The complete mitochondrial genome sequence of the pathogenic yeast *Candida* (*Torulopsis*) *glabrata*

Romain Koszul[a,1,*], Alain Malpertuy[a,1], Lionel Frangeul[b], Christiane Bouchier[c], Patrick Wincker[d], Agnès Thierry[a], Stéphanie Duthoy[c], Stéphane Ferris[c], Christophe Hennequin[a,e], Bernard Dujon[a]

[a] *Unité de Génétique Moléculaire des Levures (URA 2171 du CNRS, UFR 927 Université Pierre et Marie Curie), Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France*
[b] *Annotation Platform, Génopole, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France*
[c] *Genomics Platform, Génopole, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France*
[d] *Génoscope – Centre National de Séquençage, 2 rue Gaston Crémieux, B.P. 191, 91006 Evry Cedex, France*
[e] *Laboratoire de Parasitologie–Mycologie, CHU Amiens, 80054 Amiens Cedex 1, France*

Received 6 November 2002; accepted 21 November 2002

First published online 10 December 2002

Edited by Horst Feldmann

Abstract We report here the complete sequence of the mitochondrial (mt) genome of the pathogenic yeast *Candida glabrata*. This 20 kb mt genome is the smallest among sequenced hemiascomycetous yeasts. Despite its compaction, the mt genome contains the genes encoding the apocytochrome b (*COB*), three subunits of ATP synthetase (*ATP6*, *8* and *9*), three subunits of cytochrome oxidase (*COX1*, *2* and *3*), the ribosomal protein *VAR1*, 23 tRNAs, small and large ribosomal RNAs and the RNA subunit of RNase P. Three group I introns each with an intronic open reading frame are present in the *COX1* gene. This sequence is available under accession number AJ511533.

Key words: DNA sequence; Mitochondrial DNA; *Candida glabrata*

1. Introduction

Mitochondrial (mt) genomes of yeasts show a high degree of diversity in size, gene content and organization. The well-studied 80 kb mt genome of *Saccharomyces cerevisiae* is the largest one entirely sequenced [1] compared to the ones of *Candida albicans* (40 kb), *Yarrowia lipolytica* (48 kb) [2], *Pichia canadensis* (27 kb, synonymous name *Hansenula wingei*) [3] and the fission yeast *Schizosaccharomyces pombe* (19 kb). Despite its large size, the *S. cerevisiae* mt genome lacks the seven genes encoding the reduced nicotinamide adenine dinucleotide (NADH) ubiquinone oxidoreductase subunits, found in the 48 kb mt genome of *Y. lipolytica* [2]. The mt genome of *S. cerevisiae* exhibits long segments of non-coding DNA compared to the compact genome of *S. pombe*.

As part of the complete sequencing program of the pathogenic yeast *Candida glabrata* (formerly called *Torulopsis glabrata*), we have determined the complete mt DNA sequence of this organism. With a size of 20 kb, the mt genome of *C. glabrata* is the smallest sequence determined among the hemiascomycetous yeasts. Its small size was already estimated at about 19 kb many years ago [4,5]. Genetic elements present in *C. glabrata* mt genome have been the object of several studies over the last years. Respiratory-defective mutants were isolated, due to large deletions in mt DNA [4] or to loss of the entire mt genome [6], as observed in the rho$^-$ and rho^0 mutants of *S. cerevisiae*, respectively. A recent study performed on clinical isolates led to identification of an azole-resistant petite mutant [7]. The authors suggest that this petite mutant was selected by fluconazole therapy. The first detailed physical map of *C. glabrata* mt DNA was first elaborated by Sriprakash [6], and a completed map, integrating tRNA gene locations, has been published by Clark-Walker et al. [8]. Transcripts from *C. glabrata* mt DNA were mapped by Clark-Walker and Sriprakash, showing evidence of polycistronic transcription [9]. This study was followed by the identification of a nonanucleotide transcriptional control signal in intergenic regions, showing striking similarity with the known consensus sequence of *S. cerevisiae* [8]. In the same study, a *C. glabrata*-specific dodecanucleotide motif involved in transcript processing was also characterized. The authors also identified 23 tRNA genes by sequence comparison with *S. cerevisiae* tRNA sequences [8]. An RNase P activity was detected in mt extracts of *C. glabrata*, and Shu et al. have proposed a 227 mt sequence from the *C. glabrata* mt genome to be the RNA subunit of the mt RNase P enzyme [10]. The VAR1 gene of *C. glabrata*-type strain CBS 138 was previously characterized by Ainley et al. [11]. In *S. cerevisiae*, VAR1 contains a variable number of Asn codon repeats. The *C. glabrata* VAR1 shows a high similarity with the sequence of a small size allele, *var1[40.0]*, of *S. cerevisiae*.

From the complete 20 063 bp circular sequence determined in this work, we report the presence of eight protein-coding genes: respectively *VAR1* for the ribosomal protein var1p, *ATP6*, *ATP8*, *ATP9*, three ATP synthase subunits, *COB* for the apocytochrome b and *COX1*, *COX2*, *COX3* for the cytochrome c oxidase subunits I, II, III. Like *S. cerevisiae*, *C. glabrata* mt DNA does not contain genes encoding NADH: ubiquinone oxidoreductase (complex I), whereas such genes are found in other *Candida* species. The *COX1*

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

[1] These authors contributed equally to this work.

0014-5793/02/$22.00 © 2002 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.
doi:10.1016/S0014-5793(02)03749-3

gene contains three group I introns, named here Cg.COX1.1, Cg.COX1.2, Cg.COX1.3, each with an intronic open reading frame (ORF) of the dodecapeptide family (for a review see [12]). This genome also contains 23 tRNA genes, the large and small rDNA subunits (LSU and SSU) and the RNA subunit of the mt ribonucleoprotein enzyme RNase P enzyme. No replication origin of the *S. cerevisiae* type was found.

2. Materials and methods

2.1. Strain and cultures

The type strain of *C. glabrata* (CBS 138 = ATCC 2001 = IFO 0622) initially isolated from human feces was chosen for this project. Culture of *C. glabrata* was done in YPGlu medium (Bacto Yeast Extract Difco 1%, Bactopeptone Difco 1%, glucose 2%) at 30°C.

2.2. Genomic library construction

Library construction was essentially as described in Blandin et al. [13]. Briefly, a random genomic library of the *C. glabrata*-type strain CBS 138 was made by nebulization of purified total DNA (DNA Nebulizer, GATC, Germany). The DNA fragments were purified from agarose gel and ligated into the unique *Sma*I site of the *Escherichia coli* plasmid vector pBAM3 (derivative from pBluescript KS and constructed by R. Heilig). Actual size distribution of inserts was estimated on a random sample of 96 clones whose plasmid DNAs were purified, digested, and electrophoresed as described by Blandin et al. [13]. No empty vector was found and 86.7% of the clones contained inserts 3–5 kb long (average size of all inserts is 3.8 kb).

