because the sequences composing this contig are strictly identical in its first part but then suddenly diverge over the last 800 bp. They appear to be no full-length retrotransposons because the common region of 1.8 kb does not present any obvious open reading frame. Two RSTs of this contig contain, respectively, one tRNA gene and two tandemly repeated tRNA genes. Using this contig for similarity search, we identified five other RSTs bearing putative LTR elements among the whole set of *C. tropicalis* RSTs. Two of them also contain tRNA genes. In *S. cerevisiae*, *Ty3* transposable elements are known to preferentially integrate in a distance-specific manner, 16–19 bp upstream from tRNA genes [19]. If a similar mechanism is maintained in *C. tropicalis*, the LTRs associated with tRNA genes in *C. tropicalis* are probably *Ty3* LTRs.

3.5. Mitochondrial DNA

C. tropicalis RSTs were also compared to the complete mitochondrial DNA sequence of *S. cerevisiae* and to the translation products of four other fungi: *Allomyces macrogynus*, *Pichia canadensis*, *Podospora anserina* and *Schizosaccharomyces pombe* (respective accession numbers: U41288, D31785, X55026 and X54421). A total of 87 RSTs, distributed in eight contigs plus two singletons, were similar to the mitochondrial genes encoding 21S rRNA, 15S rRNA, tRNA-glu, cytochrome oxidase subunits 1, 2 and 3, apocytochrome *b*, NADH dehydrogenase subunits 1, 2 and 4 and ATPase subunits 6, 8 and 9.

3.6. Comparison with the proteome of S. cerevisiae: identification of C. tropicalis protein-coding genes, confirmation of genetic code and determination of codon usage

The first search to identify genes encoding proteins in *C. tropicalis* was performed against the *S. cerevisiae* translation products, as described in [13]. All RSTs were translated with the universal genetic code in spite of the fact that, in *C. tropicalis*, the CUG codon is read as a serine and not a leucine [7]. It would have been more appropriate to use the *C. tropicalis* code, however the relatively low occurrence of the CUG codon (0.3% among the 12 221 serine codons) reduced the impact of this difference on the genes.

A total of 1060 RSTs (45.6% of the total set) contain at least one gene or part of a gene having significant similarity with 1130 distinct *S. cerevisiae* ORF products (18.5% of 6123). The average value of the percentage identity of *blastx* alignments between the two species is 47.05%. In total, a minimum of 928 and a maximum of 1012 genes of *C. tropicalis* were identified by homology with *S. cerevisiae*. The uncertainty about gene number is due to the existence of several RSTs matching non-overlapping parts of the same *S. cerevisiae* gene (see [13] for details). All genes identified in this work are novel, with the exception of 14 of them corresponding to previously characterized genes. An exhaustive list of the entire *S. cerevisiae* genes having homologs in *C. tropicalis* is available at http://cbi.genopole-bordeaux.fr/genolevures.

The nuclear genetic code of *C. tropicalis* was studied using validated *blastx* alignments obtained with *S. cerevisiae* proteins (for details, see [13]). Results confirm that the CUG codon is read as a serine and not a leucine in *C. tropicalis*.

The relative synonymous codon usage (RSCU) of *C. tropicalis* was calculated as described in [13] for each amino acid and is available at the Genoscope site http://www.cns.fr). Consistent with the low GC content of its genome, *C. tropi-*

calis tends to favor AT-rich codons. Among the codons encoding arginine for example, AGA is 10-fold more frequent than CGA.

3.7. Identification of additional C. tropicalis protein-coding genes

In total, 36 additional genes of *C. tropicalis* were identified (Table 1) following the method described in [13]. Among them, the closest homolog most often belongs to an ascomycetous species, specifically to *Candida albicans* whose five genes encoding adherence factor or agglutin-like proteins which are important pathogenic factors [20]. In four cases, *C. tropicalis* gene products share a higher similarity with a bacterial or an archeal protein than with an eukaryotic protein. Two of them have a particularly high amino acid identity with bacterial proteins only (see Table 1 for details), corresponding to a hypothetical monooxygenase of *Rhizobium* sp. (53% identity) and to a formamidase of *M. methylotrophus* (52% identity), respectively.

The *C. tropicalis* RSTs remaining without homolog after these analyses, were compared using the *tblastx* program, to a set of sequences of *C. albicans* consisting of 1631 contigs all longer than 2 kb and corresponding to a total of 14.9 Mb (Assembly 4, http://www-sequence.stanford.edu/group/candida). The results indicate that 243 *C. tropicalis* RSTs present significant homology (P-value $< 10^{-9}$) on fragments longer than 100 amino acids, and that 41 of them share more than 70% of amino acid identity with their *C. albicans* counterpart without known homolog until now.

