Evolutionary Origin of Nonuniversal CUG^{Ser} Codon in Some Candida Species as Inferred From a Molecular Phylogeny

Graziano Pesole,* Marina Lotti,† Lilia Alberghina† and Cecilia Saccone*

**Dipartimento di Biochimica e Biologia Molecolare and Centro Studi sui Mitocondri e Metabolismo Energetico Consiglio Nazionale delle Ricerche, Università di Bari, 70126 Bari, Italy and †Dipartimento di Fisiologia e Biochimica Generali, Università degli Studi di Milano, 20133 Milano, Italy*

Manuscript received May 1, 1995
Accepted for publication August 5, 1995

ABSTRACT

CUG, a universal leucine codon, has been reported to be read as serine in various yeast species belonging to the genus Candida. To gain a deeper insight into the origin of this deviation from the universal genetic code, we carried out a phylogenetic analysis based on the small-subunit ribosomal RNA genes from some Candida and other related Hemiascomycetes. Furthermore, we determined the phylogenetic relationships between the tRNA^{Ser}CAG, responsible for the translation of CUG, from some Candida species and the other serine and leucine isoacceptor tRNAs in *C. cylindracea*. We demonstrate that the group of Candida showing the genetic code deviation is monophiletic and that this deviation could have been originated more than 150 million years ago. We also describe how phylogenetic analysis can be used for genetic code predictions.

DEVIATION from the universal genetic code was first discovered in 1979 in vertebrate mitochondria (BARREL *et al.* 1979). The myth of a frozen universal genetic code was further shattered by the discovery of genetic code changes in several other nuclear and mitochondrial genomes (OSAWA *et al.* 1990, 1992).

In 1989 KAWAGUKI *et al.* (1989) found that CUG, a universal leucine codon, is assigned as serine in the nuclear code of the yeast *Candida cylindracea*. More recently, OHAMA *et al.* (1993) have shown that the nonuniversal CUG leucine codon is found in various yeast species all belonging to the genus Candida.

In one of our laboratories five lipase (triacylglycerol acyl hydrolase) genes from *C. cylindracea* (also named *C. rugosa*) have been recently sequenced (LOTTI *et al.* 1993). Lipases, which catalyze the breakdown of triacylglycerols to free fatty acids and glycerol, have been classified as serine hydrolases since their active site contains the Ser-His-Asp/Glu catalytic triad with the serine enclosed in a highly conserved (Ala)Gly-X-Ser-X-Gly motif. It was found that, interestingly, the catalytic Ser is always encoded by CTG in all five lipase genes. These findings have stimulated our interest to gain a deeper insight into the origin of the deviation from the universal genetic code in these organisms. To do so it was first necessary to assess their taxonomic position on molecular grounds.

The genus Candida includes many species that are closely related to other yeast species, such as Saccharomyces and Torulopsis, all belonging to the group Hemiascomycetes. The phylogenetic relationships among higher fungi are rather unclear; from molecular analyses conflicting data have been obtained with respect to the traditional nonmolecular classification (BERBEE and TAYLOR 1992).

We report here a phylogenetic analysis of some Candida and other related yeast species based on the 18S rRNA, aimed at clarifying both the phylogenetic position of the Candida species and the origin of the universal genetic code variation. Furthermore, we also determined the phylogenetic relationships between the tRNA^{Ser}CAG, responsible for the translation of CUG in some Candida species and the other serine and leucine isoacceptor tRNAs in *C. cylindracea*, to shed light on the possible mechanisms responsible of the deviation from the universal genetic code.

MATERIALS AND METHODS

The 18S rRNA and the Candida tRNA sequences analyzed in this study have been extracted from release 40 of the EMBL database. Multiple alignments have been obtained with the program PILEUP and manually adjusted after visual inspection with the program LINEUP (GCG 1993). Distance measures between Candida tRNAs have been determined by applying a stochastic model of gene evolution, the Stationary Markov Clock (SMC) model (SACCONE *et al.* 1990). The phylogenetic tree has been calculated from pairwise distances with the program NEIGHBOR and DRAWGRAM of the PHYLIP package (FELSENSTEIN 1993).