2.3. Sequencing reactions and data processing

The complete *C. glabrata* genome sequence was determined by a shotgun strategy (to be published elsewhere). The sequencing was performed by both the Genoscope and the Génopole of the Institut Pasteur. The sequencing strategy used by the Genoscope is described in [14], and all sequences were performed on LiCor 4200L DNA sequencers (dye primers). The Genomics Platform sequencing strategy is as follows. Plasmid DNA purifications were performed using the Montage Plasmid Miniprep$_{96}$ kit (Millipore). Sequencing reactions were performed, from both ends of DNA plasmid, using ABI Prism BigDye Terminator cycle sequencing-ready reaction kits and run on a 3700 Genetic Analyzer (Applied Biosystems). The trace files were base-called using *phred* [15,16] and a quality file for each sequence was obtained. We retain sequences with segments of at least 100 bp called with a quality over 20 per base. In order to eliminate empty vectors and possible contaminations by non-*C. glabrata* DNA, we used the *cross_match* program [15] and *blastn* [17] to compare with sequences of the sequencing-vector, *E. coli* genome, transposons (like Tn10), lambda phage and a few other potential contaminants. Sequences showing a 100 bp long segment with 85% identity or more were eliminated. Possible cross-well contaminations were searched by *blastn* comparisons of each sequence to the sequences obtained from the neighboring wells. In case of significant match (> 99% identity over > 80% of the sequence) one of the duplicated sequences was eliminated.

2.4. Sequence assembly

blastn comparison of the *C. glabrata* 188 000 reads with the mt sequences from *S. cerevisiae*, *P. canadensis*, *Y. lipolytica*, *C. albicans* and *S. pombe* (accession numbers: NC_001138, NC_001762, NC_002659, NC_0002653 and NC_001326, respectively) was per-

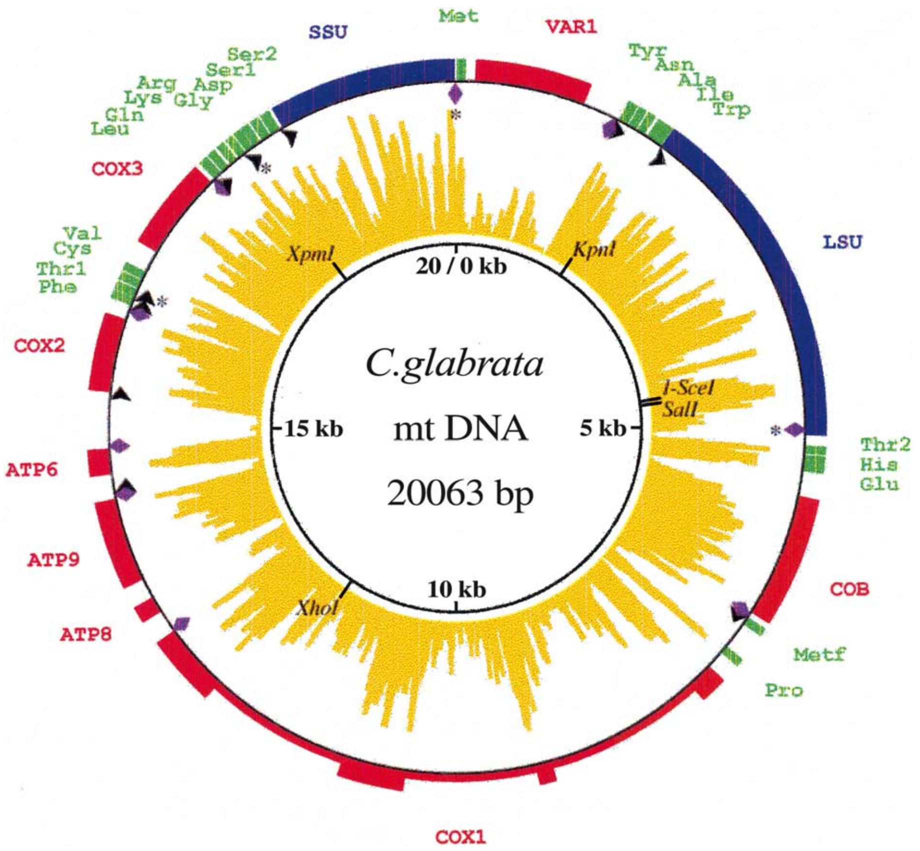

Fig. 1. Sequence-derived map of the mitochondrial genome of *C. glabrata*. The protein and large RNA coding genes are represented respectively with red and blue large blocks. tRNA genes (green) are indicated by their amino acid. The black arrow flags represent the transcriptional unit origins. The purple diamond flags are the transcription processing motifs. A '*' is added for degenerated motifs (see Table 5A,B). Thin blocks symbolize introns in the *COX1* gene. The origin of the coordinates was placed 9 bp after the end of *SSU* and 12 bp before tRNA-met1, because no unique restriction site was found in intergenic regions. All genes are transcribed on the same DNA strand. The inner yellow motif represents GC content, with a scale from 0 to 46% and a mean of 17.7%.

formed to identify the *C. glabrata* mt sequences. Sequence assembly was performed using the *Phred Phrap consed* package with the 2704 mt sequences identified [15,16,18].

2.5. Annotation

We used the *blastx* program [17] with the mt genetic code of *S. cerevisiae* to compare the *C. glabrata* mt sequence with the available mt proteins from *S. cerevisiae*, *P. canadensis*, *Y. lipolytica*, *C. albicans* and *S. pombe*. The RNA genes were identified by comparison with *S. cerevisiae* using the *blastn* program. The RNase P subunit was placed at the position described by Shu et al. [10]. For tRNA genes we used *blastn* alignments to compare our mt sequences with *S. cerevisiae* tRNA genes, and we also used the program *FAStRNA* [19].

3. Results and discussion

3.1. General organization

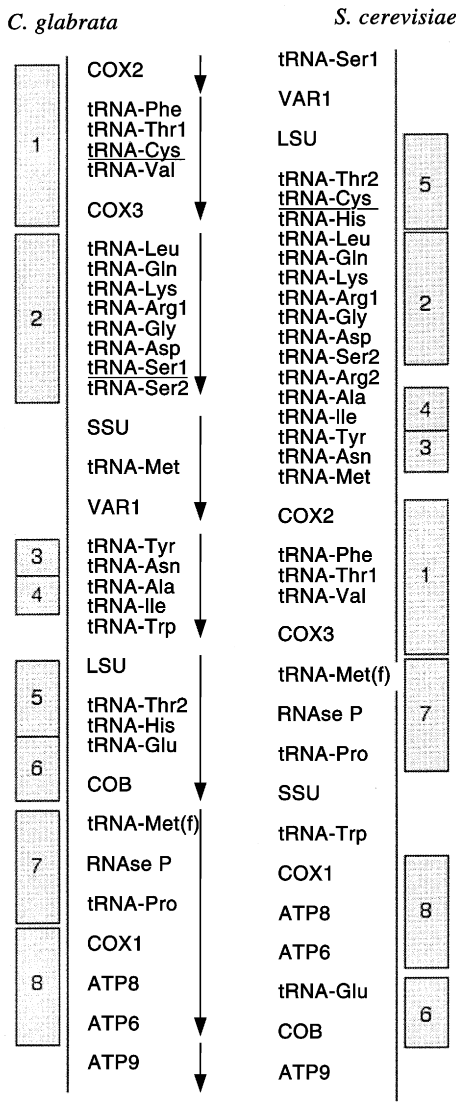

Fig. 2. Comparison of *C. glabrata* and *S. cerevisiae* mt genome maps. Conserved gene series are represented by gray numbered boxes in each genome. Underlined gene names represent the translocation of a single gene in an otherwise conserved block. Arrows indicate positions of identified *C. glabrata* transcriptional units and RNA processing signals. Degenerated motifs are not shown.