3.8. Sequences with spliceosomal introns

Prior to this work, three genes of *C. tropicalis* were reported to have introns [21]. In this study, sequences of *C. tropicalis* homologous to *S. cerevisiae* genes having introns were analyzed by searching for the 5′ and 3′ splice sites and the branchpoint consensus sequence at positions similar to their *S. cerevisiae* homologs. In this way, we identified putative introns in 10 *C. tropicalis* genes, most of them encoding ribosomal proteins (Table 2). Interestingly, *C. tropicalis* introns are systematically shorter than those described in *S. cerevisiae*. Moreover, one gene of *C. tropicalis* (RST BD0AA010C10DP1, Table 2) was identified to contain an intron whereas its *S. cerevisiae* counterpart, *YEL023c*, is devoid of an intron.

Nine other *C. tropicalis* genes do not possess any intron at the positions corresponding to their *S. cerevisiae* counterparts and for six other genes, alignments were too uncertain to allow a definitive conclusion on the possible presence of an intron.

3.9. Expansion of gene families

Some sequences revealed the existence of genes belonging to families in *C. tropicalis* when their homologs are unique in *S. cerevisiae* (for details, see [17] and [22]). For all of them, the level of nucleotidic identity is low enough to consider that the *C. tropicalis* genes are not heteroalleles. In five cases, at least two different genes exist in *C. tropicalis* because two different RSTs are similar to the same portion of a *S. cerevisiae* gene: *POX1* (encoding acyl-CoA oxidase), *CAR1* (arginase) *JEN1* (carboxylic acid transporter protein) *SSU1* (sulfite sensitive protein) and *YMR155w* (unknown function). In another case, at least three *C. tropicalis* genes are homologous to a unique *S. cerevisiae* gene (*DUR3* encoding the urea transport protein), two of them being tandemly duplicated. Other cases of tandemly duplicated genes were also found: the genes homologous to *YJL212c* (similar to *S. pombe ISP4* encoding sexual differentiation process protein), to *YGR197c* (*SNG1* encoding nitrosoguanidine resistance), to *YLR284c* (*EHD1* encoding enoyl-CoA hydratase), to *YGR184c* (*UBR1* encoding ubiquitin-protein ligase), and to *YMR058w* (*FET3* encoding cell surface ferroxidase).

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 her help in the library construction, Fredj Tekaia to have performed bioinformatics analysis and Alexis Harington for critical reading of the manuscript. B.D. is a member of Institut Universitaire de France.

References

[1] Ahearn, D.G. (1978) Annu. Rev. Microbiol. 32, 59–68.
[2] Odds, F.C. (1987) Crit. Rev. Microbiol. 15, 1–5.
[3] Picataggio, S., Deanda, K. and Mielenz, J. (1991) Mol. Cell. Biol. 11, 4333–4339.
[4] Oh, D.K. and Kim, S.Y. (1998) Appl. Microbiol. Biotechnol. 50, 419–425.
[5] Kanai, T., Hara, A., Kanayama, N., Ueda, M. and Tanaka, A. (2000) J. Bacteriol. 182, 2492–2497.
[6] Kawaguchi, Y., Honda, H., Taniguchi-Morimura, J. and Iwasaki, S. (1989) Nature 341, 164–166.
[7] Sugita, T. and Nakase, T. (1999) Syst. Appl. Microbiol. 22, 79–86.
[8] Kamiryo, T., Mito, N., Niki, T. and Suzuki, T. (1991) Yeast 7, 503–511.
[9] Doi, M., Homma, M., Chindamporn, A. and Tanaka, K. (1992) J. Gen. Microbiol. 138, 2243–2251.
[10] Blandin, G., Llorente, B., Malpertuy, A., Wincker, P., Arctiguenave, F. and Dujon, B. (2000) FEBS Lett. 487, 31–36 (this issue).
[11] Artiguenave, F. et al. (2000) FEBS Lett. 487, 13–16 (this issue).
[12] Osumi, M. and Kazama, H. (1978) FEBS Lett. 90, 309–312.
[13] Tekaia, F. et al. (2000) FEBS Lett. 487, 17–30 (this issue).
[14] Lander, E.S. and Waterman, M.S. (1988) Genomics 2, 231–239.
[15] Johnston, M. et al. (1997) Nature 387 (Suppl.), 87–90.
[16] Souciet, J.-L. et al. (2000) FEBS Lett. 487, 3–12 (this issue).
[17] Blandin, G., Tekaia, F., Durrens, P., Aigle, M., Bolotin-Fukuhara, M. et al. (2000) FEBS Lett. 487, 76–81 (this issue).
[18] Brodeur, G.M., Sandmeyer, S.B. and Olson, M.V. (1983) Proc. Natl. Acad. Sci. USA 80, 3292–3296.
[19] Hani, J. and Feldmann, H. (1998) Nucleic Acids Res. 26, 689–696.
[20] Calderone, R.A. and Braun, P.C. (1991) Microbiol. Rev. 55, 1–20.
[21] Ueda, M., Sanuki, S., Kawachi, H., Shimizu, K., Atomi, H. and Tanaka, A. (1997) Arch. Microbiol. 168, 8–15.
[22] Llorente, B., Durrens, P., Tekaia, F., Aigle, M, Artiguenave, F. et al. (2000) FEBS Lett. 487, 101–112 (this issue).