The 18S rRNA have been analyzed with the maximum-likelihood method under the molecular clock hypothesis with DNAMLK program of the PHYLIP package (FELSENSTEIN 1993) because they have dissimilar base compositions, *i.e.*, do not obey the stationary condition, a prerequisite for SMC (SACCONE *et al.* 1990). DNAMLK program has been used assuming empirical base frequencies and transitions to be dou-

Communicating editor: Graziano Pesole, Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, via Orabona, 4, 70126 Bari, Italy. E-mail: graziano@area.ba.cnr.it

TABLE 1

List of yeast 18S rRNA used for the phylogenetic analysis

Organism	Accession number	CUG	Genome G+C%[a]
Candida albicans	X53497	Ser[c]	36, 0
Candida glabrata	X51831	Leu[b]	39, 9
Candida guilliermondii	M60304	Ser[b,d]	44, 2
Candida kefyr	M60303	Leu[b]	41, 3
Candida krusei	M55528	Leu[b]	44, 6
Candida lusitaniae	M55526	Leu[b]	44, 7
Candida maltosa	D14593	Ser[b,c]	36, 2
Candida parapsilosis	M60307	Ser[b,c,d]	40, 5
Candida tropicalis	M55527	Ser[b,d]	36, 0
Candida tropicalis	M60308	Ser[b,d]	36, 0
Candida viswanathii	M60309	Ser[b]	43, 3
Galactomyces geotrichum	X69842	Leu[b]	
Hansenula polymorpha	M60310	Leu[b]	48, 1
Kluyveromyces lactis	X51830	Leu[b]	41, 1
Neurospora crassa	X04971	Leu[b,c]	
Candida pelliculosa	X58054	?	36, 4
Pichia membranaefaciens	X58055	Leu[b]	43, 0
Saccharomyces cerevisiae	J01353	Leu[b,c]	40, 0
Schizosaccharomyces pombe	X58056	Leu[b,c]	42, 0
Yarrowia lipolytica	M60312	Leu[b,c]	50, 0

The genome-G+C% and CUG assignment are reported when known.

[a] KREGER-VAN RIJ (1984).

[b] CUG codon assignment prediction based on the evolutionary analysis reported in the present paper.

[c] CUG codon assigned to serine based on *in vitro* or *in vivo* translation assay systems (OHAMA *et al.* 1993; SUZUKI *et al.* 1994; SANTOS and TUITE 1995; SUGIYAMA *et al.* 1995)

[d] CUG codon assigned to serine based on the finding of a $tRNA^{Ser}$(CUG) gene.