Table 1
Genes of *C. glabrata* mitochondrial DNA

Genetic element	Intron	Start	End
tRNA-met1		13	86
VAR1		181	1 200
tRNA-tyr		1 558	1 643
tRNA-asn		1 661	1 732
tRNA-ala		1 752	1 823
tRNA-ile		1 858	1 930
tRNA-trp		1 936	2 006
LSU		2 016	5 108
tRNA-thr2		5 203	5 275
tRNA-his		5 290	5 362
tRNA-glu		5 366	5 437
COB		5 672	6 829
tRNA-met2		6 889	6 959
RNase P RNA		6 974	7 200
tRNA-pro		7 219	7 290
COX1 (exon1)		7 463	7 702
	CgCox1.1	7 703	9 162
I-CgII		7 703	8 818
COX1 (exon2)		9 163	9 308
	CgCox1.2	9 309	10 503
I-CgIII		9 310	10 443
COX1 (exon3)		10 504	11 088
	CgCox1.3	11 089	12 422
I-CgIIII		11 090	11 929
COX1 (exon4)		12 423	13 056
ATP8		13 275	13 421
ATP6		13 614	14 396
ATP9		14 631	14 861
COX2		15 384	16 067
tRNA-phe		16 211	16 282
tRNA-thr1		16 292	16 364
tRNA-cys		16 383	16 455
tRNA-val		16 471	16 543
COX3		16 756	17 565
tRNA-leu		17 616	17 697
tRNA-gln		17 715	17 785
tRNA-lys		17 818	17 889
tRNA-arg		17 915	17 987
tRNA-gly		18 004	18 075
tRNA-asp		18 085	18 156
tRNA-ser1		18 180	18 262
tRNA-ser2		18 281	18 367
SSU		18 404	20 053

All the cox1 introns are group I introns and contain an ORF. All genes are oriented clockwise in Fig. 1. Coordinates of protein-coding genes correspond to the start and stop codons. Coordinates of tRNA genes are deduced for secondary structure (Fig. 3). Coordinates of other RNA genes are determined by homology to *S. cerevisiae* (rRNA) or from literature (RNase P [10]). Coordinates of intron-encoded proteins start arbitrarily at the first codon of the intron in phase with the stop codon of the intronic ORF.

The mt genome of *C. glabrata* (strain CBS 138) consists of a 20 063 bp circular DNA molecule (Fig. 1). The resulting sequence presents a high quality, due to a 90× coverage except for a 100 bp region in the *COX1* gene. A specific polymerase chain reaction (PCR) amplification was performed on this region and the resulting PCR product was sequenced to finalize the sequence assembly. The mt genome sequence contains four genes encoding subunits of respiratory chain complexes, the three ATPase subunits, *VAR1*, the large and the small ribosomal RNA genes, the RNase P ribosomal subunit gene and 23 tRNA genes (Table 1). All of these genetic elements are located on the same DNA strand, as often found in mt genomes, except for *S. cerevisiae* which contains a single tRNA gene on the opposite strand [1]. 32.0% of this genome corresponds to protein coding exons, 8.4% to the 23 tRNA genes and 24.8% to the ribosomal RNAs subunits. The three *COX1* introns represent 19.9% of the genome. The G+C con-

Table 2
Amino acid sequence identity between *C. glabrata* mitochondrial protein-encoding genes and other yeasts

Gene name	Sc	Pc	Yl	Ca	Sp
ATP6	73.1	56.5	48.1	45.4	43.6
ATP8	89.6	75.0	50.0	45.8	50.0
ATP9	82.9	80.3	78.9	76.3	59.2
COB	87.3	76.4	62.9	ND	55.0
COX1	91.2	83.7	64.6	ND	60.0
COX2	88.8	76.1	62.5	60.8	48.0
COX3	82.9	71.1	56.1	57.2	47.2
VAR1	57.8	24.3	–	–	–

The *C. glabrata* mitochondrial proteins have been compared to the corresponding proteins from *S. cerevisiae* (Sc), *P. canadensis* (Pc), *Y. lipolytica* (Yl), *C. albicans* (Ca) and *S. pombe* (Sp). The sequence alignments at amino acid level have been performed by clustalw with default parameters. Figures are in % identity. ND means that the corresponding sequence is not available. '–' indicates that the gene is absent.

tent is 17.6% for the whole genome, and 21.9% for the exonic sequences.

Comparison of *C. glabrata* with the *S. cerevisiae* mt map [1] shows eight major conserved gene blocks that have been rearranged, while a few other genes have also been translocated (Fig. 2). Three blocks contain only tRNA genes (e.g. blocks 2, 3 and 4), others contain tRNA genes and/or ORFs (e.g. blocks 1 and 6). Block 8 contains *COX1*, *ATP8* and *ATP6* and is conserved among *Hemiascomycetes*, as in the genus *Saccharomyces* [20] or in *Y. lipolytica* [2]. Finally, the RNase P gene is found in block 7 between the same two tRNA genes (M and P) as in *S. cerevisiae*. Note that the blocks defined by comparative analysis essentially coincide with the transcriptional units identified in the *C. glabrata* mt genome (see Section 3.6).

3.2. Mitochondrial protein coding genes

The *C. glabrata* mt genome contains eight protein coding genes, recognizable by their ORFs using the correct genetic code (see below) and by sequence comparisons with mt genomes of other *Ascomycetes*. The genes found encode apocytochrome b (*COB*), ATP synthase subunits 6, 8 and 9 (*ATP6*, *ATP8* and *ATP9*) and the cytochrome c oxydase subunits I, II and III (*COX1*, *COX2* and *COX3*). Average amino acid sequence conservation of each *C. glabrata* gene product with homologs from other yeast species is indicated by Table 2. The *COX1*, *COX2* and *ATP8* gene products are less divergent than the others. *VAR1* is the least conserved. *C. glabrata* is more closely related to *S. cerevisiae* than to the other yeast species, including *C. albicans*. This is in accordance with the notion that *C. glabrata* is not closely related to most other species designated as *Candida* [21].