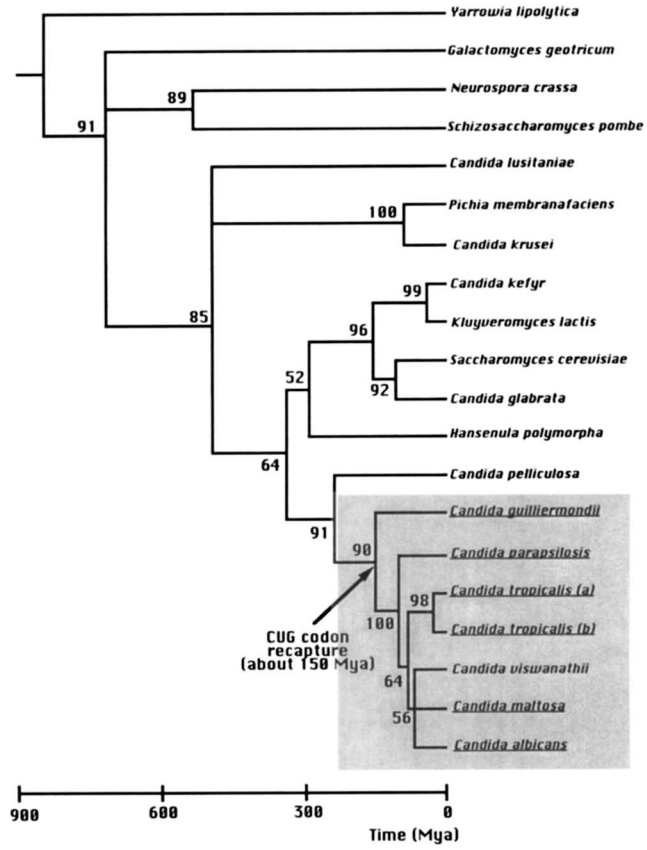

FIGURE 1.—Phylogenetic tree calculated with DNAMLK and DRAWGRAM programs (FELSENSTEIN 1993) on the 20 yeast 18S rRNA listed in Table 1. Branch lengths are proportional to genetic distances. The monophyletic subgroup of Candida (CUG-Ser species underlined) is in the shaded box. Bootstrap values out of 100 replicates are shown on each node. Only the nodes that fulfill the 50% majority rule consensus are retained in the phylogenetic tree; other nodes are collapsed into a single one.

ble the transversions). The statistical significance of the phylogenetic trees has been assessed through the bootstrap procedure carried out using SEQBOOT and CONSENSE programs of the PHYLIP package (FELSENSTEIN 1993).

RESULTS AND DISCUSSION

Table 1 reports the list of species whose 18S rRNA has been used in the phylogenetic analysis. The CUG codon assignment, when known, is shown together with the genomic guanine + cytosine (G+C) content.

Figure 1 reports the phylogenetic tree obtained by applying DNAMLK program to the 20 yeast 18S rRNAs in Table 1. *C. cylindracea* 18S rRNA does not appear in the phylogeny reported in Figure 1 since it has not yet been sequenced. It is evident that the genus Candida is polyphyletic and has a monophyletic subgroup, including *C. guilliermondii, C. parapsilosis, C. tropicalis, C. viswanathii, C. maltosa* and *C. albicans,* which shows very high levels of intersequence similarity. The other Candida species appear to be interspersed in the tree; *C. glabrata* is more closely related to *S. cerevisiae, C. kefyr* more to *K. lactis, C. krusei* to *P. membranafaciens* and *C. lusitaniae* is more closely related to the latter pair of species. A similar phylogeny (data not shown) was obtained with a distance matrix calculated according to the TAJIMA-NEI method (TAJIMA and NEI 1984) and applying NEIGHBOR/DRAWGRAM programs under both the neighbor-joining and the UPGMA option. It is worth mentioning that when carrying out the evolutionary analysis of SSU rRNA using the DNAML program, which does not assume the molecular clock, the tree topology is slightly different (data not shown). The main difference is the position of *C. lusitaniae,* which groups together with the monophyletic subgroup of Candida.

The phylogenetic tree shown in Figure 1 seriously questions the present classification of Candida species that appears to be polyphyletic. To draw more accurate phylogenetic inferences, concordant results from several genes need to be obtained.

On the basis of the phylogenetic tree shown in Figure 1, we can predict that also *C. tropicalis, C. viswanathii* and *C. maltosa* have the same CUG codon reassignment. Indeed, while this paper was in preparation, it has been reported that in *C. maltosa* the codon CUG is read as