Interestingly, by comparing with other yeasts, we found a frame shift due to the insertion of a single nucleotide (C) in the *COX2* ORF at position 544 of the nucleotidic sequence. Note that the *COX2* gene previously sequenced by Clark-Walker and Weiller [22] does not contain this frame shift (four other nucleotides also differ between our two sequences at positions 51, 52, 269 and 290, accession number: X69430) although the same strain as ours was used. A recent study on clinical *C. glabrata* isolates shows the same C insertion in all *COX2* ORFs sequenced, and that this C is also present in *COX2* mtRNA [38]. The C insertion at position 544 leads to a 227 amino acids long protein, whereas other COX2 proteins are about 252 amino acids long. The downstream part of the *C. glabrata COX2* gene after the frame shift is well-conserved with the other yeasts, suggesting that translation continues in the appropriate frame. Alternatively, it is possible that insertion of the C at position 544 is a very recent event in

Table 3
Genetic code in *C. glabrata* mitochondria and comparison of potential variant codons between ORFs of *C. glabrata* and *S. cerevisiae*

Codon	Total	Alignable positions with Sc	At aligned positions	
			Conserved in Sc	Other in Sc
TGA				
var1	3	3	3	
cob	8	8	7	TAT (Y)
cox1 exons	10	10	10	
atp6	4	4	4	
cox2	5	5	5	
cox3	9	9	9	
Total	39	38	38	
ATA				
var1	14	8	7	AAA (K)
atp8	1	1	1	
atp6	1	1	0	TTT (F)
Total	16	10	8	
CTT				
var1	1	1	0	GAA (E)
cob	1	1	0	CTA (T)
Total	2	2	0	
CTA				
var1	2	2	0	ATA (M) CTT (T)
cob	2	2	2	
cox1 exons	1	1	0	ATT (I)
atp6	1	1	1	
cox2	1	1	0	ACT (T)
cox3	1	1	0	TTA (L)
Total	8	8	3	

For each gene, we searched if the TGA, ATA, CTT and CTA codons were conserved between the two species at alignable positions. For the COX1 gene, only exonic sequences were considered.

the *C. glabrata* line of ancestry, leading to a truncated *COX2* protein.

3.3. Genetic code and codon usage

The mt genetic code is known to differ from the universal code in *S. cerevisiae* and other yeasts [23]. A few specific codons are affected. In all yeasts, the TGA codon specifies tryptophan. In *S. cerevisiae* and some other yeast mitochondria, ATA specifies methionine, whereas in other species, such as *Y. lipolytica*, ATA specifies isoleucine as in the universal code. Finally, in *S. cerevisiae* and a few related species the four CTN codons are used to translate threonine instead of leucine, due to an abnormal tRNA with eight nucleotides in the anticodon loop [24].

In order to deduce the genetic code used in the mitochondria of *C. glabrata*, we have taken advantage of the high amino acid conservation between *C. glabrata* and *S. cerevisiae* mt proteins (see Table 2) to align the corresponding genes and examine codon replacements. Results are given in Table 3. Of the 39 TGA codons used in *C. glabrata*, 38 are in conserved positions in *S. cerevisiae* and encode a tryptophan. We conclude that in *C. glabrata* mitochondrion TGA specifies tryptophan. The situation is more complex for the ATA codon, as 14 of its 16 occurrences in *C. glabrata* mitochondria are in the *VAR1* gene, which is the least conserved with *S. cerevisiae* (see Table 2). Nevertheless, eight positions in this gene could be aligned, of which seven are conserved in *S. cerevisiae*. The other two occurrences of the ATA codon are in the *ATP8* and *ATP6* genes. The first case is conserved in *S. cerevisiae* while the other is replaced by TTT (phenylalanine). In total, eight ATA codons are conserved between *C. glabrata* and *S. cerevisiae*, suggesting that the ATA codon specifies methionine in *C. glabrata* as in *S. cerevisiae*.

Finally, a total of eight CTA and two CTT codons have been observed (Table 3). The two CTT codons are not conserved in *S. cerevisiae*, but one is replaced by CTA encoding threonine. Only three of the eight CTA codons are conserved in *S. cerevisiae*, one is replaced by CTT encoding threonine and, interestingly, another one by ACT, also encoding threonine. Thus, we regard as most probable that the CTN family encodes threonine in *C. glabrata* as in *S. cerevisiae* (see also below for the analysis of tRNA genes). In conclusion, *C. glabrata* presents the same three deviations from the universal code as *S. cerevisiae*.

Codon usage in *C. glabrata* exonic ORFs shows a strong bias towards codons ending in U or A, which increases considerably in intronic ORFs (Table 4). This bias is so strong that *C. glabrata* actually uses only 41 of the 62 possible codons in its coding exons (44 if one considers the intronic ORFs). In addition, *C. glabrata* lacks the capacity to read the four codons of the CGN family, in agreement with the lack of tRNA specific for this family (see below). Incomplete genetic codes are also observed for mitochondria of other organisms but were not described before for a yeast.

3.4. tRNA genes

Sequence comparisons with *S. cerevisiae* lead us to identify a set of 23 tRNA genes (Fig. 3), all on the same DNA strand as other genes. Sixteen of these 23 tRNA genes could also be identified using the *FAStRNA* program [19]. *C. glabrata* has essentially the same set of mt tRNAs as *S. cerevisiae* [1] except for the complete absence of tRNA Arg2 (anticodon ACG), which is consistent with the lack of CGN codons (Table 4). Sequences of tRNA genes of *C. glabrata* are highly conserved with *S. cerevisiae*, particularly the anticodon sequences that are all conserved. Interestingly, the abnormal tRNA-thr1 of *S. cerevisiae*, which has eight nucleotides [25] in the anticodon loop and reads the CTN codon family as threonine, is also highly conserved in *C. glabrata*. As in *S. cerevisiae*, this tRNA gene possesses eight nucleotides in its anticodon loop (Fig. 3) and is the sole tRNA encoded in the mitochondrion able to read the CTN codon family. Contrary to *S. cerevisiae*, however, the gene encoding this tRNA is oriented like all the other mt genes in *C. glabrata*. As in *S. cerevisiae*, the tRNA-thr1 gene is located between the two tRNA Phe and tRNA Val genes. Inversion of the tRNA Thr1 gene may have

Table 4
Codon usage in protein coding genes of *C. glabrata* mitochondria

Codon	AA	exon	intron	Codon	AA	exon	intron	Codon	AA	exon	intron	Codon	AA	exon	intron
TTT	F	86	35	TCT	S	32	15	TAT	Y	118	104	TGT	C	12	8
TTC	F	62	1	TCC	S	–	–	TAC	Y	7	3	TGC	C	2	1
TTA	L	297	120	TCA	S	81	16	TAA	*	8	1	TGA	W	39	12
TTG	L	–	1	TCG	S	–	–	TAG	*	–	2	TGG	W	–	–
CTT	T	2	9	CCT	P	45	11	CAT	H	49	10	CGT	R	–	–
CTC	T	–	–	CCC	P	1	1	CAC	H	–	–	CGC	R	–	–
CTA	T	8	8	CCA	P	34	2	CAA	Q	47	5	CGA	R	–	–
CTG	T	–	–	CCG	P	1	–	CAG	Q	2	1	CGG	R	–	–
ATT	I	209	111	ACT	T	49	11	AAT	N	178	222	AGT	S	30	14
ATC	I	21	4	ACC	T	–	1	AAC	N	6	2	AGC	S	–	1
ATA	M	16	69	ACA	T	49	4	AAA	K	72	118	AGA	R	45	15
ATG	M	74	5	ACG	T	–	–	AAG	K	3	3	AGG	R	–	–
GTT	V	42	3	GCT	A	70	3	GAT	D	47	27	GGT	G	87	18
GTC	V	–	–	GCC	A	4	1	GAC	D	–	–	GGC	G	–	–
GTA	V	96	9	GCA	A	46	1	GAA	E	15	18	GGA	G	46	3
GTG	V	2	1	GCG	A	–	–	GAG	E	–	–	GGG	G	–	–

The codon content was calculated separately from the eight protein coding genes (exon) and the three COX1 intronic ORFs (intron). Note the complete absence of CGN codons and of 14 other codons (see text). Except for the TTY and ATR codons in exons, there is always a considerable excess of codons ending with A or T in their third position among synonyms.

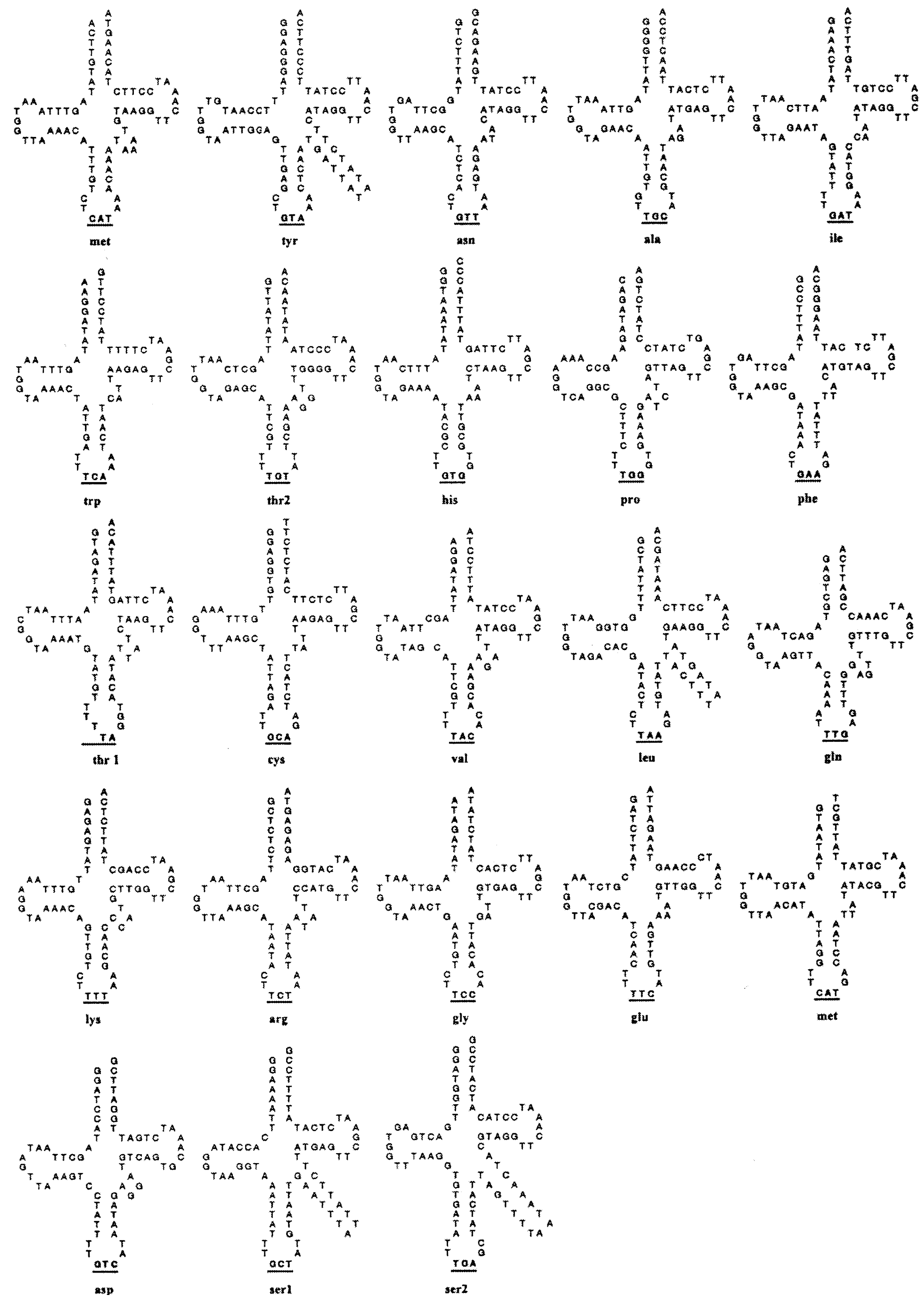

Fig. 3. Clover-leaf structures of the 23 mitochondrially encoded tRNAs of *C. glabrata*. Anticodons are underlined and the corresponding amino acids are indicated. tRNAs are listed according to the map positions of their gene (see Fig. 1). Note the absence of arg tRNA for the CGN codons, and the presence of an extranucleotide in the anticodon loop of tRNA thr1.

Table 5A
Cis-acting conserved and consensus sequences involved in transcription: transcription initiation signal

−1 base	Begin	Central octamer	End	+1 base	Located before
A	1 535	TATAAGTA	1 542	A	Y, N, A, I, W tRNA genes
G	2 008	id	2 015	A	LSU, T2, H, G, COB
T	6 872	id	6 879	A	fM, RNase P, P, COX1, atp8, atp6
A	14 464	id	14 471	A	atp9
A	15 329	id	15 336	A	cox2
A	16 197	id	16 204	A	F, T1, C, V, COX3
A	16 287	id	16 294	G	overlap thr1
T	17 598	id	17 605	A	L, Q, Y, R
G	17 986	id	17 993	A	G (overlap arg), D, S1, S2
A	18 386	id	18 393	A	SSU

Transcription initiation consensus sequences. Just three sequences differ from the WTATAAGTA *S. cerevisiae* consensus sequence, with two of them overlapping other informative sequences.
Coordinates of the first and last nucleotides of the consensus signal refer to Fig. 1 and Table 1.