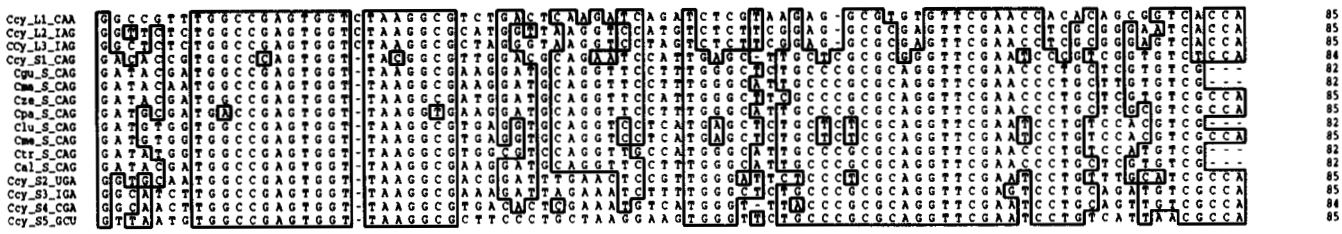

FIGURE 2.—Multiple alignment of the serine and leucine tRNA isoacceptors of *C. cylindracea* and of other Candida tRNA^{Ser}-(CAG) available in the database. Ccy, *C. cylindracea* (accession number, S42406); Cgu, *C. guilliermondii* (D17533); Cal, *C. albicans* (D13706); Cma, *C. maltosa* (D26074); Cze, *C. zeylanoides* (D12941); Cpa, *C. parapsilosis* (D14890); Clu, *C. lusitaniae* (D17534); Cme, *C. melibiosica* (D12940); Ctr, *C. tropicalis* (D14890). The *C. cylindracea* tRNA nomenclature is according SUZUKI *et al.* (1994). Ser-1 is responsible for translation of the codon CUG.

serine in agreement with our prediction (SUGIYAMA *et al.* 1995). Furthermore, after a database scanning we found three unpublished sequences for the tRNA^{Ser}-(CAG) of *C. guilliermondii, C. tropicalis* and *C. lusitaniae.* The finding of a tRNA^{Ser}(CAG) in *C. guilliermondii* and *C. tropicalis* is again in agreement with our prediction and suggests that the CUG codon reassignment is likely to have occurred before the common ancestor of *C. guilliermondii* (see Figure 1). On the contrary, the presence of a tRNA^{Ser}(CAG) in *C. lusitaniae* is unexpected, as this species, according to the philogenetic tree shown in Figure 1, is distantly related to the group of Candida showing deviation from the universal genetic code. This discrepancy can be explained by an incorrect species definition for this tRNA further supported by the striking observation that the presumed *C. lusitaniae* tRNA^{Ser}-(CAG) is 100% identical to that of *C. melibiosica* (see Figures 2 and 3).

As to sequence of events involved in the reassignment of the CUG codon, YOKOGAWA *et al.* (1992) suggested that the leucine-CUG codon could have disappeared together with a loss of the corresponding tRNA^{Leu}-(CAG) and that it was subsequently recaptured by serine. Further work (SUZUKI *et al.* 1994) has shown that the codon change of CUG from leucine to serine was caused by mutational events in the anticodon of a tRNA gene and not in the tRNA synthetase gene.

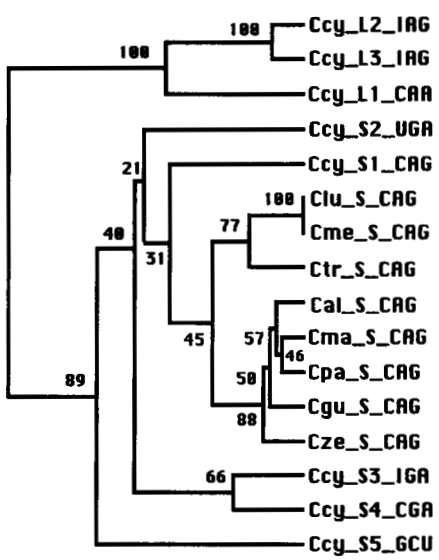

FIGURE 3.—Phylogenetic tree of the tRNA sequences in Figure 2 calculated with SMC (SACCONE *et al.* 1990), NEIGHBOR/UPGMA and DRAWGRAM programs (FELSENSTEIN 1993). Branch lengths are proportional to genetic distances.

It is therefore important to clarify the relationship between tRNA^{Ser}(CAG) and the isoacceptor tRNAs for serine and leucine. Figure 2 shows the multiple alignment, constructed according to their secondary structure, of five tRNA^{Ser} and three tRNA^{Leu} previously sequenced by SUZUKI *et al.* (1994) in *C. cylindracea* and all the other tRNA^{Ser}(CAG) sequences available in the data bank. The evolutionary position of *C. cylindracea* can be reliably predicted within the monophyletic group of Candida (shaded box in Figure 1) on the basis of their common deviation from the universal genetic code for CUG, as well as on the basis of the tRNA tree shown in Figure 3. The evolutionary relationships between these tRNAs, determined by using the SMC method, are depicted in the phylogenetic tree in Figure 3. The phylogenetic tree clearly supports a common origin for the tRNA^{Ser}(CAG) that are also more closely related to other serine tRNAs than to leucine tRNAs. In support of these findings, all tRNA^{Ser}(CAG) share a guanosine at position 33 and have a D-loop 10 nucleotides long, as other serine tRNAs, whereas leucine tRNAs have an 11 nucleotide D-loop.

The closer relationship of tRNA^{Ser}(CAG) to other tRNA^{Ser} than to tRNA^{Leu} suggests that tRNA^{Ser}(CAG) could have originated by gene duplication from a seryl-tRNA and thus has not evolved from a tRNA^{Leu}(CAG) as suggested by (SANTOS and TUITE 1995). Indeed, it is striking to note that some Candida tRNA^{Ser}(CAG) genes (*e.g., C. cylindracea, C. tropicalis, C. maltosa*) are interrupted by an intron in the anticodon loop so that it is possible intron acquisition and splicing played a role in the creation of the anticodon CAG (YOKOGAWA *et al.* 1992).

A unique feature in the tRNA^{Ser}(CAG) of *C. cylindracea* is that it is found in multiple copies (at least five), close in the genome, and all of them have the triplet CCA at the 3′ termini of the tRNA gene. The presence of the CCA triplet in the DNA suggests the possibility that during the evolution of the *C. cylindracea* genome

TABLE 2

CUG, serine and leucine codon usage calculated for 179 genes coding for proteins from the six *Candida* species with CUG coding or supposed to code for serine (shaded box in Figure 1) and other related yeast species

				Leu						Ser				
Organism	Genes	Codons	CUG	CUA	CUC	CUU	UUA	UUG	UCA	UCC	UCG	UCU	AGC	AGU
C. albicans	75	39,138	72	92	73	351	1318	1496	855	428	209	936	143	563
C. cylindracea	5	2749	92	0	113	37	0	118	3	22	33	4	65	14
C. glabrata	6	1140	1	15	3	10	53	17	33	14	2	16	6	9
C. guilliermondii	1	103	0	0	0	0	0	5	1	1	0	1	0	0
C. kefyr	16	5486	15	53	13	32	49	288	40	116	29	152	20	34
C. krusei	1	192	0	2	3	3	5	3	2	5	0	3	1	1
C. maltosa	23	9922	19	13	22	92	346	493	92	115	35	179	19	112
C. parapsilosis	10	2153	3	4	10	31	205	34	50	12	10	58	9	52
C. pelliculosa	6	2772	8	22	10	14	104	87	91	30	5	60	9	31
C. tropicalis	26	12,600	18	12	31	96	227	745	107	248	25	277	37	90
C. utilis	5	1468	24	9	27	18	1	59	12	25	18	33	13	7
H. polymorpha	22	7207	294	22	115	81	14	117	27	112	159	80	63	22
K. lactis	30	16,426	62	177	43	168	639	286	261	113	66	486	64	289
S. cerevisiae	891	459,247	4133	5511	1975	4638	11,252	14,237	7440	6843	3621	11,297	3628	5557

a multiplication process of the $tRNA^{Ser}$(CAG) gene took place via integration into DNA of reverse-transcripted $tRNA^{Ser}$(CAG) genes.

The data so far available suggest that the variation of the genetic code could have occurred as follows: (1) in an ancestor of the present CUG^{Ser} Candida, probably having a low genomic G+C content, CUG became unused and thus the corresponding $tRNA^{Leu}$ was lost; (2) a new tRNA gene was originated by gene duplication of a $tRNA^{Ser}$ gene; and (3) after the generation of the CAG anticodon sequence, possibly through point mutation in the anticodon loop of the duplicated $tRNA^{Ser}$ and/or intron acquisition, the new tRNA gene recaptured the CUG for serine.