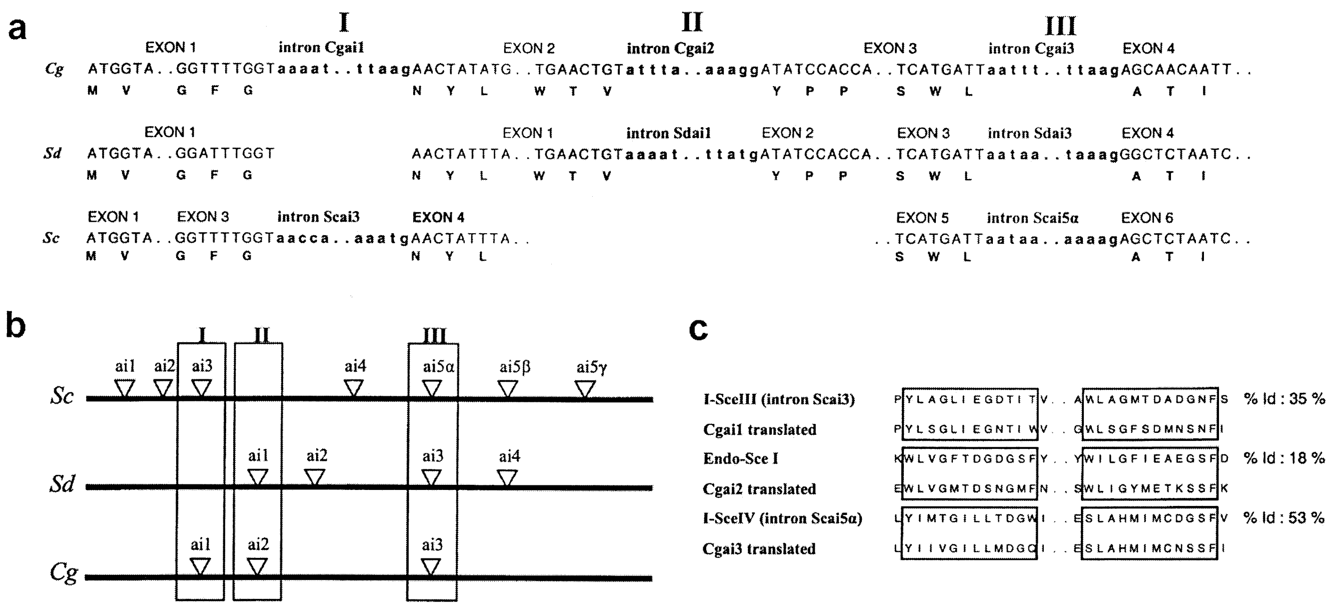
a
I
II
III
EXON 1
intron Cgai1
EXON 2
intron Cgai2
EXON 3
intron Cgai3
EXON 4
Cg ATGGTA..GGTTTTGGTaaaat..ttaagAACTATATG..TGAACTGTattta..aaaggATATCCACCA..TCATGATTaattt..ttaagAGCAACAATT..
M V G F G N Y L W T V Y P P S W L A T I
EXON 1
EXON 1
intron Sdai1
EXON 2
EXON 3
intron Sdai3
EXON 4
Sd ATGGTA..GGATTTGGT AACTATTTA..TGAACTGTaaaat..ttatgATATCCACCA..TCATGATTaataa..taaagGGCTCTAATC..
M V G F G N Y L W T V Y P P S W L A T I
EXON 1
EXON 3
intron Scai3
EXON 4
EXON 5
intron Scai5α
EXON 6
Sc ATGGTA..GGTTTTGGTaacca..aaatgAACTATTTA.. ..TCATGATTaataa..aaaagAGCTCTAATC..
M V G F G N Y L S W L A T I
b
I
II
III
ai1
ai2
ai3
ai4
ai5α
ai5β
ai5γ
Sc
ai1
ai2
ai3
ai4
Sd
ai1
ai2
ai3
Cg
c
I-SceIII (intron Scai3) PYLAGLIEGDTITV..AWLAGMTDADGNFS % Id : 35 %
Cgai1 translated PYLSGLIEGNTIWV..GWLSGFSDMNSNFI
Endo-Sce I KWLVGFTDGDGSFY..YWILGFIEAEGSFD % Id : 18 %
Cgai2 translated EWLVGMTDSNGMFN..SWLIGYMETKSSFK
I-SceIV (intron Scai5α) LYIMTGILLTDGWI..ESLAHMIMCDGSFV % Id : 53 %
Cgai3 translated LYIIVGILLMDGQI..ESLAHMIMCNSSFI

Fig. 4. *COX1* gene. a: Intronic insertion sites along *C. glabrata COX1* gene, compared to *S. douglasii* (Sd) and *S. cerevisiae* (Sc). Capital letters are for exon sequences and bold lower case letters for intron sequences. Amino acid translation of exons is indicated. Cgai1, 2, 3 are for introns Cg Cox1.1, 1.2 and 1.3 of *C. glabrata* according to the nomenclature of Sc and Sd. b: Insertion sites of introns in the *COX1* gene in various yeast species. This diagram is adapted from Tian et al. [33]. Arrows indicate where the introns are inserted in *S. cerevisiae* [30], *S. douglasii* [33] and *C. glabrata*. Sites corresponding to *C. glabrata* introns are numbered. c: Alignment of the dodecapeptide motifs found in the *C. glabrata COX1* intronic ORF products with endonucleases encoded in *S. cerevisiae* mitochondria. The overall amino acid sequence identity is indicated for each alignment. *I-CglII* intronic proteins align weakly but enough to allow identification of dodecapeptide motifs.

been correlated with the insertion of the tRNA Cys gene in *C. glabrata*. All of these tRNA genes were already reported [8] by homology comparison with *S. cerevisiae*.

3.5. Other RNA genes

The well-conserved genes for the ribosomal RNA subunits (LSU for 21S and SSU for 15S) were identified by comparison with other species. A 227-nucleotide long RNA gene located between tRNA-fMet and tRNA-Pro genes was proposed to encode the RNA subunit of the *C. glabrata* mt ribonucleoprotein enzyme Ribonuclease P (RNase P) required for 5′ end maturation of mt tRNAs [10]. It has been established that RNase P was cotranscribed with its flanking tRNA gene in *C. glabrata* [26].

3.6. Transcriptional units

All mt genes of *C. glabrata* have the same orientation. Transcription of *S. cerevisiae* mt DNA is polycistronic and is initiated at several sites characterized by the 5′-WTATAAGTA-3′ consensus sequence [27,28]. The same consensus sequence is found in *C. glabrata*, as previously reported by Clark-Walker et al. [8]. Seven of these conserved motifs occur on the Watson strand (Table 5A), two additional ones can be found if one accepts a G as the first position and a third additional one with a G as the last position. This last motif also overlaps 3 bp of the tRNA Thr1 5′ end. Similarly, the motif found at position 17 986 also overlaps the 3′ end of the tRNA Arg gene with 3 bp. All other motifs are in intergenic regions. Unlike *S. cerevisiae*, which carries thr1-tRNA on the Crick strand, in *C. glabrata* the consensus sequence WTATAAGTA is only found on the Watson strand, suggesting that no transcription occurs on the Crick strand. When placed on the map, and considering only the non-overlapping consensus transcriptional motifs, we can define eight transcriptional units in *C. glabrata* mitochondria (Fig. 1). This is to be compared with the 19 active transcript initiation sites de-

Table 5B
Cis-acting conserved and consensus sequences involved in transcription: transcript processing signal

Begin	Dodecanucleotide sequence	End	Located after
1 521	TATAATATTCTT	1 532	var1
6 852	id	6 863	cob
13 077	id	13 088	cox1
14 426	id	14 437	atp6
14 895	id	14 906	atp9
16 175	id	16 186	cox2
17 580	id	17 591	cox3
Degenerated motifs:			
5 049	TATAATATcta	5 060	LSU
20 042	TtTAATATTCTT	20 053	SSU

Transcript processing signal: the dodecanucleotide sequence found in *S. cerevisiae* is also present seven times in *C. glabrata*, with two similar degenerated sequences at the end of rRNA genes.
Coordinates of the first and last nucleotides of the consensus signal refer to Fig. 1 and Table 1.