Directional mutation pressure could have played an important role in step 1 above. Indeed, strong AT- or GC-pressures, which are know to heavily affect the codon usage at the level of the silent codon positions, could remove completely some codons by converting them into synonymous codons with concomitant disappearance or inactivation of the corresponding tRNA. Thus, the codon CUG, as it might have been the case for other CG-rich codons, disappeared after having been converted into the synonymous codons CUW or UUR.

The change of the universal serine codon UCN or AGY to CUG (step 3) could not occur through a single mutation event, but it must pass through an intermediate codon *e.g.*, UUR (Leu) or CCN (Pro). Consequently, the gene having such an intermediate codon might become nonfunctional. In the case of lipase genes of *C. cylindracea*, it is interesting to note that the catalytic serine is coded by CUG, and that out of the 24 CUG-Ser, 13 are in conserved positions in the lipase family (LOTTI *et al.* 1993). We can thus speculate that either the codon change was made possible by the presence of multiple copies of lipase genes, so that a deleterious mutation would not affect the viability of the cell, or that lipase was a dispensable gene in certain evolutionary period. To clarify this issue, it would be necessary to sequence the lipase genes in other closely related Candida species and to determine their copy number. It should be also quite interesting to analyze the codon usage of some housekeeping genes (*e.g.*, coding for enzymes of the glycolitic pathway).

The Candida species in which CUG is read as serine have quite a heterogeneous G+C genome content, from 63% of *C. cylindracea* to 36% of *C. albicans*. Table 2 reports the usage of CUG and other codons for leucine and serine in 11 Candida species based on a total of 179 genes whose sequences are present in the data bank.

It is striking to note that in *C. cylindracea* the most used serine codon is by far CUG, which accounts for ~40% of the serine codons. The exceptionally high usage of CUG in *C. cylindracea* might be explained by its overall high C+G genome content and by the existence of multiple genes for $tRNA^{Ser}$(CAG). Indeed, SUZUKI *et al.* (1994) identified in *C. cylindracea* five different $tRNA^{Ser}$(CAG). It would be very interesting to see if the codon usage pattern of *C. cylindracea* lipases actually represents a general codon usage pattern.

CUG is highly used also in *C. utilis* whereas in all the other cases, CUG is a very rare codon. This cannot be explained only on account of the low overall C+G content of the relevant genomes, as we observe, for example, in *C. albicans*, that the other serine codons ending in C or G (*e.g.*, UCC, UCG) are more abundant than CUG. The rare occurrence of CUG in several CUG^{Ser} Candida species (*e.g.*, *C. albicans* or *C. parapsilosis*) could be a remnant of the CUG disappearance produced by the AT-pressure on their ancestor. Indeed, after steps 2 and 3 in the CUG^{Ser} Candida ancestor, the genetic code variation was frozen and thus became independent from the C+G genome content.

In the attempt to date the genetic code change based on the phylogenetic tree in Figure 1, we made use of a vertebrate time scale, which is the only reliable measure available. We thus had to impose two *a priori* assump-

tions: (1) 18S rRNAs evolve according to the molecular clock and (2) vertebrate and yeast rRNAs have the same nucleotide substitution rate. Then by fixing the time of divergence between Xenopus and mammals at 350 million years ago (Mya) (MCLAUGHLIN and DAYHOFF 1972), the genetic code change can de dated at over 150 Mya that corresponds to the time of the common ancestor of all Candida species with the CUG codon reassignment.

The present paper shows a clear example of the relevance that evolutionary analysis and phylogenetic reconstruction play in some biotechnological applications. Indeeed, in this case the phylogenetic classification of some yeast species allows us to shed light on the change of the universal genetic code as well as to predict a nonuniversal code in some Candida species. The knowledge of the genetic code usage is very important for biotechnological applications that rely on the expression of heterologous genes.

We thank RON UPHOFF for providing information on yeast genomic G+C%. This work was partially financed by Ministero Universitä e Ricerca Scientifica e Tecnologica, Italy and Progetto Finalizzato Ingegneria Genetica, Consiglio Nazionale delle Ricerche, Italy.

LITERATURE CITED

BARREL, B. G., A. T. BANKIER and J. DROUIN, 1979 A different genetic code in human mitochondria. Nature **282:** 189–194.

BERBEE, M. L., and J. W. TAYLOR, 1992 Detecting morphological convergence in true fungi, using 18S rRNA gene sequence data. BioSystems **28:** 117–125.

FELSENSTEIN, J., 1993 PHYLIP (Phylogeny Inference Package). Department of Genetics, University of Washington, Seattle.

GCG, 1993 Program Manual for the GCG Package. Genetic Computer Group, Madison, WI.

KAWAGUCHI, Y., H. HONDA, J. TANIGUCHI-MORIMURA and S. IWASAKI, 1989 The codon CUG is read as serine in an asporogenic yeast *Candida cylindracea.* Nature **341:** 164–166.

KREGER-VAN RIJ, N. J. W. (Editor), 1984 *The Yeasts: A Taxonomic Study,* Ed. 3. Elsevier, Amsterdam.

LOTTI, M., R. GRANDORI, F. FUSETTI, S. LONGHI, S. BROCCA *et al.,* 1993 Cloning and analysis of *Candida cylindracea* lipase sequences. Gene **124:** 45–55.

MCLAUGHLIN, P. J., and M. O. DAYHOFF, 1972 Evolution of species and proteins: a time scale, pp. 47–66 in *Atlas of Protein Sequence and Structure,* edited by M. O. DAYHOFF. National Biomedical Research Foundation, Silver Springs, MD.

OHAMA, T., T. SUZUKI, M. MORI, S. OSAWA, T. UEDA *et al.,* 1993 Non-universal decoding of the leucine codon CUG in several *Candida* species. Nucleic Acids Res. **21:** 4039–4045.

OSAWA, S., A. MUTO, T. H. JUKES and T. OHAMA, 1990 Evolutionary changes in the genetic code. Proc. R. Soc. Lond. B **241:** 19–28.

OSAWA, S., T. H. JUKES, K. WATANABE and A. MUTO, 1992 Recent evidence for evolution of the genetic code. Microbiol. Rev. **56:** 229–264.

SACCONE, C., C. LANAVE, G. PESOLE and G. PREPARATA, 1990 Influence of base composition on quantitative estimates of gene evolution. Methods Enzymol. **183:** 570–583.

SANTOS, M. A. S., and M. F. TUITE, 1995 The CUG codon is decoded *in vivo* as serine and not leucine in *Candida albicans.* Nucleic Acids Res. **23:** 1481–1486.

SUGIYAMA, H., M. OHKUMA, Y. MASUDA, S.-M. PARK, A. OTHA *et al.,* 1995 *In vivo* evidence for non-universal usage of the codon CUG in *Candida maltosa.* Yeast **11:** 43–52.

SUZUKI, T., T. UEDA, T. YOKOGAWA, K. NISHIKAWA and K. WATANABE, 1994 Characterization of serine and leucine tRNAs in an asporogenic yeast *Candida cylindracea* and evolutionary implications of genes for tRNA^{Ser}CAG responsible for translation of a non-universal genetic code. Nucleic Acids Res. **22:** 115–123.

TAJIMA, F., and M. NEI, 1984 Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1:** 269–285.

YOKOGAWA, T., T. SUZUKI, T. UEDA, M. MORI, T. OHAMA *et al.,* 1992 Serine tRNA complementary to non-universal serine codon CUG in *Candida cylindracea:* evolutionary implications. Proc. Natl. Acad. Sci. USA **89:** 7408–7411.

Communicating editor: W.-H. LI