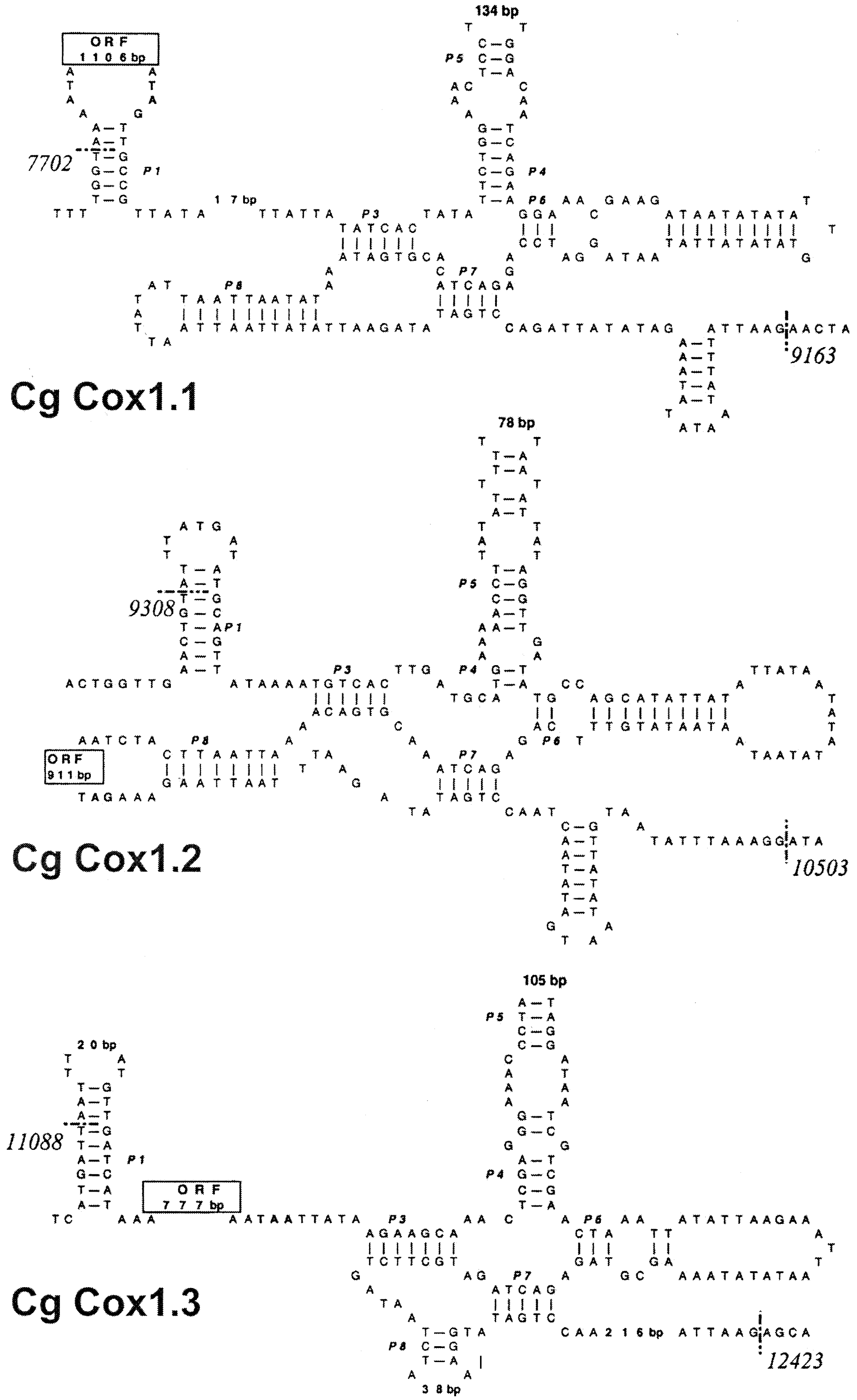

Fig. 5. Predicted secondary structure of the three group I introns of the *C. glabrata COX1* gene. The exon/intron limits are represented by dotted lines and numbers corresponding to the coordinates in the mt genome (refer to Fig. 1). The three introns each contain a long ORF, fused in frame with the upstream exon and terminating in an unstructured region of the intron, designated as 'ORF'. Corresponding size of the non-represented sequences is indicated in bp. Base pairs are represented by bars. Stems are designated according to standard group I intron 1 nomenclature [37].

scribed in *S. cerevisiae* [1], a significant difference that may be correlated with the genome size difference.

Furthermore, a transcript processing signal composed of 12 nucleotides similar to those found in *S. cerevisiae* [29] was identified by Clark-Walker et al. [8]. This dodecanucleotide motif is found seven times in *C. glabrata*, and two degenerated sequences with one single base change were also reported [8] (Table 5B). Interestingly, these two degenerated motifs are located at the end of LSU and SSU genes, respectively, while the seven non-degenerated motifs are located after the stop codon of major protein coding genes (*COB*, *COX1*, *ATP6*, *ATP9*, *COX2*, *COX3* and *VAR1*). No such motif is found after *ATP8*. Once again, all the motifs are found on the same DNA strand.

If one assumes that the transcript processing signal immediately upstream of the next origin is the transcript termination, the longest polycistronic transcript of *C. glabrata* appears to be the 7600 nucleotide long one starting with the tRNA-fmet gene followed by RNAse P gene, tRNA Pro gene, *COX1*, the *ATP8* and *ATP6*. Another long polycistronic transcript (ca. 4800 nucleotides) contains LSU, three tRNA genes and *COB*, and another one (ca. 3200 nucleotides) contains SSU and *VAR1*. Two other polycistronic transcripts each include a cluster of tRNA genes only (L, Q, K, R, G, N, S, S and Y, N, A, I, W, respectively) or in addition to *COX3* (F, thr1, C, V). There are two monocistronic transcripts corresponding to *ATP9* and *COX2*, respectively. If one ignores the degenerated processing sites, only one transcript (tRNA Met(f) to *ATP6*) contains an internal processing site (between *COX1* and *ATP8*).

3.7. Group I introns

The *COX1* gene product is highly conserved among yeast species (Table 2). This gene generally contains a number of group I or group II introns. In *S. cerevisiae* up to seven introns have been found in the *COX1* gene [30,31] in different laboratory strains. Three introns were found in *Kluyveromyces lactis* [32] and four in *Saccharomyces douglasii* [33]. Introns can be inserted at conserved positions in all species (e.g. *ai5alpha* between *S. cerevisiae*, *K. lactis*, or *S. douglasii* [33]), or be specific to a particular species (e.g. intron *ai1* of *S. cerevisiae*). *COX1* introns belong to the group I (e.g. *S. cerevisiae ai3*, *ai4*, *a15alpha*, *ai5beta*, *ai5gamma*) or to the group II (e.g. *S. cerevisiae ai1*, *ai2*) of self-splicing introns ([34] and for review, see [35]). In both groups, intronic RNA acquires complex secondary structures leading to the formation of a catalytic core necessary for the two transesterification steps [36]. Comparative sequence studies and mutagenesis experiments have been performed on the numerous group I and group II sequences from a variety of organisms, leading to the precise formulation of detailed rules about the secondary structures of both group I and group II introns (reviewed by [36,37]). Group I introns often contain long intronic ORFs, frequently in phase with the upstream exon and coding for maturases or DNA endonucleases. Group I intron-encoded proteins can be classified in four families, characterized by distinct consensus motifs [35].

In *C. glabrata*, the *COX1* gene contains three introns (Fig. 4a,b) that we have designated CgCox1.1, CgCox1.2 and CgCox1.3. Each of them contains an ORF, in frame with the upstream exon. The three introns can be folded into RNA secondary structures (Fig. 5) which possess the characteristic features of group I introns [37]. In all cases the upstream exon ends with a T which, upon folding of the P1 stem, pairs with a G inside the intron, as typical for group I introns. The P4, P6 and P8 stems are recognizable as well as the P3–P7 pseudoknot. The introns end with a G, as required for the second step of the splicing mechanism. The major part of the intronic ORFs occurs in different loops in the three introns (Fig. 5). The intronic ORFs encode lysine and leucine-rich polypeptides in which dodecapeptide motifs can be recognized. When aligned with various intronic ORFs from the dodecapeptide family coding for meganucleases in *S. cerevisiae*, the translated ORFs from CgCox1.1 and CgCox1.3 match significantly with I-SceIII and I-SceIV, respectively (Fig. 4b). CgCox1.2 does not align with more than 25% identity with any group I intronic ORF from *S. cerevisiae*, but two dodecapeptide motifs were nevertheless identified.

4. Concluding remarks

The *C. glabrata* mt genome appears to be the smallest known so far among hemiascomycetous yeasts. Despite its small size, we observed similar gene content compared to *S. cerevisiae*. The replication origin of *C. glabrata* mt genome remains unknown. Clark-Walker et al. have suggested, by analogy with *E. coli*, that the A+T-rich decanucleotide region between the *ATP9* and *COX2* genes may be an origin for mt DNA replication in *C. glabrata* [8]. Because *C. glabrata* presents a lack of genetic markers, few genotyping studies have been performed. In a recent study dealing with South and North American strains, Sanson and Briones used the sequence of the mt *COX2* gene. They were able to distinguish two clusters of *C. glabrata* strains in accordance with their geographical origin [38]. The complete mt genome could provide new potentialities for molecular genotyping.

Acknowledgements: We thank H.-V. Nguyen for providing the type strain of *C. glabrata*. We thank our colleagues of Unité de Génétique Moléculaire des Levures for fruitful discussions, and Laurence Ma, Magali Tichit and Eliana Ochoa for their technical assistance. We also thank Véronique Anthouard, Marc Baudoin, Bernard Caudron and Fredj Tekaia for help in the computing process, and Jean Weissenbach for his continuous support of the yeast sequencing programs. B.D. is a member of Institut Universitaire de France. R.K. is recipient of a CNRS BDI fellowship. This work is supported by CNRS GDR 2354.

References

[1] Foury, F., Roganti, T., Lecrenier, N. and Purnelle, B. (1998) FEBS Lett. 440, 325–331.

[2] Kerscher, S., Durstewitz, G., Casaregola, S., Gaillardin, C. and Brandt, U. (2001) Comp. Funct. Genom. 2, 80–90.

[3] Sekito, T., Okamoto, K., Kitano, H. and Yoshida, K. (1994) Nucleic Acids Symp. Ser. 31, 233–234.

[4] O'Connor, R.M., McArthur, C.R. and Clark-Walker, G.D. (1976) J. Bacteriol. 126, 959–968.

[5] Clark-Walker, G.D. and Sriprakash, K.S. (1981) J. Mol. Biol. 151, 367–387.

[6] Sriprakash, K.S. and Batum, C.M. (1981) Curr. Genet. 4, 73–80.

[7] Bouchara, J.P., Zouhair, R., Le Boudouil, S., Renier, G., Filmon, R., Chabasse, D., Hallet, J.N. and Defontaine, A. (2000) J. Med. Microbiol. 49, 977–984.

[8] Clark-Walker, G.D., McArthur, C.R. and Sriprakash, K.S. (1985) EMBO J. 4, 465–473.

[9] Clark-Walker, G.D. and Sriprakash, K.S. (1983) EMBO J. 2, 1465–1472.

[10] Shu, H.H., Wise, C.A., Clark-Walker, G.D. and Martin, N.C. (1991) Mol. Cell. Biol. 11, 1662–1667.

[11] Ainley, W.M., Macreadie, I.G. and Butow, R.A. (1985) J. Mol. Biol. 184, 565–576.

[12] Dujon, B. (1989) Gene 82, 92–114.

[13] Blandin, G., Llorente, B., Malpertuy, A., Wincker, P., Artiguenave, F. and Dujon, B. (2000) FEBS Lett. 487, 76–81.

[14] Artiguenave, F. et al. (2000) FEBS Lett. 487, 13–16.

[15] Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) Genome Res. 8, 175–185.

[16] Ewing, B. and Green, P. (1998) Genome Res. 8, 186–194.

[17] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J. Mol. Biol. 215, 403–410.

[18] Gordon, D., Abajian, C. and Green, P. (1998) Genome Res. 8, 195–202.

[19] el-Mabrouk, N. and Lisacek, F. (1996) J. Mol. Biol. 264, 46–55.

[20] Groth, C., Petersen, R.F. and Piskur, J. (2000) Mol. Biol. Evol. 17, 1833–1841.

[21] Kurtzman, C.P. and Robnett, C.J. (1997) J. Clin. Microbiol. 35, 1216–1223.

[22] Clark-Walker, G.D. and Weiller, G.F. (1994) J. Mol. Evol. 38, 593–601.

[23] Jukes, T.H. and Osawa, S. (1990) Experientia 46, 1117–1126.

[24] Li, M. and Tzagoloff, A. (1979) Cell 18, 47–53.

[25] Dirheimer, G. and Martin, R.P. (1990) in: Chromatography and Modifications of Nucleosides, Part B; Biological Roles and Function of Modification (Gehrke, G.K. and Kuo, S.K., Eds.), pp. B197–B264, Elsevier, Amsterdam.

[26] Shu, H.H. and Martin, N.C. (1991) Nucleic Acids Res. 19, 6221–6226.

[27] Christianson, T. and Rabinowitz, M. (1983) J. Biol. Chem. 258, 14025–14033.

[28] Osinga, K.A., De Vries, E., Van der Horst, G.T. and Tabak, H.F. (1984) Nucleic Acids Res. 12, 1889–1900.

[29] Osinga, K.A., De Vries, E., Van der Horst, G. and Tabak, H.F. (1984) EMBO J. 3, 829–834.

[30] Bonitz, S.G., Coruzzi, G., Thalenfeld, B.E., Tzagoloff, A. and Macino, G. (1980) J. Biol. Chem. 255, 11922–11926.

[31] Hensgens, L.A., Bonen, L., de Haan, M., van der Horst, G. and Grivell, L.A. (1983) Cell 32, 379–389.

[32] Hardy, C.M. and Clark-Walker, G.D. (1991) Curr. Genet. 20, 99–114.

[33] Tian, G.L., Michel, F., Macadre, C. and Lazowska, J. (1993) Gene 124, 153–163.

[34] Michel, F. and Dujon, B. (1983) EMBO J. 2, 33–38.

[35] Belfort, M. and Perlman, P.S. (1995) J. Biol. Chem. 270, 30237–30240.

[36] Cech, T.R. (1990) Annu. Rev. Biochem. 59, 543–568.

[37] Michel, R. and Westhof, E. (1990) J. Mol. Biol. 216, 585–610.

[38] Sanson, G.F. and Briones, M.R. (2000) J. Clin. Microbiol. 38, 227–235.